ENCYCLOPÆDIA
BRITANNICA



MACROPÆDIA

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



THE UNIVERSITY OF CHICAGO

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

The New

Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 14

Knowledge in Depth

FOUNDED 1768
15TH EDITION



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

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

Printed in U.S.A.

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

Peking

Peking (Bei-jing in Pin-yin romanization) is both a very old and a very new city—old and rich in its cultural and artistic heritage but new and thriving since 1949 as the capital of the People's Republic of China. Few cities in the world have served for so long as the political headquarters and cultural focus of a country as immense as China. The city has constituted an integral part of the history of China over the past eight centuries, and there is scarcely a building of any age in Peking that has not made its contribution toward the evolution of the country. As Peking is itself a China in miniature, it is impossible to understand China without a knowledge of this ancient city, which in 1970 had a population of 7,600,000 and a metropolitan area of 6,900 square miles (17,800 square kilometres).

More than 2,000 years ago, a site near present-day Peking was already an important military and trading centre for the northeastern frontier of China. Not until the Mongol dynasty (AD 1279 to 1368) was a successor city—called Ta-tu—to become the administrative capital of China. During the reign of the first emperor of the Ming dynasty (1368 to 1644), Nanking became the capital, and the old Mongol capital was renamed Pei-p'ing (Peace in the North); the third Ming emperor, however, restored it as the Imperial seat of the dynasty and gave it a new name, Peking (Northern Capital). It remained the capital until the 20th century, when, after the successful campaign of the Chinese Nationalist troops against warlords in Peking in 1927, Nanking was selected as the national capital, and Peking once again resumed its old name—Pei-p'ing—a name still used by the Nationalist government in Taiwan.

In spite of frequent political changes in China, the city throughout the earlier decades of the 20th century remained the most flourishing cultural centre in the nation. Peking's importance was never fully realized, however, until the city was chosen as the capital of the People's Republic of China in 1949, and this new political status has over the past two decades added much vitality to the ancient city. Indeed, few cities in the world have ever experienced such rapid growth in population and geographic area, as well as in industrial and other activities. Combining both historical relics of an ancient culture and new urban construction under a socialist system, Peking has become the showplace of modern China.

This article is divided into the following sections:

History
The contemporary city
- Physical organization
- Transportation
- Demography
- Housing and architecture
- The economy
- Political and governmental institutions
- Public utilities
- Health
- Education
- Cultural life
- The media
- Recreation

HISTORY

With but few interruptions, Peking has been the capital of China for some 700 years, and in the number of years as the Imperial capital it is exceeded only by Ch'ang-an (Hsi-an) in Shensi Province and Lo-yang in Honan Province. In prehistoric times the area around Peking was inhabited by some of the earliest human beings of whom we have knowledge. Between 1918 and 1939, the fossil remains of Peking man (formerly *Sinanthropus pekinensis*; now known as *Homo erectus pekinensis*), who lived about 500,000 years ago, and of Upper Cave man, who lived about 50,000 years ago, were unearthed at Chou-k'ou-tien, a village 34 miles southwest of Peking.

While long periods in Peking's early history necessarily remain blank, it is certain that, some 3,000 years ago, Neolithic communities settled down on or near the site where Peking now stands. During the Warring States period (481–221 BC) of the Chou dynasty (from about 1122 to 221 BC), one of the powerful feudal states, the kingdom of Yen, established its capital named Chi near the present city of Peking; this was the first capital city to be associated with the site. The city was destroyed by the troops of Shih Huang Ti, founder of the Ch'in dynasty (221 to 206 BC).

Chi—the capital of Yen

During the Ch'in dynasty, the Yen capital was incorporated into one of the 36 prefectures then established throughout the country. A new town was built during the Han dynasty (206 BC to AD 220) and was known as Yen. But, throughout the Han dynasty and the turbulent centuries that followed, the place remained a provincial town, witnessing most of the time the fateful struggle between the Han Chinese to the south and the nomadic Hsiung-nu, or Huns, to the north.

During the period of Three Kingdoms (AD 220 to 264), the city was again called Yen. The northern border of ancient China ran close to the present city of Peking, and northern nomadic tribes frequently broke in from across the border. Thus the area that was to become Peking became an important strategic as well as a local political centre.

For nearly 300 years (from the end of the Chin dynasty in 317 to the beginning of the Sui dynasty in 581), the northern territory, including the site where Peking now stands, was largely under the control of invading nomads. It was not until the T'ang dynasty (618 to 907) that it was recovered by the Han people and known as Yu-chou. By the middle of the T'ang dynasty, measures were taken to prevent the nomadic Tangut tribes of Tibet, the Hsi Hsia and the Khitans (a Manchurian people), from raiding the border lands and the local capital. The position of Yu-chou consequently became increasingly important. On the fall of T'ang China, a number of states emerged in North China. One of these states was established by the Khitans; after destroying the city of Yu-chou, the Khitans established the Liao kingdom (947 to 1125) and built one of their capitals on approximately the same site, calling it Nan-king (the Southern Capital) to distinguish it from other capitals in their Manchurian homeland (also distinct from the modern Nanking or Nan-ching in the Yangtze Valley). The Liao capital was bounded by a square wall, almost 14 miles in circumference and some 32 feet high. It had eight gates and enclosed a fine Imperial palace in the centre, which indicated the strong influence of Chinese city planning.

In the middle of the 12th century, when the Juchen, a Manchurian people from the Amur Valley, defeated the Liao and established the state of Chin, the Liao capital was rebuilt as the new Chin capital and renamed Chung-tu (Central Capital). Chung-tu under the rule of the Juchen (Golden Tartars) was constructed on a larger scale, with spendidly decorated palaces and halls.

Between 1211 and 1215 the Mongols, under the leadership of Genghis Khan, one of the great conquerors of

history, repeatedly attacked and finally took the city from the Chin. In the battle the palaces of Chung-tu were set on fire and blazed for over a month. When all China fell to the Mongol hordes, Kublai Khan (1215 to 1294), a successor to Genghis Khan and founder of the Yüan dynasty (1279 to 1368), determined to build a new capital at Peking, abandoning the old city of Karakorum in Mongolia. In 1272 he named the new capital Ta-tu (Great Capital); under Mongol control it became for the first time the political centre of all China.

The Mongol capital of Ta-tu

Ta-tu was larger than any of its forerunners and was rebuilt slightly northeast of the old site. The square of the outer wall measured about 18 miles in length and enclosed an area of more than 20 square miles. The city walls were built with pounded earth, and once a year labourers were called in to repair them with mud. The Imperial Palace, which was approximately to the west of the existing one, was situated in the southern half of the capital city. The chief palace architect at the time was an Arab, appointed by Kublai Khan. The city of Ta-tu exemplified the imposing and variegated architecture of the Mongol period. The square walls and the 12 gates were all modelled on the Chinese plan, but the inner chambers and living quarters were often in the style of Mongolia or of Central Asia. It was at this time that a canal, the T'ung-hui Canal, was dug and connected to China's Grand Canal so that boats transporting the tribute rice from the provinces south of the Yangtze could sail into one of the new lakes inside the city. Ta-tu, which had magnificent Imperial palaces and treasures drawn from every corner of the country, was the scene of stupendous feasts given by the khan (ruler) on state occasions. These characteristics and the well-organized post-stages on the roads leading to the city astounded Marco Polo (*q.v.*), the Venetian traveller who visited Ta-tu in the 1280s.

In the middle of the 14th century, Chu Yüan-chang headed a peasant revolt that overthrew the Mongol dynasty and established the Ming dynasty (1368 to 1644). He moved the capital to Chin-ling, near his own hometown in Anhwei Province, and called it Nanking; Ta-tu was renamed Pei-p'ing (Northern Peace) and was placed under his son's rule. On Chu's death (1399) the throne should have passed to his grandson in Nanking, but his son, Prince Yen, who ruled Pei-p'ing, usurped the throne. In consequence, in 1403 the city was renamed Peking (Northern Capital), and in 1420 it was officially made the capital city of the Ming dynasty.

The 15th-century city

Peking in the Ming period grew on a yet grander scale than under the Mongols. The present city walls, moats, palaces, and temples were mostly built in the 15th century. The old city of Ta-tu, including the palaces, was largely demolished. The new city was situated further south, thus leaving the northern part of the Mongol city derelict while at the same time slicing off one gate from the east and west walls, respectively. In 1553, an outer city wall was begun, to include the increasing number of inhabitants living outside the city, but, when the entire construction was subsequently found to be too costly, the plan was abandoned on the completion of the south wall, thus producing the present shape of the walled city. Unlike the city wall of pounded earth of Mongol times, the walls of the Ming city were faced with a layer of bricks, in order to prevent weathering.

In 1644 Peking was taken over by Li Tzu-ch'eng, who led a peasant uprising against the Ming regime. Li's army held it for only 40 days, for the Manchus were simultaneously preparing an incursion south of the Great Wall, and—thanks to the treachery of a Ming general who opened the gate—they swept down on the city. Peking fell intact and in the same year was declared the Manchu capital by Shun-chih, the first emperor of the Ch'ing (also known as Manchu) dynasty (1644 to 1911).

Peking remained superficially the same throughout Ch'ing times. The city plan was unaltered, though many palaces, temples, and pavilions were added outside the walls to the west, notably the Yüan Ming Yüan (old Summer Palace), built in the 17th century, and the Yi Ho Yüan (new Summer Palace), built in the late 19th century. The old Summer Palace was completely destroyed by fire in 1860 by the allied British and French troops in Peking during the "Arrow" War (1856–60). In the same year, as a result of the treaties of Tientsin in 1858, a permanent British embassy was established in the city, and a legation quarter, situated to the southeast of the palace ground, was reserved for British and other embassies. The legation quarter was besieged for nearly two months by the Boxer rebels in 1900.

After the revolution of 1911, Peking remained the political centre of the Republic of China until 1928, when the Nationalists moved the capital to Nanking, at which time Peking was renamed Pei-p'ing once again. The city was occupied by the Japanese from 1937 to 1945. In 1949, when the Communists established the People's Republic of China, Peking (with its old name restored) was chosen as the capital of the new regime.

THE CONTEMPORARY CITY

Physical organization. *The city site.* Peking is situated at the northern apex of the triangular North China Plain and lies at an elevation of between about 100 and 130 feet above sea level. To the north lies the fringe of the Mongolian Plateau; to the northeast rises the mountain range of the Yen Shan (Yen Mountains), part of which forms the eastern part of the concave arc that circles the Peking lowland from the northeast to the southwest and that is known to geologists as the "Bay of Peking."

The city was built at the mouth of this embayment, which opens onto the great plain to the south and east, and between two rivers, the Yung-ting and the Ch'ao-pai, which eventually join to empty into the Gulf of Chihli, some 100 miles to the southeast of Peking. Peking is a natural gateway on the long-distance land communication route between the North China Plain and the northern ranges, plains, and plateaus. To the south of the city the plain spreads out for about 400 miles until it merges into the lower valley and the delta of the Yangtze. On the east the plain is bounded by the sea, except for the break caused by Shantung Hills; on the west it is flanked by the T'ai-hang Shan, which constitute the eastern edge of the plateau of loess (loamy material deposited by wind). Since Peking stands at the apex of the triangle, routes running across the great plain naturally converge on the city.

Since the dawn of Chinese history, the Yen Shan range has constituted a formidable barrier between the North China Plain to the south, the Mongolian Plateau to the north, and the Liao Ho (Liao River) Plain in the southern region of the Northeast (formerly Manchuria). A few passes, however, cut through the ranges—the most important being Nan-k'ou, Ku-pei-k'ou, and Shan-hai-kuan—and are so situated that all roads leading from Mongolia and the Northeast to the North China Plain are bound to converge on Peking. For more than 700 years, therefore, Peking has been an important terminus of the caravan routes leading to and from the vast Central Asian hinterland.

The metropolitan district

Since the early 15th century, the city of Peking and its surrounding territories have been organized as a metropolitan district of enormous size, having a governor—formerly appointed by the emperor himself—equal in rank to a provincial governor. This special district organization was continued by the Ch'ing dynasty and by the People's Republic of China. The present metropolitan boundary was established in 1959 and covers a territory of approximately 6,600 square miles. The metropolis consists of eight *ch'u* (municipal districts) in the centre and nine annexed *hsien* (counties) in peripheral areas. It may be divided into three concentric zones, based on urban functions. The central zone coincides with the walled city; it is mainly occupied by old palaces, government buildings, commercial districts, and old residential areas and makes up roughly 0.4 percent of the total metropolitan area. Four of the *ch'u* lie within the walled city. The second zone, the near suburb, comprises three *ch'u* immediately surrounding the walled city; it provides sites for new factories, schools, government buildings, and workers' dormitories. The outer fringe of this zone is intensively cultivated and supplies vegetables and fruits to

the population of the central zone. The near suburb accounts for 7.6 percent of the metropolitan area. The third zone, the far suburb, consists of one *ch'u* and nine annexed *hsien*. This zone constitutes 92 percent of the metropolitan area and functions as the economic base—supplying coal, lumber, construction material, dairy products, water, and some grain crops to the urban population in the central zone and the near suburb. The far suburb of Peking is mainly under the administration of people's communes.

The environment. Though the distance from Peking to the Gulf of Chihli is only about 113 miles, the general air circulation in Peking is mainly from the northwest throughout the year. Maritime effects on Peking's weather are meagre. The climate is clearly of the continental monsoon type that occurs in the temperate zone. Winters are cold and dry because of the Siberian air that passes over the Mongolian Plateau and moves southward into China proper. In summer the warm and humid air from the southeast often penetrates into North China, bringing Peking most of its annual precipitation.

Local topography has a great effect on Peking's climate. Being located in the lowland area and shaded by mountains, the city is a little warmer in winter than other areas of China on the same latitude; it has only five months in a year in which the mean monthly temperature is lower than 50° F (10° C). Wind direction in Peking is also influenced by topography, with changes occurring from day to night. Generally, there are more southerly winds in the day but northerly or northwesterly winds at night.

Temperature and rainfall

Based on meteorological data from 1841 to 1956, the annual mean temperature of the city was 53° F (about 12° C). The coldest month was January, when monthly mean temperatures were 24° F (about −4° C), and the warmest month was July, when the monthly mean was 79° F (26° C). In an average year, there are 132 days of freezing temperature in Peking (between October and March); the mean amount of precipitation is 25 inches (635 millimetres), with 75 percent of the annual total falling in the summer months from June to August. July is ordinarily the wettest month of the year, with monthly mean precipitation averaging about nine inches.

One of the characteristics of Peking's rainfall is its variability. In 1959—an extremely wet year for Peking—the total precipitation amounted to 55 inches, whereas in 1891—an extremely dry year—only seven inches of rainfall was recorded. The average number of rainy days is about 80, and the average relative humidity for the city is 57 percent.

Winter in Peking is long and usually begins in October, when northwesterly winds gradually gain strength. Until March, this seasonal wind system dominates Peking; coming from the Mongolian desert, the wind is cold and dry, bringing little snow or other precipitation. The monthly mean temperature from December to February is below 32° F (0° C), and during this period it is not unusual for a daily mean temperature of 14° F (−10° C) or lower to last for days. Spring, the windiest season, is short and rapidly becomes warm. Total precipitation from March to May is less than three inches. Mainly because of the prevailing high winds, the evaporation rate in spring averages 24 inches—about nine times the total spring precipitation—and frequently is sufficient to cause droughts that are harmful to agriculture. In addition to being the season of torrential rains, summer is rather hot, with the average temperature for June, July, and August being above 75° F (24° C). Autumn begins in late September and is a pleasant, though short, season when the sky is clear and the temperature comfortable.

Vegetation

The mountain areas of Peking belong to the temperate deciduous forest zone, while the plains area belongs to the wooded steppe zone. Because of continuous deforestation by man for several centuries, however, only the mountain areas in the northeast and the western parts of the metropolitan area are still covered with mixed forests, composed mainly of pines, oaks, and Manchurian birch. Different types of forest, divided into vertical zones, can be seen at higher elevations. On the lower slopes of many hills to the west of the city, human activities have resulted in the disappearance of forest cover so that only bushes and shrubs dot the landscape. On sunny slopes at altitudes of between about 2,300 and 5,600 feet, a variety of species occurs, including Manchurian birch (*Betula mandshurica*), Dahurian birch (*Betula davurica*), trembling poplar (*Populus tremula*), Mongolian oak (*Quercus mongolica*), and Liaotung oak (*Quercus liaotungensis*). Between 5,600 and 6,250 feet above sea level, the mixed forest of truncated maple (*Acer truncatum*) and trembling poplar replaces all other species. Above 6,250 feet, goat willow (*Salix caprea*) becomes the dominant tree. A larger part of the plain areas of Peking has been either cultivated or occupied by various settlements so that it is for the most part bare of any natural vegetation. Occasionally, some small groves of planted trees may be seen near graveyards or in the vicinity of villages; these are mainly composed of mixed woods consisting of oil pine, Chinese juniper, Chinese cypress, willow, elm, and Chinese locust.

The city plan. The city proper of Peking essentially consists of two walled cities, the northern inner city and the southern outer city. The inner city, also known as Tatar City, lies approximately on the site of the Mongol city of Ta-tu; it is in the form of a square, with walls nearly 15 miles in length. The outer city, also known as the Chinese City, was added during the region of the Ming emperor Chia-ching (1521 to 1566); it is in the form of an oblong adjoining the inner city, with walls 14 miles in length, including four miles of the southern wall of the inner city. Within the inner city is the former Imperial City, also in the form of a square, which has red plastered walls six and a half miles in length. Within the Imperial City, in turn, is the moated "Forbidden City," with walls two and a quarter miles long. The Forbidden City contains the Imperial Palaces, which have been converted into a museum.

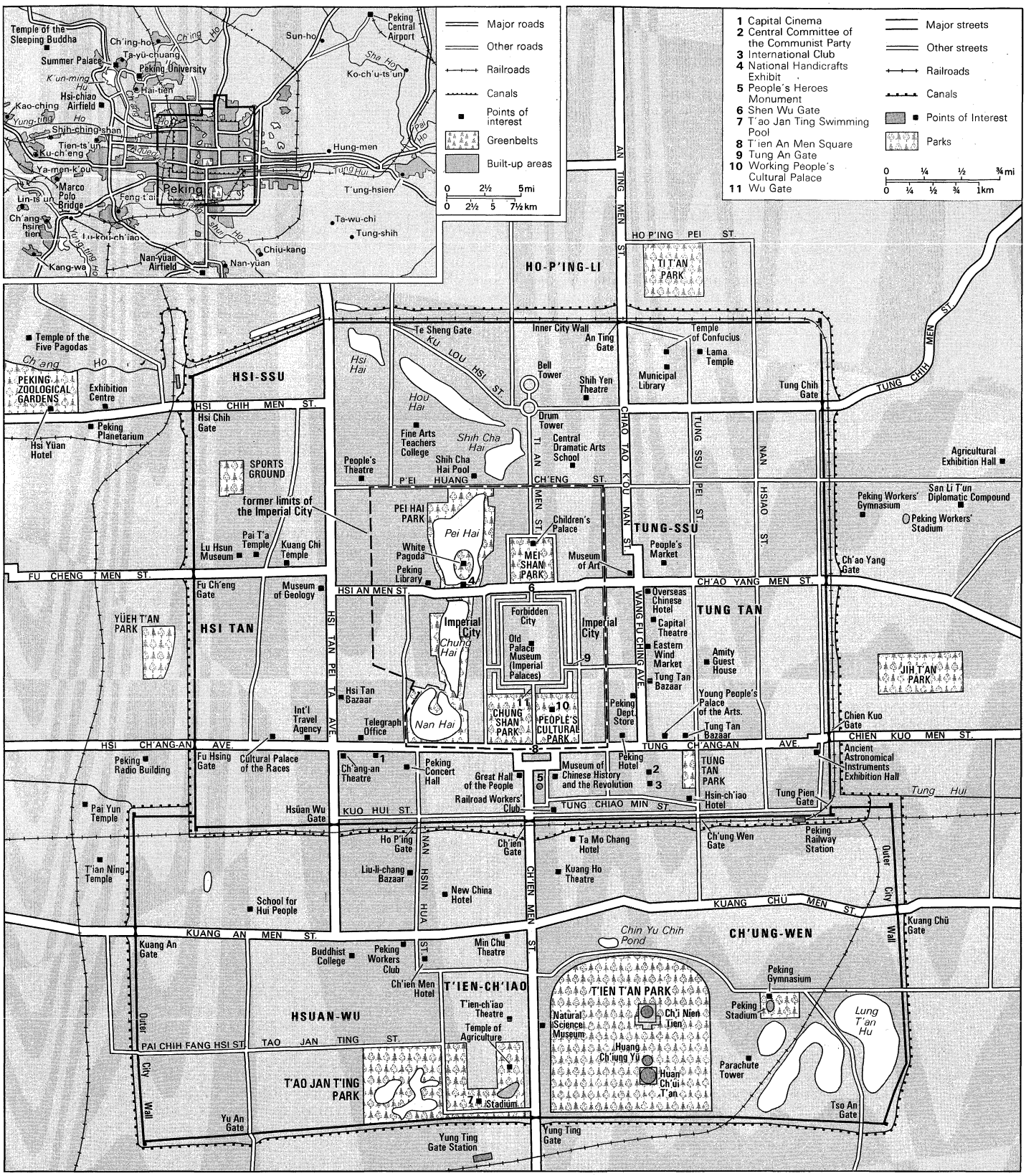
Peking represents, better than any other existing city, the heritage of Chinese architectural achievement. During each dynasty in which Peking was the capital, care was consistently taken to preserve tradition unchanged when the city was rebuilt or remodelled. Few cities in the world can thus rival Peking in the regularity and harmony of its city plan.

The north–south axis

The plan is composed about a single straight line, drawn north and south through the centre of Peking, on which the internal coherence of the city hinges. All the city walls, important city gates, main avenues and streets, religious buildings, and daily shopping markets are systematically arranged in relation to this central axis. Because the central axis has historically signified the authority of the ruling dynasty, many official buildings, public grounds, and city gates are located along this line. From north to south this line passes through the Chung Lou (Bell Tower); the Ku Lou (Drum Tower); Mei Shan (Coal Hill) Park; the Forbidden City, including the Imperial Palaces; T'ien An Men Square; Ch'ien Men (Front Gate); T'ien-ch'iao (section of town); and Yung Ting Men (Gate of Eternal Stability).

The symmetrical layout of the city to the east and west of this line is quite striking. In front of the palaces, the Temple of the Imperial Ancestors (now the People's Cultural Park) on the east side of the axis is balanced by the Altar of Earth and Harvests (now the Chung Shan Park) on the west. Farther away from the palaces, the Tung Tan (East Bazaar) is balanced by the Hsi Tan (West Bazaar); these form the main business districts. In the outer city, the T'ien T'an (Temple of Heaven) Park and the Ti T'an (Altar of the Earth) Park are similarly situated on both sides of the axis. Of the 16 city gates constructed in Ming times, seven are located on each side of the north–south line, and two are situated on the line itself.

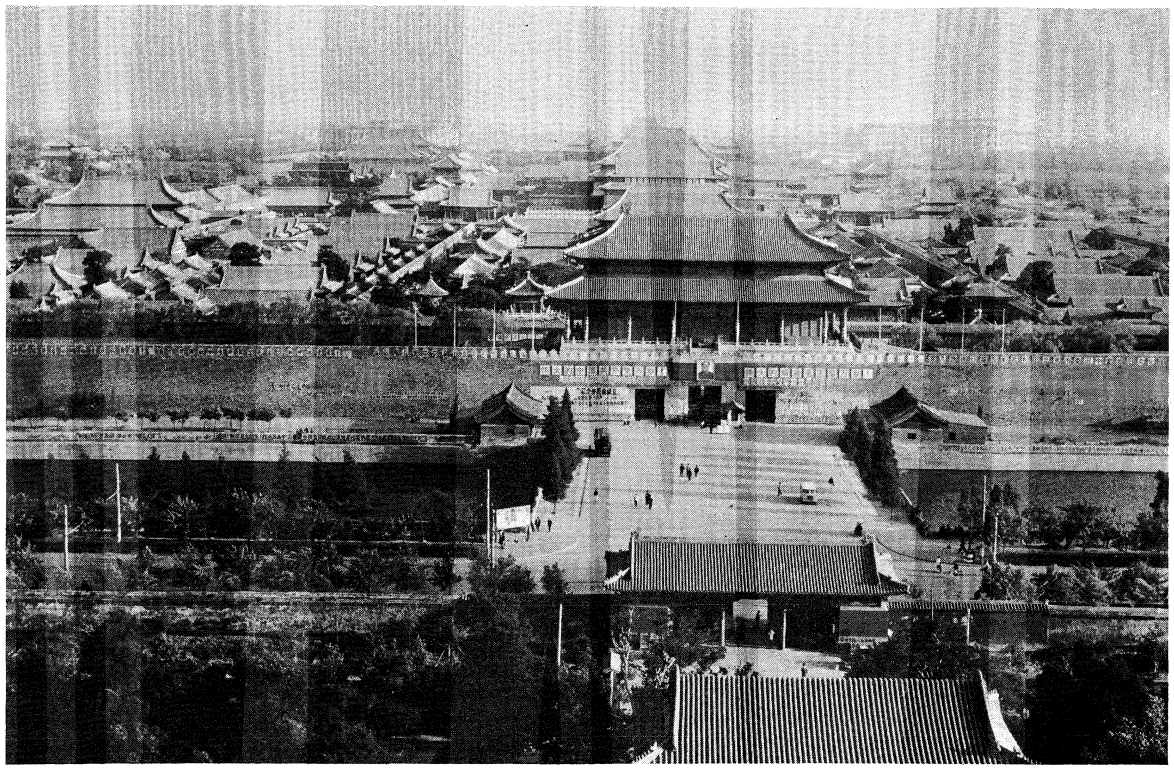
The main avenues of Peking, whether running north–south or east–west, connect the gates on the opposite walls and divide the whole city into a rectangular grid. Within the walls, buildings are constructed around a courtyard or series of courtyards, with every important building facing south. Buildings often stand behind one another along a north–south line, with small courtyards in between. The prevailing southern orientation of buildings has its climatic functional basis and also appears to



The city of Peking and (inset) its metropolitan area.

have been sanctified or conventionalized early in the Bronze Age in connection with ancestral ceremonies and with the worship of Heaven and Earth.

Since 1949, the greatest change in the city's appearance has been the extension of its streets immediately outside the old city walls. On the west side of the old city, an area extending about a mile out from the Fu Hsing Men has become a new extension of the Hsi Ch'ang-an Avenue and is used primarily for government offices. To the north, government offices are situated in an area of approximately the same width. Toward the Summer Palace, to the northwest, is an area where the most important universities and research institutes of the country are located. To the north of the city, the outlying districts between An Ting Men and Te Sheng Men are being developed as a new housing area adjoining the educational district in the



Forbidden City as seen from Mei Shan. The Great Hall of the People on T'ien An Men Square is visible in the background at right and the Museum of Chinese History and the Revolution at left.
Paolo Koch—Rapho Guillumette

northwest. Plans call for factories to be located around the east and south suburbs, although the southern suburbs still retain a predominantly rural aspect.

Rail communications

Transportation. Peking is the railroad centre of China, forming the terminus of four major lines and three local lines in the national railroad network. The four major lines radiating from Peking provide connections with Shenyang (Liaoning Province), Shanghai Municipality, Wu-han (Hupeh Province), and Pao-t'ou (Inner Mongolian Autonomous Region), respectively, each city being a major national industrial and commercial centre. The local lines are of regional importance and serve outlying districts within the metropolitan area; they are the Peking–Ku-pei-k'ou line to the north, the Peking–T'ung-hsien line to the east, and the Peking–Ta-tai line to the west. Peking is linked by direct express with several other large urban centres, such as Tsingtao (in Shantung), T'ai-yüan (in Shansi), Sian (in Shensi), Lan-chou (in Kansu), Ch'ang-ch'un (in Kirin), and Man-chou-li (in Heilungkiang). Under international agreements, Peking is also periodically connected by express train with Moscow; P'yŏngyang, North Korea; Ulaanbaatar, Mongolian People's Republic; and Hanoi, North Vietnam.

There are four major railroad stations in the city: Peking, Hsi Chih Men, Te Sheng Men, and Tung Pien Men. Peking Station, completed in 1959, is by far the largest and most important of these; located almost two miles from the heart of the city near the eastern end of the main thoroughfare, Tung Ch'ang-an Avenue, it has the capacity to serve more than 200,000 people a day and is the terminal and transfer point for all express lines. The new facilities were built to replace the Ch'ien Men Station (now the Railroad Workers' Club), built in 1906 outside the front gate of the inner city. The new station has more than 2,800,000 square feet of floor space in its central concourse and two wings, while its 18 waiting rooms can accommodate 14,000 people at one time. The roof of the beamless concourse is constructed in the form of a hyperbolic shell supported on four columns; it reaches the height of a ten-story building. Platform No. 1, at which the international trains pull in, is wide enough to hold a military review and is equipped with cinematographic and television facilities. The other three stations are mainly for local commuting trains.

While railroads connect Peking with distant cities, highways are the principal means of communication with communities within a 125-mile radius. There are 12 major highways radiating from the city, along which regular bus services run to Tientsin (in the Tientsin Municipality, an autonomous area in Hopeh Province), Chou-k'ou-tien, Nan-k'ou, the Ming Tombs (all in the Peking Municipality), and T'ang-shan and Ch'eng-te (in Hopeh). Five major bus depots are located in the walled city.

Peking is also the centre of China's civil airlines. The nation's major air routes radiate from Peking to Shanghai, Canton (in Kwangtung Province), Ch'eng-tu (Szechwan), Chungking (Szechwan), Urumchi (Sinkiang Uighur Autonomous Region), Harbin (Heilungkiang), K'unming (Yunnan), Nan-ning (Kwangsi Chuang Autonomous Region), Lhasa (Tibetan Autonomous Region), and Hsi-ning (Tsinghai).

Peking's intracity communications and commuting services are provided by a network of a city-owned system of buses and trolleys. The public transport system carries about 2,000,000 riders every day. The city's population is served by almost 120 routes with a route mileage of about 1,250 miles and by more than 1,000 buses and 430 trolleys. At places to which a large number of people are carried, such as the Shih-ching-shan steel mill and the Ch'ang-hsin-tien locomotive works (both in the outlying areas), bus routes run through the plant itself. The trolley routes run through the densely populated districts of the inner and outer walled cities; the pattern is much influenced by the symmetrical arrangement of the city plan. Bus routes also run throughout the walled city and the suburban areas. Bus fares are determined by distance, and monthly passes good for both trolleys and buses are available. There are also about 200 taxicabs on the city streets. For short-distance intracity travel, bicycles are very popular, constituting the bulk of street traffic.

Demography. Since 1949, Peking has become one of the most rapidly growing urban centres in China. The population change in historical times is illustrated by Table 1.



Young people painting Maoist slogans on the pavement in front of the Ch'ien Men railway station in 1967 (now the Railroad Workers Club).
Der Stern—Black Star

Population growth

During the dynastic period, when Peking was the capital, from the 13th to the beginning of the 20th century, the population of Peking seems to have increased very slowly in numbers, but since the downfall of the Manchu dynasty it has increased at an accelerating rate. The rapid increase during the four decades from 1910 to 1949 may be attributed to several factors. First, the rural disorder on the North China Plain during the rule of the warlords caused extensive migration to the walled city of Peking; second, the loss of Manchuria to Japan in 1931 made Peking a shelter for thousands of refugees from the northeastern provinces; third, the establishment of numerous institutes of higher education, sponsored by the government or by foreign missionaries, stimulated the growth of the student population in a city that was a cultural as well as a political capital.

Table 1: Peking Population Growth

year	population
1280	401,000
1578	707,000
1913	728,000
1916	801,000
1930	1,300,000
1936	1,551,000
1948	1,722,000
1949	2,360,000
1953	2,768,000
1958	4,148,000
1959	6,800,000
1970	7,600,000

The phenomenal increase in Peking's population did not occur, however, until the city had resumed its role as the national capital. In 1949 Peking, with a population of 2,360,000, was the fourth largest city in the country; ten years later a total of 6,800,000 made it the second largest. This increase resulted primarily from three circumstances: the necessary migration of thousands of government workers to the new capital, the rapid rate of industrial development in the municipality, and the extension of the municipal boundary in 1959 to include large rural areas. The 1953 census indicated that Peking had 708,000 persons living in rural areas—about 25 percent of the total population of the municipality. With the expansion of the city boundary in 1958 to include nine counties within the metropolis, it is likely that at least one-third of the total population may now be considered rural.

Housing and architecture. *Housing.* To cope with the rapid growth of population, some 200,000,000 square feet of floor space of new housing was constructed in the 1950s and 1960s for office and factory workers. During the time of the first five-year plan (1953 to 1957), new housing projects were concentrated in the western outskirts of the city, where apartment buildings were erected near the newly built government offices outside the Fu Hsing Men. After 1958 the construction of a large group of multi-unit housing estates was begun in the northern outlying districts between the city gates of An Ting and Te Sheng, centring on the new residential district of Ho-p'ing-li. The Ho-p'ing-li housing development contains primary and secondary schools, nurseries, hotels, and recreational facilities, as well as more than 100 four- or five-story apartment buildings. In addition, there are many groups of single-family houses in the northern suburbs, with associated parks, theatres, and recreational centres. All these buildings are supplied with gas and water, in contrast with the older parts of the city, where the existence of old stone paving had hindered the development of modern structures and facilities.

In the area outside the city gate of Chien Kuo, to the east of the city, apartment buildings accommodate the families of office workers employed in nearby government office buildings. This area has also been selected as the new diplomatic district and contains many foreign embassies, together with a number of Western-style houses for the housing of diplomatic representatives and their families.

In the older districts inside the city walls, many dilapidated houses have been pulled down in municipal housing-renewal projects and have been replaced by multistory apartment buildings. Urban-renewal projects, however, had, by the early 1970s, been unable to match the growth of population. In consequence, many traditional house compounds—originally designed centuries ago to house officials' families during the Manchu dynasty—were repaired or renovated and subdivided to provide quarters for three or four families in each compound; each family in a compound faces a public courtyard and shares a common front gate with other families.

Workers' community housing

In order to minimize commuting traffic in the metropolis, many factories in the eastern and southern outskirts of the city have erected four-story apartment buildings for workers' living quarters. These workers' residences constitute independent communities of their own and are so located that they are easily accessible from the place of work yet are far enough away to be free from noise and smoke. Rents in these community housing units are rather low compared with the total family income of average workers. A 1965 survey by the Peking Housing Bureau of 5,000 families living in houses operated by the city or various public enterprises found that the average rent amounted to about 6 percent of each family's monthly budget.

Building types and architectural features of note. Peking's heritage of Chinese architectural achievement is exemplified by both private housing and public buildings. As the whole city was laid out upon a rectangular street pattern symmetrically arranged around a centrally located palace, almost every dwelling unit in the city is rectangular in form, with the four sides squarely facing the cardinal directions. Most houses in the inner city were designed as residences of former officials and their families, and almost every dwelling compound is surrounded by high walls, with an open courtyard in the centre flanked by houses on the eastern, western, and northern sides, usually one story high. The former residences of high-ranking officials were composed of two or three compounds, interconnected along a north–south axis.

Stepping over the high wooden gate-sill of the front gate of a large compound, the visitor will find a brick screen wall located a few feet inside the gate—a structure that was supposed to shut off the intrusion of evil spirits, as well as to prevent curious passersby from looking inside. Beyond the screen is the outer courtyard, or service courtyard, which is flanked by houses to the east and west. Here, in former days, were the kitchen and living quarters for the gatekeeper, servants, and any visiting guests and relatives. A red-painted gate leads through the north wall of the outer court into the main part of the house. This is built around three sides of the main courtyard, which, usually shaded by a large tree, was the centre of the family's life. All the windows look inward to the courtyard, and a double door opens into the courtyard from each of the three wings. The windows extend from three feet above the ground up to the deep, overhanging eaves. As they face south, the rooms in the main building get the maximum possible sunshine in winter, and the eaves provide a pleasant shade in summer, when the sun is high. The wing at the northern end of the court was intended for the head of the family and his wife. It is divided into three compartments, the central one being the living or community room, and the smaller rooms at either side being the bedroom and study. The rooms facing east and west—three on each side of the court—were for married sons and their families. This is the basic plan of all the old houses in Peking. Larger families built an extra courtyard behind the main house, for it was regarded as essential that all the existing generations should live together. Since 1949, however, as mentioned, a great many of the old-style houses have been adapted for use by several families.

While the style and architecture of private dwelling units are uniform throughout the city, the public buildings and temples are characterized by a variety of designs and structures. As the nation's political and cultural centre for some 700 years, Peking possesses more buildings of historical and architectural significance than any other contemporary city in China. Since 1949 many new government and municipal buildings, combining both traditional and Western architecture, have added to the city's landscape.

The Great Hall of the People

Perhaps the most imposing structure constructed in the heart of the city since 1949 is the Great Hall of the People, an equivalent of the Capitol, in Washington, D.C. The Great Hall is located on the southwestern side of the T'ien An Men Square and is an immense building with tall columns of gray marble set on red marble bases of floral design. It has a flat roof with a golden-yellow tile cornice over green eaves shaped like lotus petals. Its frontage is 1,100 feet long—about the equivalent of two city blocks. Its floor space is 6,062,000 square feet—more than that of all the palaces in the old Forbidden City. The grand auditorium, seating 10,000, is where the National People's Congress holds its sessions. Other components are a banquet hall for 5,000, huge lobbies, and scores of meeting rooms and offices for the standing committee of the congress. The base of the building is of pink granite, and its walls are apricot yellow. Inside the building, the ceiling and walls are rounded. The focus of the lighting system is a red star surrounded by golden sunflower petals.

Among the historical and religious structures in Peking, the T'ien T'an (Temple of Heaven), located in the outer city, is unique both for its unusual geometric layout and because it represents the supreme achievement of traditional Chinese architecture. Entering from the western gate of the temple, along a path that is about 1,600 feet in length and shaded by ancient cypresses, one comes to a raised passage about 1,000 feet long. This broad walk connects the two sets of main buildings in the temple enclosure. To the north lies the Ch'i Nien Tien (Hall of Prayer for Good Harvests) and to the south the Huang Ch'iung Yü (Imperial Vault of Heaven) and the Huan Ch'ui T'an (Circular Mound Altar). The three buildings are built along a straight line. A bird's-eye view of the enclosure would show that the wall to the south is square, while the one on the north side is semicircular. Such a pattern symbolizes the traditional Chinese belief that Heaven is round and Earth square.

The Ch'i Nien Tien, built in 1420 as a place of heaven worship for the emperors, is a lofty, cone-shaped structure with triple eaves, the top of which is crowned with a gilded ball. The base of the structure is a triple-tiered circular stone terrace more than 63,000 square feet in area. Each ring has balustrades of carved white marble, which gives the effect of lace when seen from a distance. The roof of the hall is deep blue, resembling the colour of the sky. Without the use of steel and concrete, the entire structure, 125 feet high and about 100 feet in diameter, is supported by 28 massive wooden pillars. The four central columns, called the "dragon-well pillars," represent the four seasons; there are also two rings of 12 columns each, the inner ring symbolizing the 12 months and the outer ring the 12 divisions of day and night, according to a traditional system. The centre of the stone-paved floor is a round marble slab that has a design of a dragon and a phoenix—traditional symbols of king and queen. The hall has no walls, only partitions of open latticework doors.

The Huang Ch'iung Yü, first erected in 1530 and rebuilt in 1752, is a smaller structure some 64 feet high and about 50 feet in diameter. The circular building has no crossbeam, and the dome is supported by complicated span work. Its decorative paintings still retain their fresh original colours.

South of the enclosure lies the Huan Ch'ui T'an, built in 1530 and rebuilt in 1749. The triple-tiered white stone terrace is enclosed by two sets of walls that are square outside and round inside; thus, the whole structure forms an elaborate and integrated geometric pattern. The inner terrace is 16 feet above the ground and about 100 feet in diameter; the middle terrace is about 165 feet across and the lowest terrace some 230 feet across. Each terrace is encircled by nine rings of stones. Both the Huang Ch'iung Yü and the Huan Ch'ui T'an were erected to portray the geometric structure of the heaven, as conceived by the architects of the Ming dynasty. After 1949, the whole enclosure of the T'ien T'an was repaired; it is now a public park.

The economy. One of the main differences between the Imperial capital of former times and present-day Peking is that the Imperial capital was a consumption rather than a production centre, receiving supplies of all kinds from other parts of the country. In the 1950s and 1960s, however, Peking became one of the nation's principal industrial centres. The first five-year plan emphasized the



Children sweeping the steps at the Ch'i Nien Tien, or Hall of Prayer for Good Harvests, central building of the Temple of Heaven, Peking.
Emil Schulthess—Black Star

reconstruction and industrialization of the capital city; by 1959, there were almost 100 industrial units employing 1,000 to 5,000 workers each and 15 that employed more than 5,000.

Industrial activities. Among the large industrial establishments is the Shih-ching-shan Iron and Steel Works, located about 15 miles west of the walled city. The Shih-ching-shan plant was originally started in 1920, based on local ore and anthracite in the Western Hills; after the Japanese occupation of 1937, it produced a meagre amount of pig iron. Since 1958, the plant has been enlarged with a blast furnace, a steelworks, electric furnaces, and other equipment. With an annual capacity of 1,900,000 tons of iron and 1,500,000 tons of ingot steel, it is among the largest steel plants in China. Its high-quality steel production supplies such growing Peking industries as machine building, electrical engineering, and precision instrument manufacturing. A number of smaller finishing mills have also been established to produce such items as cold-drawn bearing steel and flat spring steel for tractor and automobile accessories, seamless tubes for high-pressure boilers, and magnetic steel for machine tools and radios. Peking is also an important centre of machinery manufacture, with plants established for manufacturing cranes, locomotives, automobile parts, tractors, wheat harvesters, and mining and printing machinery. Most of these plants were built in the suburbs east and south of the walled city, where extensive tracts of level land were available and where the prevailing northwest winds would carry industrial pollutants away from the densely populated areas. With the rapid expansion of cotton cultivation in Hopeh Province in the 1950s and 1960s, Peking has also become one of China's major textile centres. Peking's textile industry includes the manufacture of cotton and woollen fabrics and piece goods, serge, and several types of synthetic fabric, including polyester and nylon. To meet the needs of agriculture in the large rural area within Peking's administrative jurisdiction, insecticides, chemical fertilizers, and plastic sheets are also produced.

Iron and steel manufacturing

As a former Imperial capital, Peking possesses a rich accumulation of art treasures, produced by a variety of arts and handicraft industries that were intimately connected with court life and Imperial needs. Today, much of the traditional handicraft industry has not only been preserved but has been reorganized and re-equipped. Among the city's well-known handicraft products are rugs and carpets, porcelain and chinaware, jade and ivory sculpture, brass ware, enamel and lacquer ware, lace, and embroidery.

Commercial activities. With the rapid growth of industry and population in the city, many service industries, operated mainly by government agencies, have also greatly expanded.

Since 1949, there has been a slow but steady increase in tourist services provided in Peking. Many new hotels and hostels have been built, and old ones have been renovated and enlarged to meet the demands of the public; among these are the International Hotel, Hsin-ch'iao Hotel, Ho-p'ing Hotel, Ch'ian Men Hotel, Nationalities Hotel, the Overseas Chinese Hotel, Tung-ssu Hotel, and the enlarged Peking Hotel. Peking now has altogether more than 300 hostels and hotels; more than a dozen of them meet international standards. The China International Travel Service, which is state-owned, especially serves foreign tourists. Its general office is located on Hsi Ch'ang-an Avenue, and it has branches in all major cities in the nation, including one in Hong Kong. It makes reservations for airline and train tickets, hotel accommodations, and can also arrange for food, interpreters, cable service, and foreign monetary exchange. The general office of the Overseas Chinese Tourist Service is located in the inner city; its main function is to take care of overseas Chinese who are touring China, and it has more than 40 branch offices throughout the country.

Banking institutions

China has a nationwide and centralized system of banking, in which The People's Bank of China, a state-owned institution, plays a key role. This bank, with its head office in the inner city, functions as the agent of the national treasury. All funds of state-owned industrial enterprises, as well as of national, provincial, and local governments, the People's Liberation Army, and the network of cooperatives, are deposited with the People's Bank. The bank centralizes the financial resources of the government and the major sectors of China's economy. It uses the working capital at its disposal for the operations of the economy and the government. The People's Bank thus acts as cashier of the national budget and of a large part of the nation's financial operations. Through its branch offices and savings account centres, the bank also serves the daily needs of people in the city and in the suburbs. Working in cooperation with the People's Bank and under its supervision are two other older banking institutions—the Bank of Communications, which has the task of financing and supervising all basic construction projects, such as new railways, and the Bank of China, which specializes in international trade and foreign exchange.

The urban economy. Although Peking has been the capital of China for hundreds of years, it is far distant from the nation's traditional key economic area, the productive Yangtze Valley. The task of feeding the large urban population in Peking—not an easy one—was facilitated by the opening and maintaining of the Grand Canal and by the use of sea transport for importing grain from the south. Since the first decade of the 20th century, railways have played an important role in bringing food supplies to the capital; with the completion of additional

trunk lines since 1949, Peking, as mentioned, has become a major railway centre. Today, food grains and industrial raw materials from almost every corner of the nation can easily be shipped to Peking. The city, however, has striven for a high degree of self-sufficiency in secondary food supplies, such as vegetables, fruits, fish, and poultry, and in a number of construction materials. The expansion of municipal boundaries in 1958 and 1959 was partially aimed at this goal, and in the early 1970s the municipality was succeeding in maintaining an adequate supply of vegetables and fruits.

Vegetables are grown by the people's communes in a six-mile-wide belt, totalling more than 35,000 acres of land, that encircles the city and is covered by a network of irrigation channels. Some 600 miles of power lines provide power for the several thousand electric pumps that have made intensive farming possible. Hothouses, where vegetables are grown in the winter months, have 17,000 compartments, each occupying an average area of 130 square feet. This large farm belt is also one of the most highly mechanized agricultural areas in China.

In the 1950s and 1960s, many state farms near Peking have converted some of the hill slopes to the west of the city into large orchards, some of which are 1,700 acres in area. Several communes have also reclaimed large areas of wasteland to the north and east of the city for fruit tree planting. At present, the city's fruit-growing region ranges from the Great Wall in the north to the Grand Canal in the south and from the Yen Shan in the northeast to the Yung-ting Ho (Yung-ting River) in the southwest. While the hill areas produce large quantities of pears, persimmons, apples, chestnuts, and walnuts, the newly reclaimed lowlands are covered with vineyards and peach orchards. There are six state farms for fruit growing within the metropolitan area; of these, the Nan-kou State Farm, north of the city, is the largest. With more than 40 varieties of fruit grown in the metropolitan area, the ripening season lasts from early May to September. Most of Peking's fruit is channelled to the retail market through two distribution centres. Some of the fruit grown around Peking is sent to canneries, and some is put into cold storage; large quantities of "Aromatic Rose" grapes are sent to the Peking winery. Part of the Peking fruit crop is shipped by train to the cities of the Northeast and Inner Mongolia.

The business districts. Because of the symmetrical layout of the city, Peking lacks the compact central business district that characterizes most cities in the Western world. Historically, the market areas of the city were situated at two street intersections to the southwest and southeast of the centrally located Imperial Palace, and today the Tung Tan and Hsi Tan districts are still major shopping centres. Since 1949, however, a number of multistory department stores have been erected along Wang-fu-ching Avenue and at Hsi Tan, supplying a wide variety of goods for daily needs. The China Department Store, a state-owned enterprise, is the largest department store in China and has branch stores in major cities throughout the country. The International Friends (Quaker) Service Center serves foreign visitors and specializes in high-quality clothes, souvenirs, and gifts. In addition to these modern establishments, there are a dozen traditional bazaars that still serve an important local function. The Peking bazaars have evolved through a long period of history, and each has developed its own reputation for special commodities and services.

Shopping facilities

The Tung Tan Bazaar, located on the busy Wang-fu-ching Avenue, is the largest bazaar in the city. It includes about 30 different trades, which are carried on in more than 600 separate stalls. The bazaar is famous for the traditional handicrafts produced in the city, such as ivory, jade, and stone carvings, antiques and curios, lacquer ware, necklaces, cloisonné (a kind of enamel work) and porcelain, gauze lanterns, Peking sweets (dried and sugared fruits, such as sugared dates, sugared lotus, and sugared apples), and preserved fruits. The bazaar contains a number of restaurants and cafés serving a variety of Chinese regional foods.

The Liu-li-chang Bazaar is located just outside the city gate of Ho P'ing in the outer city. Because colourful glazed tiles were made here in the Ming dynasty, the area acquired its present name (which means "the glazier's shop"), but during the later part of the 18th century it gradually became a market for curios, antiques, old books, paintings, works of ancient Chinese calligraphers, and paper. It is still the centre for traditional art shops. The Ch'ien-men Administrative Department of the Peking Handicrafts Company has set up a workshop there that makes reproductions of traditional Chinese paintings, ancient bronze and iron vessels and statuettes, pottery and porcelain ware, and stone and wood carvings.

Inside the city gate of Yung Ting and near the T'ien T'an is the T'ien-ch'iao Bazaar. As early as the 16th century, poor people of the city began to come to T'ien-ch'iao to buy and sell secondhand goods. Open-air markets were also held here in the evenings. After 1900, when the railway station in the southern suburbs was moved to the vicinity of the T'ien T'an, peddlers and small traders, folk artists, storytellers, and musicians began to appear in large numbers, and the place became a bustling market for fish, poultry, fruit, and vegetables, as well as a recreational area for poor people. This bazaar has some of the best food shops in the city and remains a folk-entertainment centre for the metropolis.

Paolo Koch—Rapho Guillumette



Stalls in Peking's main market.

The Hsi Tan Bazaar, established in 1932 around Hsi-tan-pei-ta Street in the western section of the city, is one of the most diverse markets in the city. The bazaar is composed of five different sections, each offering special commodities or services. The commodities include sports equipment, musical instruments, books, magazines, food supplies, hardware, and electric appliances; there is also a complete department store, as well as a number of recreational establishments.

Political and governmental institutions. As the national capital, Peking houses the most important political institutions in the nation. These include the National People's Congress, the supreme organ of state power, and the State Council, the highest executive organ of the state, which carries out the decisions of the congress. The State Council has under its jurisdiction administrative depart-

The State Council

ments, the number of which varies according to the needs of the work. They include the ministries and commissions in charge of foreign affairs, internal affairs, public security, national defense, justice, finance, culture, health, education, nationality affairs (concerning minority groups in the country), agriculture, and various branches in industry.

In addition to the above, there are some organs of state concerned with the maintenance of law and order—the People's Courts and the People's Procuratorate, both of which are located in Peking.

Peking is one of the three independent municipalities in China (the others being Shanghai and Tientsin), and there is no governmental tier between it and the central government. Peking is divided administratively into eight municipal districts in the centre and nine annexed counties in the peripheral areas. The highest organ of the municipal administration is the People's Congress of Peking Municipality. The representatives for the congress, numbering 564, are elected among groups of workers, peasants, teachers, merchants, office workers, the armed forces, public institutions, and religious circles. The congress elects the members of the Peking People's Council, the executive organ of the municipal government, including the mayor and deputy mayors. It ensures that the municipal government carries out the decisions and resolutions of the congress, discusses and approves the budget and financial statements submitted by the municipal government, and examines the report on the administrative work of the municipality. The composition of the congress in 1955 is shown in Table 2.

Table 2: Composition of the People's Congress of Peking Municipality

groups	number of deputies
Workers	168
Employees in the education and cultural fields	136
Government employees	65
Industrialists and businessmen	43
Democratic parties	30
Peasants	27
Medical and health workers	26
Engineers and technicians	26
Co-op employees	12
Dependents of veterans	12
Armed forces stationed in Peking	10
Religious groups	9
Total	564

The congress included a percentage of women, minority groups, and model workers elected by various industrial plants.

The Peking People's Council in 1955 consisted of 47 members, including the mayor and his deputy mayors. They control 47 administrative units that cover all fields of life in the city—health and sanitation, public security, education, employment, industry, etc. As a subdivision of the municipality, each district or township has its own people's council. Below the urban district are a number of street offices of the district people's council. Below street offices are residents' committees, which are self-governing welfare organizations established on a neighbourhood basis. The residents cooperate on matters of common interest, such as problems of local hygiene, culture and recreation, sports, studies in current events, and literary classes. The residents' committee is also in charge of the neighbourhood's population census, vital statistics, population movements, children's nurseries, and street beautification.

Public utilities. Peking is not on a river, and all water for municipal consumption has to be brought from elsewhere. Some water, however, comes from shallow wells, which are common throughout much of the region. These provide some villages with drinking water and supply water for irrigation.

Water supplies

From the earliest history of the city, use has been made of the rivers and springs in the neighbourhood. To the east is the Pai Ho, which, after being joined by the Sha Ho, turns southward. The tumultuous Yung-ting Ho is a large river, draining the Shansi Plateau and the northwest of Hopeh. After following a twisting course through the mountains, it reaches the Peking plain, passes under the Marco Polo Bridge, nine miles southwest of Peking, and then turns southward to meet the Pai Ho north of Tientsin. The flow of the Yung-ting Ho is irregular; in the rainy season it rises rapidly, carrying with it large quantities of silt, which raise the level of the riverbed considerably. At Marco Polo Bridge it is 50 feet above the level of the city, thus constituting a hazard when the river is in flood but facilitating canalization and irrigation. Several springs rise at the foot of Hsiang Shan (Fragrant Hills) and on the Yü-ch'üan Shan (Jade Fountain Hill), both to the northeast of the city. During the Manchu dynasty these springs were tapped by means of an aqueduct that conveyed water for the city moat and for the three lakes near the Imperial Palaces. The moat around the city walls became an important means of water distribution for the municipality.

In the early 1950s, large-scale water-conservancy projects were begun to provide more water for the expanding urban area. Two large artificial lakes have since been made: the Mi-yün Reservoir, northeast of the city, and the Kuan-t'ing Reservoir, in the northwestern mountains beyond the Great Wall. These regulate the flow of the Pai Ho and of the Yung-ting Ho upstream, storing water at times of heavy discharge and releasing it when the rivers are low. Two lesser projects have also been carried out: the construction of the Ming Tombs Reservoir, whose waters feed a hydroelectric power station and irrigate the neighbouring countryside, and a hydroelectric power station near Mo-shih-k'ou, which uses the waters of the Yung-ting Ho and feeds them back into the river through an ancient canal. The hydroelectric station at Kuan-t'ing, which has been generating electricity since 1956, is the largest source of electricity for the metropolitan area of Peking.

The Peking Water Company, which was originally formed in 1908, is responsible for supplying water for the municipality. Water is obtained from the old waterworks within the walled city, as well as from several waterworks in the suburbs; altogether, the network of mains and pipes totals nearly 1,250 miles in length and extends to every newly built residential and industrial unit, including towns and villages near the Men-t'ou-kou coal mine about seven miles from the city. Public standpipes have also been installed at regular intervals throughout the city, the suburbs, and nearby rural people's communes. The Kuan-t'ing Reservoir, the largest man-made lake in the North China Plain, supplies the metropolitan area of Peking with most of its water.

Peking is one of the few cities in China equipped with a modern sewage system. An underground sewage system was installed as early as the 15th century, when nearly 195 miles of sewage network were constructed. Later this network became clogged with effluent and silt. A new sewage system was installed in the early 1950s, which, within the walled area, was partly based on the rehabilitation of the old system. All open sewers, characteristic of many Chinese cities, were eliminated, and nearly 600 miles of new pipes were laid throughout the densely populated areas of the municipality.

The headquarters of the Peking General Post Office is located on Kung-an Street, on the east side of T'ien An Men Square. It provides more comprehensive services than the post office in any Western city, handling mail, telegrams, long-distance telephone calls, and the distribution of newspapers and magazines. There are also more than 350 branch offices and stations in the city and the suburbs. The Peking General Post Office has also set up several service centres at certain busy traffic points, such as Hsi Ch'ang-an Avenue, Ch'ien Men, Peking Railway Station, T'ien An Men, Hsi-chi, and Wang-fu-ching. The service centre at Wang-fu-ching takes subscriptions to foreign newspapers and magazines and sells single issues.

The Peking Telegraph Office is housed in the Telegraph Building on Hsi Ch'ang-an Avenue. It is open day and night for domestic and international telegrams and for

long-distance telephone calls. International telegrams can be sent to all countries throughout the world in four languages: Chinese (in Chinese standard code, in which each character is represented by four numerals), Russian (in Roman alphabet), English, and French; telegrams may also be sent in any other language agreed upon by the countries to which the telegrams are sent, provided they are written in the Roman alphabet and notice is given of the language to be used.

The telephone system

Telephones first appeared in Peking in 1900, when a small exchange was established by a foreigner. In the subsequent 50 years, more than 20,000 telephones were installed. In the 1950s the Peking Telephone Administration increased its 21 exchanges to 29, and nearly all telephones in the city became automatic. Seven of the new exchanges were installed in suburbs, serving outlying districts that have experienced rapid industrial and agricultural development. In the 1960s the original five-digit dialing system was changed to one of six digits, thus increasing the city's telephone capacity from 100,000 to 1,000,000 instruments. Public telephones are located in shops in busy streets. As a great number of households in Peking cannot be reached by phone, a neighbourhood messenger service is provided in many of these street service stations. The caller gives the name and address of the person he wants to reach, leaving his own number and the time he wants to be called with the attendant who delivers the message; attendants are usually elderly men or housewives.

Health. Chinese medical services employ two systems of medicine: traditional Chinese and Western. Peking possesses some of the best hospitals of both systems in the nation.

The Peking Union Hospital, the earliest modern hospital to be established in China, was founded in 1921 as an affiliate of the Peking Union Medical College, which from 1915 onward was supported by the China Medical Board of the Rockefeller Foundation. The hospital, now renamed Peking Hospital, is a large polyclinic combined with an institute of gynecology and pediatrics. It handles about 500,000 cases annually and is the largest hospital in Peking. Since 1949, more than 30 new hospitals, clinics, and sanitariums, together covering more than 10,000,000 square feet of floor space, have been built, and there are more than 25,000 beds available in the inner districts of the metropolitan area. Among the newly established hospitals is the Friendship Hospital, located at T'ien-ch'iao in the outer city, a gift made by the Soviet people at the time of the peak of Sino-Soviet friendship in the 1950s. Until 1960, the Soviet staff not only treated patients but also gave advanced training to Chinese medical personnel according to the latest Soviet methods.

Many hospitals are clinical teaching hospitals attached to medical schools; these include the Fu Wai Hospital of the Chinese Academy of Medical Science, located outside the city gate of Fu Ch'eng; the First Affiliated Hospital of the Peking Medical College, near the city gate of Hsi An; the Affiliated Hospital of the Research Institute of Chinese Medicine, located in the western suburban town of Hsiao-yüan; and the Affiliated Hospital of the Peking Chinese Medical College, situated on Nanhsiao Street. The largest children's hospital in China, the new Peking Children's Hospital, is situated just outside the city gate of Fu Hsing. It has 600 beds and was designed to provide a homelike atmosphere for the sick child. In the hills to the west, a sanitarium in the traditional Chinese architectural style has been built for students from Asian countries; it has accommodations for up to 300 students. The most specialized hospitals are located in the old residential districts of the city and include the Stomatological Hospital (dealing with the mouth and related diseases), the Chest Surgery Hospital, the Hospital of Plastic Surgery, the old Peking Children's Hospital, the Peking Municipal Hospital of Chinese Medicine, and the Ji-sui-tan Hospital, which specializes in traumatic (accident-related) and orthopedic surgery.

Traditional medicine

Chinese traditional medicine comprises the prescription of Chinese herb drugs and the practice of acupuncture (a medical practice encompassing therapeutics and anesthesia that involves puncturing various parts of the body with needles). Doctors in both Chinese and Western systems are trained in many universities and institutes in Peking. Students in the Western-style medical schools are taught the essentials of Chinese traditional medicine, and Chinese traditional medical students are expected to be familiar with the essentials of Western medicine. Diagnoses of difficult cases are made only after consultations with specialists of both systems.

For the rural areas of the metropolis, each county has a well-equipped county hospital located in the county seat. But the basic unit of the medical service is the health centre in the commune, which is equipped with complete clinical facilities. Medical teams of doctors, nurses, and public health personnel are sent frequently from the teaching centres in the city to the communes to assist in the general medical work and to keep the local medical staff in touch with the latest medical progress and techniques.

Education. There has been no radical change in the school system in Peking since 1949. The school system is still largely composed of six years of primary education, which is quite universal, and six years of secondary education. For preschool-age children, kindergartens and nurseries are operated by factories, business enterprises, government offices, and city block cooperatives. Their main function is to permit the mother to work so that, while some of these facilities are located in the residential neighbourhoods, many are located on the premises of particular institutions of employment. In the suburban or rural areas, kindergartens are often temporarily set up during such events as the harvest; small fees are charged that help to offset the cost of meals provided.

The schools

In the early years of the present regime, efforts were made to abolish private schools and to put all education in the hands of the state. As a result, all private schools run by missionaries were taken over by the government. It was soon realized, however, that the government could not possibly provide enough schools for the millions who demanded education. The government then reversed its policy to some degree and decided to encourage the establishment of schools by local organizations, such as factories, business concerns, collectives, and communes. In the early 1970s a major part of the finance for elementary education was met by sources other than the central government, though government subsidies to schools in difficult financial situations were also considerable.

The curriculum in the primary school is simple. The early part of primary education is largely devoted to reading, writing, and mathematics, and in the last two years of the primary school, nature study, history, geography, and physical education are added to the curriculum. Closely related to primary education are movements for the liquidation of illiteracy, the phoneticization of the Chinese writing, and simplification of Chinese characters.

The secondary schools in Peking are mainly of three types: the general middle school, the normal school, and the vocational and technical school. The general middle school is of the academic type, with a curriculum designed to prepare students for college. The normal school is to train teachers for primary schools. The vocational and technical schools, created to provide skilled workers in various fields, have developed most rapidly in the city and in the early 1970s were beginning to overtake the general middle schools in number.

Institutions of higher education

Peking's position as a centre of higher education in China has been further strengthened since 1949, and within the metropolitan area in the early 1970s there were more than 20 institutes of higher learning. A scientific and educational district has emerged in the northwestern outskirts of the city, against the background of the K'un-ming Hu (K'un-ming Lake), the Summer Palace, and the Western Hills. This area contains such prestigious institutes of higher education as Peking Univerity and Ch'ing-hua University. Founded in 1898, Peking University is one of the largest composite universities in China. The university has a faculty of 2,000 and



May Day rally in T'ien An Men Square, Peking.
Der Stern—Black Star

an enrollment of more than 10,000 students, including some 300 foreign students from 30 different countries. In 1953 the university moved from its old site at Sha-t'an, in the inner city, to the present campus, which previously belonged to the missionary-established Yen-ch'ing University. The new campus has been expanded considerably to the east and south. Ch'ing-hua University has an enrollment of 6,000, with 45 percent of the students being selected from among workers in various factories throughout the country, 40 percent from communes, and 15 percent from the People's Liberation Army.

Within the academic district are also found the People's University of China, an institute of ideological training, and the Central Institute of Nationalities which accepts students from the various autonomous regions and districts. Situated in the same area are the Peking Normal College, the Peking Medical College, and the Central Conservatory of Music. Numerous institutes of higher education are also located in this area; these include the Peking Institute of Steel and Iron, the Peking Institute of Aeronautical Engineering, the Peking Petroleum Institute, the Peking Institute of Geological Prospecting, the Peking Institute of Forestry, and the Peking Mechanical Institute of Agriculture. The Chinese Academy of Sciences, the highest research institute in the nation, is also located in this district.

Foreign-language institutions

Besides technical colleges, there are a number of foreign-language institutions in Peking where foreigners are employed to teach alongside their Chinese colleagues. The First Foreign Language Institute is the largest institute of its kind in China and has an enrollment of 2,000. The Second Foreign Language Institute was set up in 1964 for training personnel for the New China News Agency. There are also a number of small institutes of language training, run by the Ministries of Foreign Affairs and Foreign Trade and by the Peking Radio Station. The Peking Institute of Foreign Languages is an institute at the secondary level, which accepts children as young as eight years of age and enrolls about 1,000 students; 500 of them learn English, while others study French, German, Spanish, Arabic, Russian, and Japanese.

Cultural life. As the cultural focus of China for several centuries, Peking possesses what are probably the finest cultural institutions in the nation, including libraries, museums, and theatres.

To the east of T'ien An Men Square within the People's Cultural Park is found the Working People's Palace of Culture (formerly the Temple of the Imperial Ancestors), where the tablets of the emperors were displayed. The temple, like the T'ai Ho Tien (Hall of Supreme Harmony) and the Imperial Palaces in style, was built in three stone-work tiers, each with double eaves. On either side are two rows of verandas surrounding a vast courtyard large enough to hold 10,000 people. Exhibitions of economic and cultural achievements, both of China and of foreign countries, are frequently held in the three halls. Each exhibition lasts several weeks and draws hundreds of thousands of visitors. Lectures by leading scholars on science, literature, and the arts are frequently held there, and reports are presented by nationally known heroes of labour. There is also an art school for adults in the park; it has five departments—for literature, drama, music, fine arts, and dance. The park also serves as the home of amateur art groups, who specialize in dancing, orchestral music, or choral singing. Persons who are found to have particular artistic talent may be sent from here to a specialized school for further study. In the northeast corner is a sports ground that can accommodate more than 4,000 spectators. In a grove in the southeast corner is a large open space for dancing.

The Palace Museum

The Imperial Palace Museum is housed in the main buildings of the former Imperial Palaces. Many of the halls are kept as they were in dynastic times, each constituting a museum in itself; others are kept for the display

of some of the priceless treasures from China's past. The Imperial Palaces, with their golden roofs, white marble balustrades, and red pillars, stand in the heart of Peking and are surrounded by a moat and walls with a tower on each of the four corners.

The palaces consist of outer throne halls and an inner court. North of the three tunnel gates that form the front gate, called the Wu Men (Meridian Gate), a great courtyard lies beyond five marble bridges. Further north is a massive double-tiered hall, the T'ai Ho Tien (Hall of Supreme Harmony), once the throne hall. A marble terrace rises above the marble balustrades that surround it, upon which stand beautiful, ancient bronzes in the shapes of cauldrons, cranes, turtles, compasses, and ancient measuring instruments. The T'ai Ho Tien is the largest wooden structure in China.

Behind it, beyond another courtyard, is the Chung Ho Tien (Hall of Complete Harmony), where the emperor paused to rest before going into the throne hall (the Hall of Supreme Harmony). Beyond the Chung Ho Tien is the last hall, the Pao Ho Tien (Hall of Preserving Harmony), after which comes the Nei T'ing (Inner Court). The Inner Court was used as the emperor's personal apartment. It contains three large halls, the Ch'ien Ch'ing Kung (Palace of Heavenly Purity), the Chiao T'ai Tien (Palace of Heavenly and Earthly Intercourse), and the K'un Ning Kung (Palace of Earthly Tranquillity).

The Ch'ien Ch'ing Kung is divided into three parts. The central part was used for family feasts and family audiences, audiences for foreign envoys, and funeral services; the eastern section was used for mourning rites and the west section for business of state. The other two palaces, one behind the other, were Imperial family residences. The three throne halls in the Outer Court and the three main halls in the Inner Court lie along the central axis. On either side are smaller palaces, with their own courtyards and auxiliary buildings. Behind the buildings, before the north gate of the Imperial Palaces is reached, lies the Imperial Garden. Each palace, its courtyard and side halls, forms an architectural whole.

The Museum of Chinese History and the Revolution, located on the eastern side of T'ien An Men Square, is a new museum completed in 1960. There are more than 8,000 antiquities on display, of which the greater part have been excavated only during recent years. These are arranged to give a comprehensive survey of the progress of human achievement in China, beginning with the appearance of the prehistoric Peking man some 500,000 years ago, then continuing through the last 6,000 years of Chinese history. The exhibition has three main sections, devoted to primitive society, slave society, and feudal society.

The Museum of the Chinese Revolution occupies the left wing of the museum building on the eastern side of T'ien An Men Square. Its aim is to trace the history of China since the middle of the 19th century, laying particular stress on the revolutionary movement and the part played by the Communist Party.

The Peking Planetarium, the first of its kind in China, is located outside one of the city's western gates. Completed in 1957, it is an educational centre for the dissemination of general knowledge in astronomy and natural science by means of exhibitions, forums, films, projector demonstrations, and telescopic observations of the heavens. The planetarium proper, the "Theatre of Heavens," is situated on the south side of the building and has a seating capacity of 600. Behind the main building stand an observatory and meteorological station.

The Peking Library

The Peking Library, with a collection of 7,000,000 volumes, is the largest public library in China and is located on the west bank of Pei Hai. The library has inherited books and archives from the Imperial Wen Yüan Ko library, a famous collection that has existed for more than 500 years and that, in turn, included books and manuscripts from the Imperial Library of the Southern Sung dynasty, established some 700 years ago. It also includes other collections from Imperial libraries of the Manchu dynasty, Imperial colleges, and private owners. Among them are rare copies of ancient manuscripts and books of five dynastic periods from the Sung to the Ch'ing, including 140,000 manuscript volumes on 18,000 different subjects, 8,700 copies of Buddhist sutras going back to the 6th century, old maps, diagrams, and 25,000 rubbings from ancient inscriptions on metal and stone; it also possesses the *Yung-lo ta-tien* ("Yung-lo Encyclopedia") of the Ming dynasty and a copy of the *Ssu-ku ch'üan-shu* (from the Imperial Library of Ch'ien-lung), dating from the Manchu dynasty.

The media. Peking is the chief publishing centre for newspapers, periodicals, and books in China. Almost all publishing enterprises are owned by the government. New China News Agency in Peking is the chief provider of news releases in English. There are almost 250 periodicals with national circulations published in Peking. Several foreign-language magazines, including *People's China, China Pictorial, China Reconstructs, Chinese Literature,* and *Women of China,* are also published in the capital. More than 20 national newspapers are published there. "People's Daily" (*Jen-min jih-pao*), the official organ of the Central Committee of the Chinese Communist Party, is the most influential paper in the country and reflects the ideology and policies of the party in both foreign and domestic affairs. *Kuang-ming jih-pao,* jointly published by the various minor democratic parties, the All-China Federation of Industry and Commerce, and nonparty members of the Chinese People's Political Consultative Conferences, gives special attention to culture, education, science, politics, law, and the life of the national minorities. *Ta-kung pao* emphasizes news and commentaries on commerce, finance, and international affairs. "Worker's Daily" is the official organ of the All-China Federation of Trade Unions. "China Youth News," the official organ of the Central Committee of the Communist Youth League of China, is a political and educational media for teen-agers. "Chinese Children" is an official paper for the Chinese Young Pioneers, published also by the Central Committee of the Communist Youth League of China.

There are more than 40 publishing houses in Peking. The largest is the People's Publishing House; others deal in specialized fields, such as literature, fine arts, science, and national minorities. The Foreign Language Press is responsible for introducing China and the Chinese people to the world through books, pamphlets, pictorials, and translations of literary works; most are in English.

The most important radio station in China, the Central Chinese Broadcasting Station, is also located in Peking. Using more than a dozen different frequencies and broadcasting three sets of programs simultaneously, the central station is on the air for more than 47 hours a day. Three sets of programs are coordinated and supplement each other; the first gives news and news commentaries, while the second and third broadcast different literary or artistic items. Programs reach the city's population not only through radios but also through loudspeakers and radio-listening centres that are installed in almost every public gathering place, including railroad stations, public dining halls, playgrounds, and parks.

Television services

The first television station in Peking was established in 1959 by the Peking Broadcasting System. The city's television programming was still on a rather limited scale in the early 1970s, as the number of sets was not very large. Most television sets were for public use and are found in parks with benches provided for viewers, as well as in railroad stations, hotels, and other public places. Daily television programs include those devoted to popular science, sports, opera and theater, and the entertainment of children. The Peking Television College, offering preparatory courses in mathematics, physics, chemistry, and other subjects, was instituted in 1960 under the joint auspices of the city department of education, the Peking Broadcasting System, and three universities in the area. Students of the television college, numbering several thousands, are scattered in more than 800 institutions and organizations throughout the city; most of them work in factories, mines, and research organizations or are teachers in primary and secondary schools. In every district of the city the college also set up an advisory centre where students may go for additional tutoring.

Recreation. As the residence of the Imperial families through several dynastic periods, Peking is well-known for its numerous parks and playgrounds; few cities in China have as large a proportion of land within the city walls allocated for recreational uses. Among the most popular of Peking's parks are Chung Shan Park, Pei Hai Park, Coal Hill Park (Mei Shan Park), the Summer Palace, and the Peking Zoological Gardens.

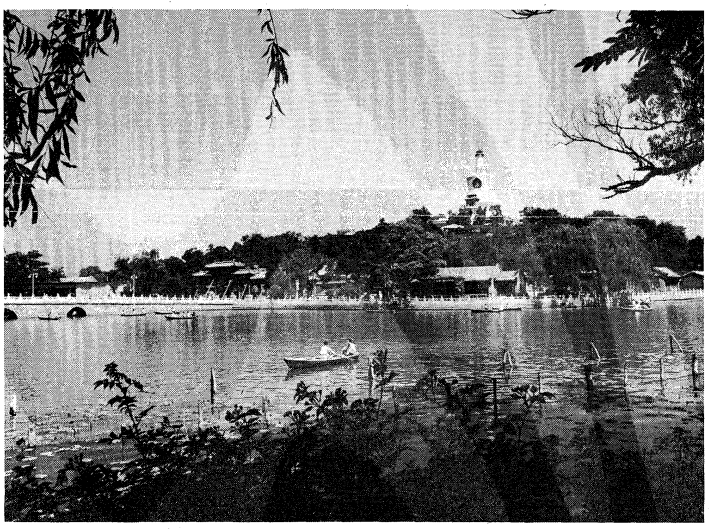
Chung Shan Park adjoins the west wall of T'ien An Men; it is the most centrally located park in Peking and encloses the former Altar of Earth and Harvests, where the emperors made offerings to the gods of Earth and agriculture. The Altar of Earth and Harvests consists of a square terrace in the centre of the park, to the north of which is the 550-year-old Hall of Worship, now the Sun Yat-sen Memorial Hall, which is used for public meetings. The Hall of Worship is the oldest wooden structure in Peking; its simple form, masterly design, and sturdy woodwork bear the characteristic marks of early Ming architecture.

The Water Pavilion, built out over a lotus pond on three sides to provide a gathering place for scholars and poets, is also located in the park. Scattered among the park's pools, goldfish enclosures, rocky hills, weeping willows, pines, cypresses, bamboos, and flowers are pavilions, kiosks, and towers, which are typical of Chinese garden landscape.

Pei Hai Park

Pei Hai Park lies to the northwest of the Forbidden City. It covers some 176 acres, half of which is water. It contains the most northerly of the three lakes that lie roughly north and south within the Imperial City. Pleasure grounds, lakes, and buildings have existed on this site for 800 years. As the lakes were deepened and dredged, the excavated earth was used to build hillocks and islands of great beauty. In 1651, a Manchu emperor built the White Pagoda, the most striking landmark in the park, on the top of a hill. Many of the buildings in the park are used for serious pursuits, as well as recreation. They include an Institute of Research on the History of Chinese Classical Literature and a Natural History Museum, as well as a branch of the National Library. The lake in Pei Hai Park is crowded with rowboats in summer and freezes over to become a natural skating rink in winter.

Paolo Koch—Rapho Guillumette



Pei Hai Park in central Peking. In the background is the White Pagoda, which dates from the Yüan dynasty.

Coal Hill Park, also known as Mei Shan Park, is a man-made hill, more than a mile in circumference, located north of the Forbidden City. The hill, from the top of which a spectacular panorama of Peking can be seen, has five ridges, with a pavilion on each. The hill was the scene of a historical tragedy when, in 1644, at the end of the Ming dynasty, the defeated Ming emperor hanged himself on a locust tree on its east slope; the tree is still standing and may be seen there to this day, located amid ancient pines and cypresses. In the northern part of Coal Hill Park is Peking Children's Palace, with recreational, athletic, and educational facilities.

The Summer Palace

The Summer Palace, lying close to the Western Hills, is about six miles northwest of Hsi Chih Men, the nearest city gate. It is the largest park on the outskirts of Peking and is famous for its artificial landscaping, which provides an inimitable blend of woods, water, hills, and architecture. The Summer Palace covers more than 800 acres, four-fifths of which consists of lake, the remainder being man-made hillocks. More than 100 buildings—halls, towers, pavilions, bridges, and pagodas—lie scattered throughout the park; a marble boat, two stories high and some 80 feet long, located at the northwest corner of the lake, is one of its attractions. Connecting the buildings and courts along the shore of the lake are a series of richly painted covered promenades.

Peking Zoological Gardens is located on the western outskirts of the city, covering an area of 175 acres. The zoo was established toward the end of the 19th century and was named the Garden of Ten Thousand Animals. It is the largest zoo in the country with more than 3,000 animals from all parts of China and the world residing in captivity there.

Peking Workers' Gymnasium, which seats 15,000, is the largest and most complete athletic facility in China. Higher than a conventional eight-story building, it presents a facade of wide, gleaming windows set in light-green walls. Spectators may enter through gates on three of its sides, the north, east, and west sides; the south side is rimmed with practice rooms. The facility is used for basketball, volleyball, badminton, table tennis, gymnastics, fencing, weight lifting, as well as other sports. It is also used for large concerts and mass acrobatic performances.

For outdoor sports, the park located on the bank of the Shih Ch'a Hai (Lake of Ten Monasteries) is the largest open field in Peking. The park includes four indoor and outdoor swimming pools, a sunbathing ground, and an artificial lake for rowing.

The Peking Workers' Stadium, where soccer games and field and track events are held, is a large compound located to the east of the Imperial City, just outside the wall.

BIBLIOGRAPHY. HIKOTARO ANDO, *Peking* (1968), is the best illustrated recent description of the city by a foreigner. *Nagel's Encyclopaedia-Guide to China* (1968) includes a lengthy section on historical features and various localities of the metropolitan area. HU CHIA, *Peking, Today and Yesterday* (1956) and *Peking: A Tourist Guide* (1960), published by the official press in Peking, provide useful readings for visitors. The historical development of the city is covered both by NIGEL CAMERON and BRIAN BRAKE in *Peking: A Tale of Three Cities* (1965); and by E. BRETSCHNEIDER in *Die Pekinger Ebene und das benachbarte Gebirgsland* (1876; Eng. trans., *Archaeological and Historical Researches on Peking and Its Environs*, 1876). LIN YU-TANG, *Imperial Peking* (1961), deals with the cultural aspects of the city, with emphasis on arts and daily life in historical times. For a detailed account of architectural features of the walled city, see OSVALD SIREN, *The Walls and Gates of Peking* (1924) and *The Imperial Palaces of Peking* (1926). FRANK DORN, *The Forbidden City* (1970), vividly associates the details of the Imperial palaces with historical events. A number of important works on scenic and architectural aspects of the city were completed in the 1930s; among them are L.C. ARLINGTON and WILLIAM LEWISOHN, *In Search of Old Peking* (1935); and JULIET BREDON, *Peking: A Historical and Intimate Description of Its Chief Places of Interest*, 2nd ed. (1931). An account of living experience in the city in the 1930s is found in GEORGE N. KATES, *The Years That Were Fat* (1952). For a geographical analysis of the strategic location of the city, see SEN-DOU CHANG, "Peking: The Growing Metropolis of Communist China," *Geogrl. Rev.*, 55:313–327 (1965); GUSTAV STRATIL-SAUER, "Zur Stadtgeographie von Peking," *Petermanns geogr. Mitt.*, 103:180–189 (1959); and SIR REGINALD JOHNSON, "Peking," *Geogrl. Mag.*, 1:185–197 (1935). The urban growth and municipal development of the city in the past two decades are described in various articles that appeared in *China Reconstructs*, a monthly periodical published in Peking. *Carte Touristique de Pekin* (1957) is a coloured map of the city for Western visitors, and includes streets, scenic spots, bus lines, tramways, and the layout of the palaces and important parks.

(S.-d.C.)

Pelecaniformes

The order Pelecaniformes comprises a number of diverse groups of aquatic birds that share the common characteristic of webbing between all four toes. There are three main groups (suborders): Phaethontes, the tropic birds; Fregatae, the frigate birds; and Pelecani, the pelicans, boobies (including the gannet), cormorants, and anhingas (or darters). All are relatively large birds: they range in length from about 40 centimetres or 16 inches (excluding the elongated central tail feathers) in the white-tailed tropic bird (*Phaethon lepturus*) to 1.8 metres (6 feet) in the Dalmatian pelican (*Pelecanus crispus*).

GENERAL FEATURES

In terms of their way of life, the pelecaniform birds fall into four adaptive groups: frigate birds—long-winged masters of piracy and aerial pursuit of surface-living marine prey; tropic birds and boobies—wide-ranging flyers that capture prey underwater by plunging from a height; pelicans—large, large-billed, long-necked, buoyant birds that fish mostly by reaching down while swimming at the surface; cormorants and anhingas—heavy-bodied, long-necked, underwater swimmers, respectively pursuing and lying in wait for their prey below the surface. The anhingas are largely confined to freshwater; pelicans and cormorants occur in both freshwater and marine habitats; and the other groups are entirely marine.

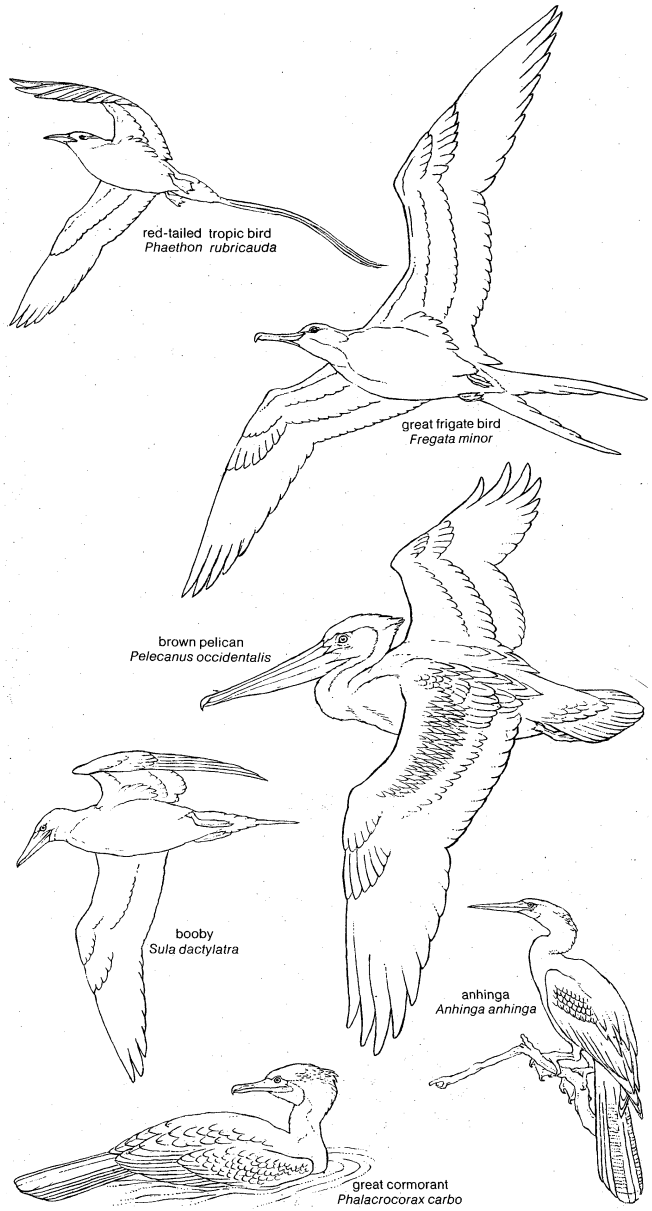
Production of guano as fertilizer

Importance to man. Most pelecaniform birds are of rather little significance to man, but the guano (excrement) of cormorants, boobies, and pelicans is an important fertilizer. Exploitation of old accumulations of guano reached its peak in the mid-19th century, and since then only the current production of guano has been available in most areas, but even this provides a substantial resource where the bird populations are large. There were an estimated 18,000,000 guano birds on the coast of Peru early in the 1960s. Of these, about 15,000,000 were guanay cormorants (*Phalacrocorax bougainvillii*), and the remainder were Peruvian boobies (*Sula variegata*) and brown pelicans (*Pelecanus occidentalis*). The harvest of guano at that time amounted to about 180,000 tons per year. The guano birds in the area feed largely on the Peruvian anchovy, and now that this fish is directly exploited on a large scale there is some competition between the birds and the fishing industry. In Southwest Africa artificial platforms have been constructed in coastal lagoons and on an offshore reef, greatly facilitating the collection of guano.

Pelicans, feeding on fish from inland and coastal waters, are among the animals whose diet tends to ensure that they will accumulate residues of insecticides (especially DDT) in their bodies. It recently has been found that among the physiological effects of these substances on birds are changes in calcium metabolism that result in their laying eggs with abnormally thin shells, or no shells at all, with the result that these eggs usually break before hatching. Effects of this kind have entirely prevented the successful reproduction of brown pelicans on the California coast during recent years, and some populations elsewhere are also showing adverse effects.

Distribution and abundance. Pelecaniform birds occur all over the world except in parts of the interior of North America, Africa, Asia, and Australia, in the high Arctic, and in most of Antarctica. Two groups, the tropic birds and frigate birds, are essentially confined to the tropics; the boobies, pelicans, and anhingas are widespread in the tropics but also penetrate far into the temperate zones; and cormorants breed from the Equator to the Arctic Circle and the Antarctic Peninsula.

The tropic birds are the most pelagic (free-flying over open ocean) of the pelecaniforms and can be seen even in the most unproductive central parts of tropical seas. Some boobies and frigate birds are found far out at sea, but most of them return to land to roost at night. Pelicans, cormorants, and anhingas do not venture far from land but may commute some distance between roosting or breeding places and their feeding grounds.



Representative Pelecaniformes.
Drawing by R. Keane

NATURAL HISTORY

Breeding grounds

Reproduction. The breeding of pelecaniform birds is essentially restricted to places free of mammalian predators. Tropic birds, boobies, and frigate birds typically breed on oceanic islands or on small islets and stacks off continental coasts. When man has introduced predators (such as cats) to isolated islands, pelecaniforms and other marine birds have often been eliminated from their traditional breeding grounds. On Ascension Island and St. Helena in the South Atlantic, for example, tropic birds, frigate birds, and boobies, which used to breed in large numbers on the main islands, have either been exterminated or driven to nesting on small offshore islets. Pelicans and cormorants are not normally found on islands far from continental land but breed on islands in lakes or offshore, or in other protected sites such as trees standing in water or cliffs. Anhingas breed in trees or bushes close to the sheltered waters where they feed.

Pelecaniform breeding, whether on the ground or in trees, is typically colonial, apparently because of the scarcity of safe places. If they are to breed at all, all the birds of an area must crowd into the available space. In practice, most colonies have a fairly definite upper limit of density determined by the distance that an incubating

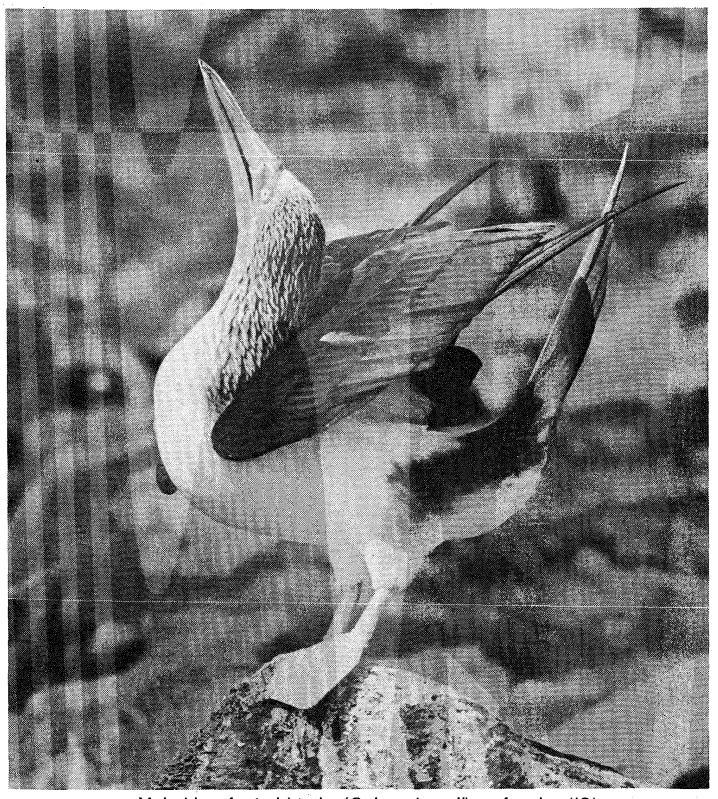
bird can reach out to repel intruders. Nests of the gannet (*Morus bassanus*) have an average density of about one per square metre; those of the guanay cormorant average about 3.5 per square metre.

In some species, colonial breeding has become obligatory, and single pairs or small groups do not breed successfully. Other species breed colonially only where there is a shortage of space for nesting. The masked booby (*Sula dactylatra*), for example, breeds in dense colonies on islets off Ascension Island but in dispersed patterns on Christmas Island (Pacific). Breeding in a number of species is normally dispersed; the red-footed cormorant (*Phalacrocorax gaimardi*) of South America, for instance, often nests on cliffs where only scattered sites are available. Similarly, tropic birds nest in holes or crevices in cliffs or under the shelter of rocks or bushes. The sites are thus normally well dispersed, and in some places there is intense competition for them.

Pair formation displays

Courtship behaviour. During courtship (except in tropic birds) the male typically selects a nest site and then displays from this site to passing females. In the male advertising display of boobies ("Sky-pointing") and of cormorants and anhingas ("Wing-waving"), the head and tail are raised and the wings partially raised. The boobies have species-specific positions of the wings and tail; the cormorants and anhingas wave either their wing tips or their whole wings, to a varying degree. In many species a particular call is associated with the display. In the "Rattling" display of frigate birds, the male leans back on his tail, extends his wings along the ground, and with his bill pointing vertically throbs his inflated scarlet throat (gular) sac, vibrates his wings, and claps his bill.

Eric Hosking



Male blue-footed booby (*Sula nebouxii*) performing "Sky-pointing," a display effected in pair formation.

In most pelecaniforms, when a female has finally joined a male at the nest site, he may go and fetch nest material, which is then added to the nest structure by the female, alone or with the assistance of the male. There is no courtship feeding in the pelecaniforms, but in tropic birds there are a few records of one adult feeding the other during incubation. Copulation normally occurs on the nest site, without any special precopulatory or postcopulatory displays. The pair-bond, once established, is reinforced by such activities as joint nest building and defense, and preening of one bird by the other. Certain displays functioning in individual recognition are performed in the air or on the ground when one of a pair returns to the nest and also occur during the "handling" of nest material, which continues during incubation. The boobies have a "Head-wagging" recognition display; the cormorants and anhingas have a comparable "Kink-throat" display; similarly, the frigate birds use a "Rattling" display.

In tropic birds, which do not normally have nest sites in the open, courtship display is aerial. Birds fly around in a group, calling, then two individuals may leave the group and fly together, the upper bird depressing his tail (especially the long central feathers) during gliding or hovering; the flights often end with one or both birds flying into a nest cavity. Tropic birds do not collect nest material but probably sometimes enlarge nest cavities with their bills or feet.

In all or nearly all pelecaniform families, mating frequently occurs with the same partner in successive years, and in many populations the same nest site may be occupied repeatedly. Since most species nest colonially, fighting over nest sites is common. It is particularly intense in the gannet and in populations of tropic birds when nest sites are in short supply.

Breeding seasons. Outside the tropics, egg laying by pelicans, boobies, and cormorants occurs typically in the spring, and the laying period extends over one to three months. In the tropics there is greater diversity in both the timing and extent of the laying period. Although many tropical populations have distinct laying periods at the same season in each year, a number of those in relatively seasonless environments have been known to lay in every month of the year. The blue-footed booby (*Sula nebouxii*) in the Galápagos, for example, is known to breed at intervals of less than a year. In other populations, laying is synchronized but occurs at nonannual intervals; *e.g.*, the brown booby (*S. leucogaster*) on Ascension Island has laying periods roughly every eight months. For the cormorants (*Phalacrocorax carbo* and *P. africanus*) of Lake Victoria, conditions are favourable for breeding during about eight months in each year, but the scarcity of islands free of predators forces the birds to occupy the few safe colonies in relays.

Nests, eggs, and young. Most pelecaniforms build substantial but untidy nests of twigs, grass, algae, or feathers; in many species guano accumulates on and around the nest as breeding progresses. Many cormorants and pelicans habitually build their nests on the ground, but most species that normally breed where trees or bushes are available build their nests in them. Certain cormorants and boobies sometimes nest on cliff ledges, and tropic birds and most boobies lay their eggs on the open ground in insignificant depressions.

Clutch size

There is much variation in clutch size. In general, species living in the unproductive tropical seas lay only a single egg. Those living in seasonally productive areas at higher latitudes, or in areas where surface waters are enriched by the upwelling of nutrient laden water from deeper layers, lay larger clutches and raise several young at a time. Tropic birds and frigate birds all have clutches of one egg, but pelicans generally lay one to three eggs, cormorants two to five, and anhingas three to five. The boobies show a more complex picture. The gannet of the temperate zones and the tropical red-footed booby (*Sula sula*) have clutches of one egg, but the other typical boobies of the tropics—the masked and brown boobies —normally lay two eggs. These two-egg species, however, normally raise only one chick, the older nestling apparently pushing the younger out of the nest soon after hatching. The second egg in these ground-nesting species seems to function as an insurance against infertility, loss, or breakage of the first, risks that may be lower in the tree-nesting, red-footed booby. The single-egg clutch of the gannet has not been satisfactorily explained, since pairs have been able to raise two chicks when an extra has been provided experimentally.

Except in the tropic birds, the eggs of pelecaniforms are small in relation to body size. In the tropic birds they are

9–13 percent of the adult weight, in the frigate birds, 6 percent, in the boobies, 3–5 percent, and in the pelicans and cormorants, 2–3 percent. In species with clutches of more than one, eggs are laid at intervals as short as 24 hours, or as long as five to six days in the case of the masked booby. In all of the groups, both parents participate roughly equally in incubation and feeding of the young. The lengths of the incubation shifts vary extensively, both within and between species. In some species (*e.g.*, many cormorants) shifts last only a few hours, but in tropic birds and some boobies shifts of three or more days are not uncommon; in the great frigate bird (*Fregata minor*), on the Galápagos, shifts last 10–15 days.

Incubation generally starts with the laying of the first egg (though not in some cormorants). First and last young may hatch several days apart and often differ substantially in size during the nestling period. The incubation period is just under four weeks in anhingas, four to four and a half weeks in cormorants, about four and a half weeks in pelicans, six weeks in tropic birds, and six to eight weeks in frigate birds.

Growth of young birds

The young, which are helpless and essentially naked at hatching, are brooded continuously for several days, until their down grows. The tropic birds are exceptional in that they have thick down at hatching and are sometimes left alone by their parents even on the day of hatching. The rate of development of the young varies dramatically and is correlated with the characteristics of the environments in which the birds live. The fledging period in the pelagic tropical masked booby is double that of cormorants of comparable body size living in the cool, productive waters off the coast of North America (17 weeks versus less than 8 weeks). Brown pelicans in North America fledge in less time (9 weeks) than white-tailed tropic birds weighing only one-tenth as much (10 weeks) and in less than half the time needed by tropical frigate birds weighing about one-third as much (more than 20 weeks).

The young of all pelecaniforms are fed on regurgitated food that they obtain by inserting their heads into the mouths of their parents. In frigate birds and tropical boobies, the parents continue to feed their young for a long period after they can fly, apparently because the young take a long time to become sufficiently skillful in hunting to survive in tropical areas where food is generally scarce. The most extreme case is that of the great frigate bird, in which the young can fly at about five months of age but sometimes are fed for more than a year after this. In the tropic birds and the gannet, however, the young are not fed by the parents after fledging; they leave their birthplace either by flying straight out to sea or by fluttering down to the water and swimming out to sea.

Molt. In the pelecaniforms, as in most birds, replacement of the feathers (molt) occurs mainly in the intervals between breeding. In a few species molt progresses concurrently with breeding activities, but normally, a complete molt commences when breeding is over. In the anhingas the primary flight feathers of the wing are shed simultaneously, but in the other groups molt is gradual. The primary and secondary feathers typically are replaced by means of two or three molt waves that start several feathers apart and move outward in parallel until all feathers have been shed. In most, if not all, pelecaniforms, feathers of the head and neck, and some on the body, are replaced a second time shortly before the next breeding season. Special feathers (crests, neck plumes, white patches) acquired at this time are shed around the time of egg laying in some species; in pelicans some head feathers are replaced yet again during incubation.

In many of the pelecaniforms the colours of the soft parts (especially the bare skin on the face) change as the birds become sexually active. In some pelicans a horny triangular plate grows on the upper mandible towards the tip before the breeding season and is shed after laying.

Ecology. *Survival and mortality.* Full-grown pelecaniform birds have few natural enemies, and although some are taken by marine predators they are generally long-lived. Nestlings, recently fledged young, and sometimes even adult birds suffer heavy mortality when food shortages occur. In many tropical areas where food is generally scarce, irregular fluctuations in the supply lead to drastic variations in breeding success. Mortality from starvation also occurs among birds of richer seas, most dramatically along the west coast of South America in years when oceanographic changes cut off the normally abundant food supply of the huge bird populations. Even under average conditions, young pelecaniforms in their first year after fledging experience much higher mortality than adults. In the European shag (*Phalacrocorax aristotelis*) more than half the young die during this period, although among adults annual mortality is only about 15 percent in males and 20 percent in females. In the British population of the gannet, about 80 percent of the fledglings die before reaching breeding age (about five years), most during their first year. Adults die at an average rate of less than 6 percent per year and have a mean expectation of life exceeding 16 years. Age records for individual banded birds include a great frigate bird that was recently found breeding at an age of more than 30 years and a masked booby breeding at about 23 years. Negative evidence suggests, however, that few individuals of these species survive much longer than the two examples.

Causes of death

Population movements. Members of the pelecaniform populations breeding at high latitudes generally move to lower latitudes in winter but do not perform transequatorial migrations. Adequate analyses of banding results have been carried out on only a few species. Great cormorants (*Phalacrocorax carbo*) from Britain do not show a well-defined migration but disperse random distances from the colonies. In all populations of gannets, juveniles move further than adults. A number of juvenile gannets from New Zealand have been found in eastern Australia only a week after leaving their natal colonies on their first flight, having made journeys exceeding 1,600 miles, at average speeds of more than 230 miles per day. Gannets that survive to breeding age usually return to the same cluster of nests where they were hatched. In a similar way, juvenile lesser frigate birds (*Fregata ariel*) from colonies in the central equatorial Pacific disperse on a broad front, many moving over 4,000 miles from their natal island. Their pattern of dispersal may be related to that of the prevailing winds in the area.

Feeding ecology. The various groups of pelecaniform birds are specialized for different ways of life and in particular for different feeding methods. Tropic birds and boobies, though they differ in many ways, are both adapted for catching prey underwater by plunging from the air. They are powerful flyers and buoyant swimmers, but, although their feet have large webs, their legs are not streamlined and they are not specialized for fast swimming. The feet are probably used underwater more for steering than for propulsion, and the birds depend mainly on the impetus of the dive to enable them to approach their prey at high speed.

Tropic birds are adept at hovering while locating prey, and probably obtain most of their food near the surface. Many boobies often dive from greater heights and probably go deeper than tropic birds; the gannet sometimes dives from more than 100 feet. Blue-footed boobies, when hunting in groups, tend to dive almost simultaneously. A disyllabic whistle is often heard from such groups as they start to dive and may be a signal given by the initiator. It has been suggested that the simultaneous plunging of several birds may confuse the fish in a school and so increase each bird's chance of catching one. Red-footed boobies, and perhaps also other boobies, catch flying fish (family Exocoetidae) in the air as well as in the water. These fish, with squid of the family Ommastrephidae, are staple foods of the tropical boobies and the tropic birds. Boobies of upwelling zones and the gannet feed almost entirely on fish.

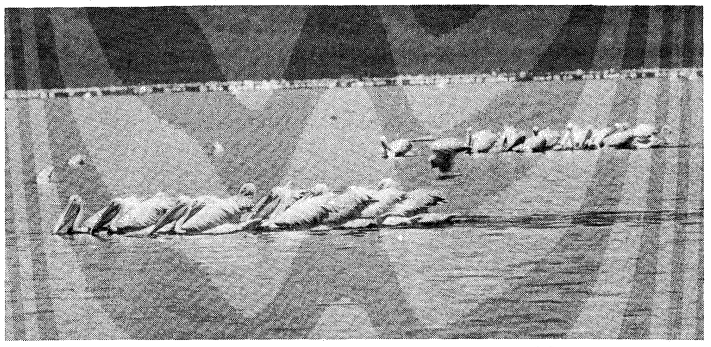
Frigate birds are among the most aerial of all sea birds. Their best known feeding habit involves piracy, in which they harry other sea birds until they disgorge their prey, after which the frigate birds catch the food in the air or pick it from the surface. They also catch prey for themselves, however, pursuing flying fish in the air and

snatching fish (and probably squid) from the surface without alighting.

Feeding by pelicans

Pelicans generally catch their prey while swimming, thrusting their long bills and long necks below the surface to scoop up fish in their distensible throat pouches. When fishing in shallow water, pelicans often cooperate to form a kind of living net. They form a crescent facing the shore, or even a circle, and then close in, splashing, paddling hard. Each bird keeps station until the fish panic and can be captured as they try to escape between the birds. Pelicans have also been observed herding and capturing ducklings in a similar way. The brown pelican, unlike the other species, often forages some distance offshore and habitually fishes by plunging from the air.

Leonard Lee Rue III—Bruce Coleman, Inc.



Flocks of white pelicans (*Pelecanus onocrotalus*) fishing cooperatively on Lake Nakuru, Kenya.

Cormorants and anhingas are adapted for underwater swimming. The cormorants pursue free-swimming or bottom-living fish, and some species also eat mollusks and other invertebrate animals. When feeding on schooling fish, cormorants often engage in mass fishing activities in which a flock advances in a long line stretching at right angles to the direction of movement, apparently with the fish fleeing ahead of them. The birds swim forward while above the surface and also while pursuing prey underwater, and laggards fly forward and land just ahead of the line. Anhingas do not pursue their prey but lie in wait underwater and then stab passing fish. Most of the fish that they eat are slow swimming and laterally flattened.

The physiological adaptations for diving have been little studied in cormorants and anhingas. It is known, however, that for cormorants feeding on the bottom the mean length of the dives is directly related to water depth. Dives of most species usually last less than half a minute, although dives as long as one minute are not uncommon. The time spent resting after a dive averages between one-half and one-third of the duration of the dive, but if dives are very long the length of the rest periods approaches that of the dives. Anhingas often stay under water for as long as two minutes, and a dive of nearly seven minutes has been recorded in a captive bird.

FORM AND FUNCTION

Distinguishing characteristics. Few generalizations can be made about the gross morphology of the pelecaniforms, since their external form reflects the diversity of their adaptations. In the larger and more aerial species, the skeleton is extensively pneumatized: nearly all the bones contain air sacs that are connected with the respiratory system. In the frigate birds, pelicans, and boobies, the major limb bones are no more than thin-walled hollow tubes with some internal struts. Pneumatization is least developed in anhingas, which need to have a high specific gravity to remain underwater while motionless.

Cushioning by air sacs

Those pelecaniforms that plunge into the water from a height (boobies, tropic birds, and the brown pelican) have a system of air sacs under the skin, connected with the lungs, that form a spongy mattress that presumably protects the bird as it strikes the water, although its presence also in species of pelicans that do not plunge suggests that the buoyancy it confers may be advantageous in itself.

The reduction of the external nostrils in most pelecaniforms is probably also connected with diving and swimming underwater. The closure, complete in boobies and in adult cormorants and anhingas, is only partial in frigate birds and pelicans; the nostrils are developed normally in tropic birds. The reduction of the nostrils in the aerial Fregatae as well as in the highly aquatic Pelecani suggests (as does the structure of the foot) that the Fregatae originated from a more aquatic stock. In the birds with nonfunctional nostrils, air enters the mouth cavity directly from the outside and then reaches the lungs in the normal way via the glottis—*i.e.*, the muscular opening to the windpipe—at the base of the tongue. In boobies and cormorants entry of air to the mouth when the bill is closed is by means of secondary external nostrils. These are slitlike openings at the angle of the mouth on each side where a horny flap (jugal operculum) at the base of the upper mandible overlaps the lower mandible.

Tropic birds, pelicans, and boobies have water-repellent plumage and are very buoyant. They swim high in the water and can reach significant depths only by acquiring momentum during aerial dives; their return to the surface is rapid. Frigate birds have wettable plumage, cannot swim or dive, and quickly become waterlogged if they alight on the water. In cormorants and anhingas the large contour feathers are wettable and trap little air when the birds are swimming underwater; this keeps buoyancy low in these birds and enables them to submerge from the surface without benefit of an aerial dive. The compressible but dense body plumage remains dry near the skin, at least in many species. The hind limbs of anhingas, which are not specialized for fast swimming, are relatively efficient for climbing and perching. Anhingas do not normally take flight from the water but climb out and dry their wings in a characteristic spread-wing posture that is also commonly used by cormorants. Once dry, anhingas show highly manoeuvrable flight and can also soar well, in contrast to the cormorants, which have a higher wing loading. Many species of pelecaniforms habitually spread their wings in "sun-bathing" postures, even when dry. The function of this behaviour is not well understood.

Physiological adaptations. Most pelecaniform birds nest in exposed situations, and special adaptations are necessary to prevent overheating of adults and young, particularly in tropical areas. The masked booby on the Galápagos, for example, stands with its back to the sun, thus shading its feet and naked throat pouch and permitting loss of heat by convection and conduction; it elevates its scapular (shoulder) feathers and droops its wings allowing maximum exposure of shaded feather areas, and it exhibits sustained fluttering of the throat pouch. Throat fluttering, which permits evaporative cooling with minimum expenditure of energy, is used under heat stress by all the pelecaniforms except the tropic birds, which do not have a naked throat pouch. The rate of throat fluttering remains roughly constant under increasing heat stress but becomes continuous instead of intermittent, and there are increases in its amplitude and in the throat area involved. All species make use of panting to increase evaporative cooling. Under heat stress the rate of breathing increases with increasing body temperature, and when it reaches a maximum the amplitude of breathing increases.

Although even the newly hatched young of certain species are capable of throat fluttering, they cannot regulate their temperature when left in the sun, until they have acquired down, and in some species not until they are several weeks old. Young tropic birds provide an exception: they have dense down at hatching and are sometimes left alone for long periods within a few days of hatching; small chicks are capable of regulating their body temperature by panting, even in the sun, but most nests are in the shade.

Salt excretion

Evaporative cooling requires much water, which is obtained from the food and probably also by drinking seawater. The resulting high intake of salt necessitates special adaptations for elimination of the excess. Marine birds have highly developed salt excretory glands, the lateral nasal glands, which function when the amount of salt taken into the body is more than the kidneys can deal

with. The glands secrete a concentrated solution of sodium chloride, which flows through ducts to the nasal cavity. In most sea birds, including the tropic birds, it then flows out through the nostrils and may drip or be shaken off the tip of the bill. In boobies and cormorants, in which the nostrils are closed, the solution trickles from the internal nares in the roof of the mouth and out at the end of the bill. In some species of cormorants that occur in both marine and freshwater habitats, the salt gland is relatively larger in individuals from marine habitats. Presumably these individuals tend to eat food with a higher salt content or drink seawater for cooling.

CLASSIFICATION

Distinguishing taxonomic features. The classification of the order Pelecaniformes is based mainly on morphological features. Recent attempts have been made to clarify relationships by analysis of egg-white proteins, and by comparative analysis of social behaviour. The parasitic feather lice (Mallophaga) found on members of the different groups also provide clues to relationships.

The six modern families of pelecaniform birds are readily distinguished, although the relationships among them are not firmly established. The families are internally homogeneous, and in each of them the modern species are often placed in single genera; the Phaethontidae, Fregatidae, Pelecanidae, and Anhingidae are usually treated in this way. The relationships of many fossil groups—indicated below by a dagger (†)—are quite uncertain.

Annotated classification.

ORDER PELECANIFORMES

Aquatic birds characterized by totipalmate feet (webs connecting all 4 toes); the 1st toe (hallux) not pointing backward as in most birds, but turned inward and joined by a web to the 2nd toe.

Suborder Phaethontes

Family Phaethontidae (tropic birds)

Medium sized (16–19 in. or 40–48 cm long). Plumage satiny white with black markings; central 2 tail feathers much elongated in adult. Sexes similar. Bill laterally compressed, slightly decurved, and with undivided horny sheath (rhamphotheca). Nostrils open in slits; nasal glands each with 2 ducts. No naked gular (throat) pouch. Neck short, 14–15 cervical vertebrae. Pelvic girdle reduced, legs short, set far back. No nest building behaviour. Single egg heavily marked with brown and black. Chick hatched covered with grey down. 1 fossil genus, *Prophaeton* (lower Eocene of England); 1 extant genus, *Phaethon*; 3 living species, worldwide in tropical oceans.

Suborder Fregatae

Family Fregatidae (frigate birds)

Large sized (31–34 in. or 79–86 cm long), light bodied. Plumage blackish, juveniles and some adults with white underparts; juveniles with white heads, sometimes tinged with orange. Wings very long and pointed, tail deeply forked. Sexes differ in size (female larger) and in coloration of plumage and soft parts. Bill long, slender, strongly hooked at tip; horny sheath divided by deep grooves. Nostrils almost closed; nasal glands each with a single duct. Small gular pouch, highly distensible and brilliant red in courting male. Neck short, with 14–15 cervical vertebrae. Furculum (wishbone) fused with coracoids and with sternum (breastbone). Pelvic girdle and leg muscles reduced; legs very short. Feet small, with webbing only between bases of toes. One white, thin-shelled egg. Young naked at birth, soon acquire thick white down. No fossil record. One extant genus, *Fregata*; 5 living species, worldwide in tropical oceans.

Suborder Pelecani (pelicans, boobies, cormorants, anhingas)

Horny sheath of bill divided by deep grooves (largely suppressed in anhingas). Naked gular pouch; facial skin and pouch colorful, especially in breeding season. Nasal glands each with a single duct. Neck long, with 17–20 cervical (neck) vertebrae; 8th and 9th have articulations permitting sharp S-bend, especially in anhingas. Pelvic girdle not reduced, legs strong. Feet large, fully webbed. Eggs thick shelled, with chalky white outer layer. Young naked at hatching.

Family Pelecanidae (pelicans)

Very large (50–72 in. or 127–180 cm long). Plumage white, brown, or grey; some species crested. Wings long and broad, tail short. Sexes similar. Bill long, hooked at tip, with enormous gular pouch suspended from highly distensible rami (lateral elements) of lower mandible. Nostrils almost closed. Furculum fused with sternum. Down of nestlings white. 10 fossil species in the genera *Pelecanus* (the earliest from lower Miocene of France) and *Liptornis* (middle Miocene). One extant genus, *Pelecanus*; 7 living species, seacoasts and freshwater lakes of southern Africa, central and eastern Eurasia, Australia, and North America; north and west coasts of South America.

Family Sulidae (boobies)

Large (26–40 in. or 66–102 cm long). Plumage of most species white with black on wings; some brown and white; juveniles largely grey or grey-brown. Wings long and narrow, tail rather long, wedge shaped. Sexes generally differ in voice, sometimes in colour of bare skin. Bill stout, conical, slightly decurved near the tip and finely serrated on cutting edges. Nostrils closed. Down of nestlings white. 18 fossil species, in the genera *Sula* (the earliest *S. ronzoni*, lower Oligocene of France), *Microsula*, *Morus*, *Palaeosula* and *Miosula*. Two extant genera, *Sula* with 6 living species, *Morus* with 1 species (sometimes considered 3 species, forming a superspecies); widespread but irregular distribution, on seacoasts and oceanic islands.

Family Phalacrocoracidae (cormorants)

Medium to large (19–40 in. or 48–102 cm long). Plumage usually blackish; some with white underparts, and a few largely grey; some species have ornamental crests or patches of white when breeding. Wings broad, tail long and stiff. Sexes similar, but females up to 30 percent lighter in weight. Bill slender, sharply hooked at tip. Nostrils closed in adults. Bony occipital style projecting upward from back of skull. Down of nestlings dark. 24 fossil species in the genera *Graculavus* (the earliest, upper Paleocene of North America), *Actiornis*, *Miocarbo*, and *Phalacrocorax*. About 28 living species (and one recently extinct), all currently included in the genus *Phalacrocorax;* certain Old World species have sometimes been placed in a separate genus *Halietor*, and the Galápagos flightless cormorant (*Phalacrocorax harrisi*) in the genus *Nannopterum*.

Family Anhingidae (anhingas or darters)

Large (34–36 in. or 86–91 cm long). Plumage mainly black and brown with pale markings on back and wings; breeding male has light-coloured plumes on head and neck. Tail very long and stiff. In some forms the sexes differ in colour. Bill long, straight, slender, sharply pointed; cutting edges serrated towards tip. Nostrils closed in adults. Occipital style smaller than in cormorants. Head small, neck long and slender. Legs short but strong. Pyloric lobe of stomach with mat of hairlike processes. Down of nestlings brown. Five fossil species, in the genera *Protoplotus* (the earliest from the Eocene of Sumatra), and *Anhinga*. One extant genus, *Anhinga*; living forms have been treated as 1, 2, or 4 species and are here considered as 4 species comprising a single superspecies.

†*Family Elopterygidae*

Fossil. Marine birds probably related to the Pelecani, perhaps especially to the cormorants. Genera *Elopteryx* (Upper Cretaceous, Maastrichtian, of Romania); *Argillornis* and *Eostega* (Eocene of Europe).

†*Family Pelagornithidae*

Fossil. Marine birds probably related to the Pelecani, perhaps especially to boobies, but not adequately known. Genus *Pelagornis* (middle Miocene of France).

†**Suborder Odontopteryges**

Fossil. Very large marine birds, with long wings and short legs. Bill long and high, with sharply pointed, toothlike projections on margins of upper and lower mandibles, continuous with bony structure of jaws (and thus not true teeth). Upper and lower mandibles with longitudinal grooves, suggesting that their horny sheaths were divided.

†*Family Odontopterygidae*

Fossil. "Teeth" slant forward. Nostrils closed or nearly so. About half the size of the pseudodontorns (see below). Genus *Odontopteryx* (lower Eocene of England).

†*Family Pseudodontornithidae*

Fossil. "Teeth" vertical—not slanted forward. Wingspread about 4.5 metres. Genera *Pseudodontornis* (Miocene); and *Osteodontornis* (upper Miocene of California), one of the best preserved of early marine birds.

†*Family Cyphornithidae*

Fossil. Based on fragmentary material, and of uncertain relationships, but may be members of this suborder. Lower Miocene of North America. Genera *Cyphornis*, even larger than

the pseudodontorns; *Palaeochenoides*, rather larger than pelicans, may be synonymous with *Pseudodontornis* (and would take priority over it); *Tympanonesiotes*, smaller than the preceding form but may be related to it.

†Suborder Cladornithes

Fossil. Relationship to pelecaniforms doubtful. One family, with 1 genus, *Cladornis* (lower Oligocene of Patagonia); known only from a poorly preserved tarsometatarsus.

Critical appraisal. The tropic birds (Phaethontes) differ in many ways from the other members of the order Pelecaniformes and may well be wrongly placed in this order. They are united to the others primarily by the totipalmate condition, but the first toe in Phaethontes is distinctly more elevated on the tarsometatarsus than in the other groups. Some of the resemblances of the Phaethontes to the Laridae (gulls and terns) in the order Charadriiformes may not result from evolutionary convergence but may indicate that the Phaethontes arose from forms more closely related to the basal stock of the Laridae than to that of the Pelecaniformes.

The frigate birds (Fregatae) are generally treated as a distinct suborder, on the basis of a number of morphological characters that they share with members of the order Procellariiformes but not with the Pelecani. Pending further evidence from protein structure, this distinction seems worth maintaining.

There is little doubt that the suborder Pelecani is made up of closely related groups. The pelicans are sometimes placed in a separate superfamily, but it is more useful to emphasize the similarities among the modern groups by keeping them together. The anhingas have sometimes been treated as a subfamily of the Phalacrocoracidae, but their morphological and ecological differences from cormorants perhaps justify separation at the family level.

The Odontopteryges show a number of similarities to the Procellariiformes as well as to the Pelecaniformes and also show differences from both. The suggestion that they deserve to rank as a separate order Odontopterygiformes will be strengthened if it can be proved (as is now suspected) that *Osteodontornis* lacked or had a much reduced first toe, since this is an essential element of the totipalmate foot of pelecaniforms.

Available evidence suggests that the stocks leading to the Procellariiformes, Odontopteryges, typical pelecaniform birds (Fregatae plus Pelecani), Phaethontes, and Laridae, diverged during the Cretaceous Period, probably at least 70,000,000 years ago.

BIBLIOGRAPHY. R.S. PALMER (ed.), *Handbook of North American Birds*, vol. 1 (1962), detailed accounts of the biology of all species of North American pelecaniforms; R.C. MURPHY, *Oceanic Birds of South America*, 2 vol. (1936), a classic work that includes excellent accounts of the guano birds of the Peru Current; D. LACK, *Ecological Adaptations for Breeding in Birds* (1968), a modern analysis that includes a useful discussion of pelecaniforms; G.F. VAN TETS, "A Comparative Study of Some Social Communication Patterns in the Pelecaniformes," *Amr. Orn. Union, Orn. Monogr. No. 2* (1965), a general discussion of behaviour and relationships of the different groups; B. NELSON, *Galapagos: Islands of Birds* (1968), readable and authoritative accounts of the ecology and behaviour of tropical pelecaniforms, especially boobies and frigatebirds; P. BRODKORB, "Catalogue of Fossil Birds: Part 1 (Archaeopterygiformes through Ardeiformes)," *Fla. St. Mus. Bull.* 7, pp. 179–293 (1963), a list of the fossil pelecaniform birds.

(N.P.As.)

Peloponnesian War

The Great Peloponnesian War (431–404 BC), between Sparta and Athens, was one of the great watersheds in the history of the Ancient Greeks, and the subject of one of the greatest historical works of all time, the *Peloponnesian War* of Thucydides of Athens.

Thirty Years' Treaty

The antecedents of the war. The Thirty Years' Treaty of 445 brought a truce between the two leading blocs of power in Greece: Athens and its empire, which comprised most of the islands and maritime states on the northern and eastern coasts of the Aegean basin, along with the Greek states of the Propontis and Black seas; and Sparta and its allies, among whom were most of the leading land powers in the Peloponnese and central Greece, and Corinth, the leading sea power in the western seas extending to Sicily. The peace lasted as long as neither bloc trespassed on the other's sphere of influence; they were about equal in strength, and the outside powers were neutral. But in 433 Athens made a defensive alliance with a recalcitrant colony of Corinth, Corcyra, which held a strategic position in the western seas and possessed a fleet of 120 triremes (galleys with three banks of oars). This alliance infringed the principle but not the letter of the Thirty Years' Treaty and resulted in a clash between Athenian and Corinthian ships when Corinth and its allies attacked Corcyra. In the ensuing winter Athens not only banned one of these allies, Megara, from access to the eastern seas, which had been guaranteed under the Thirty Years' Treaty, but also ordered one of its own allies, Potidaea, which was a colony of Corinth, to break off relations with its founder city and dismantle its seaward walls. These were no idle threats because the fleet of Athens dominated the Aegean basin. Nevertheless, Potidaea refused. "Volunteers" from Corinth and its Peloponnesian allies reached Potidaea and enabled it to beat off an Athenian attack.

Corinth, Megara, and other states now maintained that Athens had committed aggression and had thus infringed the terms of the Thirty Years' Treaty. They complained, therefore, to Sparta. The Spartans' local interests were not involved. Yet their position at the head of the Spartan alliance was endangered by Athens' moves against two of its members, Corinth and Megara. The executive authorities at Sparta had already made a secret agreement to help Potidaea, but this fact was not known to the Spartan citizens when they met in the assembly to decide their attitude. A large majority found Athens guilty of aggression. The allies of Sparta, meeting in a formal congress, confirmed Sparta's verdict. The two components of the Spartan alliance, or, as it is commonly called, the Peloponnesian League, declared themselves ready to go to war if the Athenians proved unwilling to withdraw their ban on Megara and their order to Potidaea.

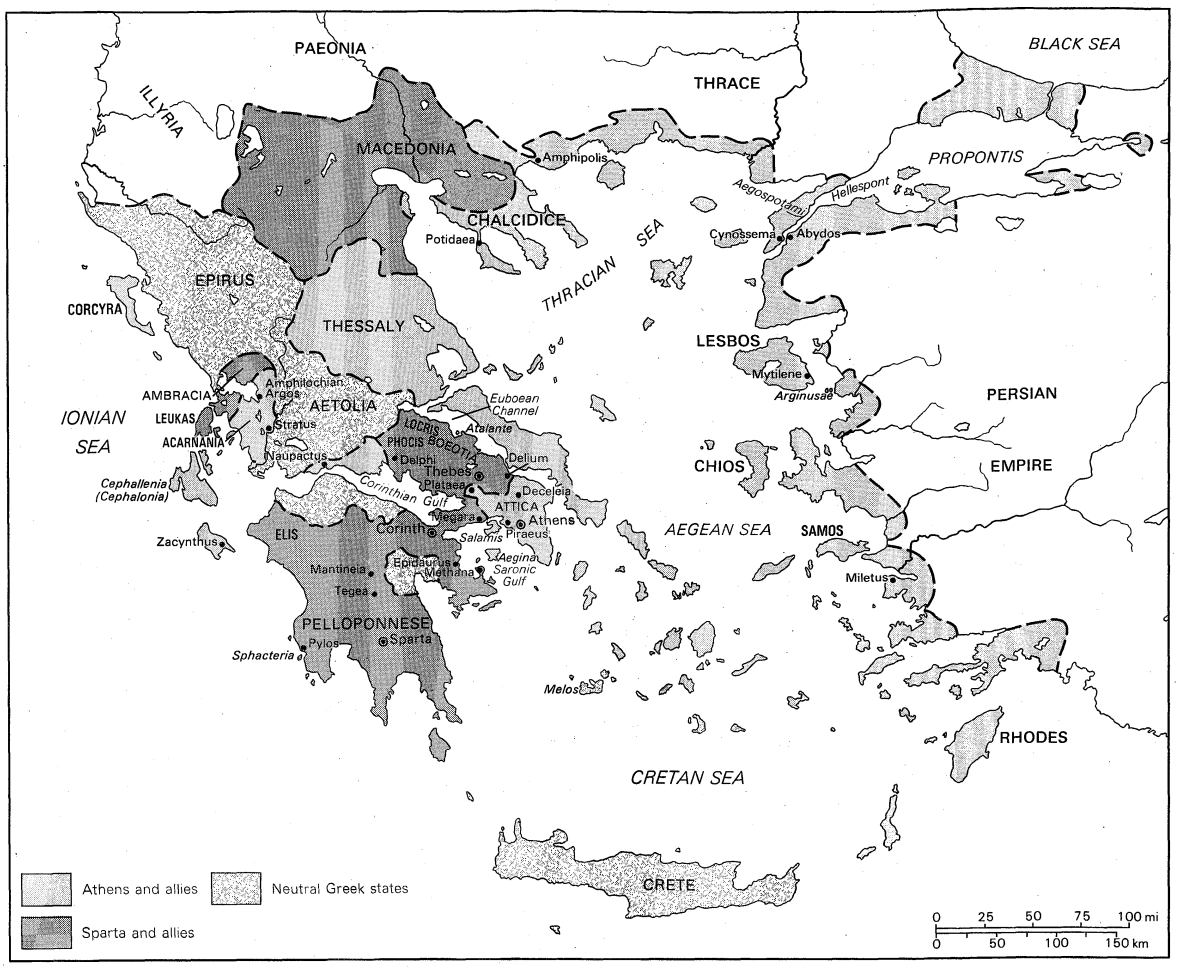
Peloponnesian League

The decision lay with Athens. The people were swayed by their leading statesman, Pericles, who had himself proposed the alliance with Corcyra and the ban on Megara. He warned the assembly against yielding to any demands made by the Peloponnesians. The majority accepted his advice, and the demands of Sparta were rejected. A diplomatic deadlock ensued. The Athenians offered to accept arbitration on the points under dispute. The Spartans refused arbitration. In spring 431 an ally of Sparta, Thebes, made a treacherous attack on an ally of Athens, Plataea. The Thirty Years' Treaty was flagrantly broken, and both blocs of power went to war. As Thucydides observed, the fundamental cause of the war was the interaction of two forces: the growing power of Athens and the growing apprehension of Sparta.

The first part of the war (the "Archidamian War," 431–421 BC). The armies of Sparta and its allies were far stronger than those of Athens. But they could be deployed for only a few months each year, because the Spartans had to keep their serfs, the helots, under control, and the allied soldiers had to maintain their farms. The Spartans hoped to force a decisive battle on land. Otherwise there was little hope of success, for Sparta and its allies had relatively small fleets and no financial resources for building large numbers of warships. On the other hand, the fleet of Athens, even without the squadrons of Lesbos, Chios, and Corcyra, was unrivaled; and the Athenians had large financial reserves and access to shipbuilding materials. Their infantry, though outmatched by the combined forces of their enemies, included a striking force of 13,000 men, free to operate at any time and capable of damaging their neighbours individually, and a reserve force of 16,000 men, able to hold the massive walls that circled Athens and the Piraeus and linked the city with its port.

The opening campaigns

Sparta's lines of communication with its allies beyond the isthmus—Boeotia, Phocis, and Locris—ran through the narrow corridor of northern Megaris, which separates Attica from the Corinthian Gulf. This corridor was



Greece at the beginning of the Peloponnesian War.
Adapted from W. Shepherd, *Historical Atlas*; Barnes & Noble Books, New York

threatened from the east by Athens and from the west by Athens' ally Plataea. In their desire to control the corridor, the Thebans attacked Plataea in March 431. The attack failed. Athens evacuated the noncombatants from Plataea and sent troops to reinforce the garrison. In May the Spartan king, Archidamus, led the allied armies slowly into Attica and ravaged part of the plain north of Athens, hoping to provoke an engagement. But on the orders of Pericles the Athenian population and the Athenian army were concentrated behind the walls of the city. Although the troops wanted to attack, Pericles held them back. He had no intention of offering battle then or later to the superior army of the enemy. He proposed to conserve his forces for a war of attrition during which the Athenian fleet would guard the empire and Athens' supplies and wear down the enemy by seaborne raids. In June, therefore, before the Spartans left Attica, a fleet of 150 triremes set off from the Piraeus. It raided the coasts of the Peloponnese and closed the mouth of the Corinthian Gulf by winning over Cephallenia (Cephalonia), an island off the entrance to the gulf, as a base. Meanwhile, a fleet of 30 triremes ravaged the coast of Locris and garrisoned the island of Atalante in the Euboean Channel, through which convoys passed to an Athenian army blockading Potidaea. Another force expelled the people from Aegina, and Athenians occupied the island, controlling the entry to the Saronic Gulf. Thus the net of naval blockade was closing round the enemies of Athens and its own sea lanes were secure.

The policy of Pericles was a wise one in the opinion of Thucydides. It safeguarded the city, the fleet, and the empire—the sources from which Athens drew its financial and naval strength. Even if it meant the loss of property in Attica, which was relatively unimportant, it involved no risks for the Athenian army; and Pericles was confident that he could maintain the morale of the people. In 431 and subsequent years the army ravaged the Megarid. In March 430 Pericles proclaimed his faith in Athens when he delivered a funeral speech in honour of the Athenian dead. But shortly afterward, when the Peloponnesian army was ravaging Attica for the second time, an unexpected blow fell on the crowded city of Athens—a terrible plague that within three years killed one-third of its best troops and maimed many others, apart from its toll among the civilian population. Pericles himself was one of the casualties. The power and the morale of the Athenians were seriously undermined; but they fought on with grim determination in Chalcidice, capturing Potidaea and instigating an extensive raid on Chalcidice and Macedonia by their ally, the Thracian king Sitalces, in 429.

Plague at Athens

While the plague raged at Athens, the initiative lay with the Spartans. They tried to oust the Athenians from their bases in the western area—Zacynthus, Cephallenia, Acarnania and Amphilochian Argos, Corcyra and, in the Gulf of Corinth, Naupactus—and thus to regain contact with Italy and Sicily, where they had allies and sources of supply. In 430, attacks on Zacynthus and Amphilochian Argos failed. In 429 a large army of Peloponnesians, Ambraciots, and light-armed levies from the tribes of Epirus invaded Acarnania but suffered defeat at Stratus. A Peloponnesian fleet of 47 warships and smaller troop-carrying vessels, which had intended to join the attack on Acarnania, was intercepted as it sailed out of the Corinthian Gulf by an Athenian squadron of 20 warships based at Naupactus. When the Peloponnesian warships formed a circle, facing outward to protect the smaller vessels at the centre, the Athenian commander Phormion ordered his warships to row round the outside of the circle, which contracted and fell into confusion. Phormion then attacked and gained a remarkable victory. The Peloponnesian fleet then threatened Naupactus. But Phormion won another victory, and reinforcements from Athens soon

reached Naupactus. Later in the year the Peloponnesians set out to raid the Piraeus by sea; but their courage deserted them, and they raided the island of Salamis instead. The other objective of Sparta in 429 was Plataea. When diplomacy failed, the Spartans built a double wall round the town and imposed a blockade that was destined to last into 427.

Mytilene revolt

In 428 the Peloponnesians made no further move in the western area, but an important opportunity arose for intervention in Aegean waters. Lesbos, one of the two allies of Athens that had a fleet, was planning to revolt and allied itself in August with the Spartans, who undertook to invade Attica that autumn and to send help overseas to Lesbos. But the Athenians moved first, blockading the leader of the revolt, the city of Mytilene, with a squadron of 40 ships and ravaging the east coast of the Peloponnese with an army transported on 100 ships. The Peloponnesians failed to invade Attica or to send help to Mytilene. During the winter Mytilene was hard pressed by the Athenians. The oligarchs in the town who had organized the revolt found themselves threatened by the democratic party, and they surrendered to the Athenian commander on condition that the Athenian people should make a final settlement. A week later a Peloponnesian squadron of 42 ships crossed the Aegean Sea, learned of the surrender, and sailed ignominiously home. The only success that Sparta gained during the years of the plague was the capture in 427 of Plataea, from which half the garrison had made a brilliant escape at night. The remainder surrendered, entrusting their fate to Spartan courts.

The strain of war lowered the standard of humanity on both sides. The Athenians, incited by a violent orator, Cleon, ordered the execution of all adult males at Mytilene and the enslavement of the rest. Next day the assembly repented. The order was countermanded just in time, but even so a thousand ringleaders were executed. The Spartans condemned and executed the captured garrison of Plataea, and the town was demolished by the Thebans. In Corcyra a terrible civil war, fanned by the intervention of a Peloponnesian squadron and later of an Athenian squadron, caused much loss of life and degenerated into guerrilla warfare, which lasted until 425. Supplies were now short in the Peloponnese. At Athens financial reserves were running low; and bitter animosity developed between the extreme democrats, led by Cleon, and the moderate party with which Thucydides and Aristophanes sympathized.

Athenian offensives

In 427–424 the Athenians abandoned the cautious strategy of Pericles and engaged in a series of offensives. An attack on Melos failed (426). A squadron of 20 ships started operations in 427 against Syracuse in Sicily, where Athens had allies, and gained control of the Strait of Messina; 40 more ships were sent in 425, but the Sicilian states made peace among themselves in 424 and so ended Athenian intervention. An enterprising general, Demosthenes, tried to extend Athens' power in western Greece by attacking Leukas and invading Aetolia, where he suffered defeat. In 426 the Spartans retaliated. Campaigning along the northern coast of the Corinthian Gulf and then joining the Ambraciots in an attack on Amphilochian Argos, they hoped to reduce Acarnania and control the northwestern mainland. But Demosthenes came up with 20 Athenian ships, took command of the Acarnanians and Amphilochians, and inflicted a double defeat on Sparta and Ambracia. His success was so great that the Acarnanians were alarmed and made a separate alliance with Ambracia. From this Athens derived no advantage.

In 425, when the reinforcements were sailing to Sicily, Demosthenes fortified Pylos, a rocky promontory on the west coast of Messenia, and stayed there with a small force. A Peloponnesian fleet of 60 ships, which had been operating at Corcyra, hastened south to Pylos and landed 420 Spartans on an island, Sphacteria, hoping to blockade the promontory. The Athenian fleet returned suddenly, defeated the Peloponnesian fleet by surprise, and isolated the Spartans on Sphacteria. Sparta then negotiated an armistice at Pylos, during which the Athenians were given possession of the Peloponnesian fleet as a guarantee of good faith. Sparta offered peace and alliance to Athens. Cleon persuaded the assembly to reject the offer in the hope of winning some decisive victory, and later in the year he was sent out to take command with Demosthenes at Pylos. There the Athenians, alleging a technical breach of the armistice, had refused to return the Peloponnesian ships but failed to land on Sphacteria. Cleon and Demosthenes, encouraged by a fire that burned the woods on the island, made a landing, fought the Spartans to a standstill, and obtained the surrender of 120 Spartiates.

Stripped of its fleet, shaken in morale, and threatened by helot risings in Messenia, Sparta was now at its lowest ebb. Athens tried to press home its advantage. During the voyage to Sicily the fleet from Pylos connived at a massacre of oligarchs at Corcyra. Nicias led successful raids by sea on the coasts of Corinth and Epidaurus and occupied Methana. In 424 he captured and garrisoned Cythera as a base for raiding Laconia. Demosthenes and Hippocrates came within an ace of capturing Megara and thus cutting the Peloponnese off from Boeotia. In the summer a triple attack was launched on Boeotia; but a failure in the timing and the slowness of the main force under Hippocrates at Delium enabled a Boeotian army of 7,000 hoplites (heavy-armed infantry) and 1,000 cavalry to force a battle, in which Hippocrates and 1,000 Athenian hoplites were killed. This disaster brought the Athenian offensive to an end.

The Battle of Delium

Meanwhile the Spartans broke out of the net in which the Athenians sought to enclose them. An able Spartan officer, Brasidas, having saved Megara from capture by Demosthenes and Hippocrates, led an army of Peloponnesian troops and freed helots into Chalcidice, an area neglected by Athens since the fall of Potidaea. There he raised city after city in revolt against Athens and in December 424 captured Amphipolis, Athens' chief base in the northwest Aegean. Athens tried to stop the landslide by concluding a year's armistice with Sparta; but it proved ineffective in Chalcidice, where Brasidas continued to detach states from the Athenian empire. Afraid of more widespread revolt, the Athenians sent an army under Cleon's command to recapture Amphipolis in 422. It was decisively defeated; Cleon was killed, but so was Brasidas. Nicias, Cleon's rival, now persuaded the assembly to accept the overtures for peace that Sparta had been making since 425. A treaty of peace was accepted in 421 by Athens and Sparta but not by four of Sparta's leading allies (Boeotia, Corinth, Elis, and Megara). Soon afterward Athens and Sparta entered into a defensive alliance, not of good will but of convenience.

The ineffective peace, 421–415 BC. These years gave the Athenians an opportunity to seduce Sparta's allies and isolate Sparta by diplomacy. They failed to use it because they were split internally by the rival policies of appeasement, led by Nicias, and of aggression, led by Alcibiades. The other states now distrusted both Athens and Sparta, but a breakaway group of Argos, Corinth, Elis, and Mantineia was too weak to withstand either great power. In the atmosphere of general mistrust the terms of the so-called Peace of Nicias were never implemented. Sparta managed to make an alliance with Boeotia, and Athens won over Argos, Elis, and Mantineia but not Corinth in 420. Sparta took military action in 418. The Spartan army of 6,000 hoplites and the levies of Boeotia, Megara, Corinth, Tegea, and lesser allies outmaneuvered the forces of Argos (numbering 6,000 hoplites), Elis, Mantineia, and some lesser states and imposed an armistice on Argos. Sparta's allies had gone home when Argos received 1,000 hoplites and 300 cavalry from its ally Athens and reopened hostilities. At Mantineia the army of Sparta, aided by Tegea alone, inflicted a decisive defeat on Argos, Mantineia, and Athens and rapidly re-established the Spartan alliance as it had been in 431. The feeble intervention of Athens in the Peloponnese with a mere 1,000 hoplites had lost the chance of diplomatic or military success on land. Meanwhile in the north, Amphipolis and Chalcidice remained independent and hostile. The war party at Athens now favoured adventure at sea. In 416 Athens attacked Melos, which refused to renounce neutrality and accept Athens' rule. When Melos capitulated, Athens massacred the adult males, enslaved the rest, and gave the island to 500

Athenian settlers. In 415 the Athenians sent out a great armada to conquer Sicily. These acts showed that their methods were unscrupulous and their ambitions were unlimited. They heralded the end of the Peace of Nicias.

The second part of the war, 415–404 BC. This was decided at Syracuse and at sea. On this element the Athenians began with complete supremacy. In all they committed 200 warships to the gamble in Sicily and left their home fleet dangerously weak. The expeditionary force in Sicily was commanded by three incompatible generals: Alcibiades, Lamachus, and Nicias. Alcibiades was soon recalled on a charge of sacrilege; he escaped and transferred his services to Sparta. Lamachus was killed in action. Nicias was soon affected by illness. In 414 the first attempt was made to surround and blockade Syracuse by land (for Athens held command at sea); it failed, mainly because a Spartan, Gylippus, with nearly 3,000 troops slipped through the Athenian lines and enabled the Syracusans to defeat the Athenian land forces, which had had only a small margin of superiority. Nicias then concentrated his army in defense of the naval base by the great harbour and advised Athens to recall the expedition. He was told to await reinforcements.

In July 413, when Nicias had lost his naval base and been defeated at sea, Demosthenes arrived with a large force. It was essential to establish military superiority at once. Demosthenes therefore launched a night attack, suffered a severe defeat, and wished to sail home. Nicias now refused to go. A month later Nicias agreed to sail; but after an eclipse of the moon on Aug. 27, 413, Nicias accepted his soothsayers' advice to stay "thrice nine days." While he stayed, the fate of the expedition was decided in the Great Harbour; the morale of the Athenian navy was broken by a single defeat, and the whole force attempted to escape overland. It suffered complete destruction. Mismanagement, indecision, and panic thus cost Athens more than 200 warships with crews totalling about 35,000 men, a fine army including 4,000 Athenians, and much material and treasure.

Athens' defeat at Syracuse

The Spartans and their allies declared war in 414. On the advice of Alcibiades, they established a fortified post at Deceleia in northern Attica that denied the Athenians the use of Attica and made them import all foodstuffs by sea at great expense. The Athenian disaster in Sicily gave Sparta hope of winning at sea. With Alcibiades' help the Spartans entered into alliance with Darius II of Persia, who could provide bases and money; they were joined by Chios, Miletus, and other states that revolted from Athens; and a small Peloponnesian fleet established itself at Miletus. The Athenians showed great energy; they built ships quickly, secured Samos as their main base, and covered their chief line of supply through the Hellespont. In 411 Abydos in the Hellespont revolted with the help of a Spartan force and Persia. At this stage the Persian king adopted a waiting game on the advice of Alcibiades, who suggested that he let the Greeks wear one another down and use his wealth to end the war at his pleasure. Alcibiades engineered his own recall to the Athenian fleet at Samos. He began by promoting an oligarchic revolution, which developed partly from the hypothesis that as an oligarchic leader he could obtain aid from Persia. Oligarchy at Samos was short-lived, but the succeeding democratic leaders appointed Alcibiades general of the fleet. Meanwhile at Athens an oligarchy seized power, ruled despotically, and negotiated with Sparta for terms of peace. It was overthrown by a popular rising, and a moderate government known as "the Five Thousand" took its place and cooperated with the democratic fleet at Samos. The slowness of Sparta and the deliberate inaction of Persia gave Athens the chance to ride out the storms of disaster in Sicily and of revolution at home. Late in 411 the Athenians won their first victory against the Peloponnesian navy at Cynossema in the Hellespont. Two more victories followed. They put the main enemy fleet out of action for three years and merged the two governments at Samos and Athens in a restored democracy.

The democratic leaders, especially Cleophon, pushed the war to a finish by rejecting the terms that Sparta repeatedly offered and by attacking those who advocated negotiation as oligarchs. Neither side had the money to outdo the other. Alcibiades came home in triumph in 407; but Persia chose to subsidize the Spartan admiral, Lysander, who equipped 90 ships and defeated an Athenian fleet. Alcibiades was held responsible by the Athenians; he fled into exile in 406. By tremendous efforts the Athenians survived further defeats at sea, manned 150 ships, and won a battle at Arginusae in which about 20,000 Greek lives were lost. The Persian policy of attrition was now coming to an end. In 405 the Persian prince Cyrus enabled Lysander to man 200 ships. He entered the Hellespont, drew the Athenian fleet of 180 ships after him, and caught them ashore in a sudden attack at Aegospotami. Only eight ships escaped. The Athenian navy was no more. The relentless advance of Lysander to the Piraeus and the blockade imposed from Deceleia starved the Athenians into final submission but only after Cleophon had been executed and the way had been prepared for an oligarchic government. In April 404 Lysander's troops began to demolish the walls of the Piraeus. Athens was to be stripped bare of empire, fleet, and fortifications, so that it should never try again to enslave the Greek states.

The fall of Athens

The Spartans owed their victory to the tenacity of their government, their excellent infantry, and their power of leadership, but also to the fighting quality of the Syracusans and to the gold of Persia, for which Sparta promised to sacrifice the liberty of the Greeks in Asia. The Athenians owed their failure to instability in government and temperament, mistaken policy, incompetent generalship, and declining morale. Although the principle of national independence was strikingly vindicated by the outcome, the general strength of the Greek states was fatally weakened; and Athens, the most progressive and civilized state, was reduced to impotence.

BIBLIOGRAPHY

Translations of Thucydides: R. CRAWLEY, *Thucydides' History of the Peloponnesian War* (1910), close to the style and diction of Thucydides; R. WARNER, *Thucydides: The History of the Peloponnesian War* (1954), in clear, modern English.

General accounts: *The Cambridge Ancient History*, vol. 5, ch. 7–13 (1927), a good narrative, written by leading scholars of the time, without reference to ancient sources; B.W. HENDERSON, *The Great War Between Sparta and Athens* (1927); J.B. BURY, *A History of Greece*, 3rd ed., pp. 390–507, rev. by R. MEIGGS (1952), concentrates on military and political aspects; N.G.L. HAMMOND, *A History of Greece to 322 B.C.*, 2nd ed., pp. 311–436 (1967), deals with military, political, and cultural aspects and gives references to ancient sources.

Special topics: D. KAGAN, *The Outbreak of the Peloponnesian War* (1969), a well-documented study; C. HIGNETT, *A History of the Athenian Constitution to the End of the Fifth Century B.C.* (1952), a detailed and critical study of wartime political systems at Athens; G.B. GRUNDY, *Thucydides and the History of His Age*, 2nd ed., 2 vol. (1948), an original and stimulating book; J. DE ROMILLY, *Thucydide et l'impérialisme athénien* (1947; Eng. trans., *Thucydides and Athenian Imperialism*, 1963), gives a clear insight into power politics at Athens; H.D. WESTLAKE, *Individuals in Thucydides* (1935, reprinted 1968), a lucid account of personalities of the war.

(N.G.L.H.)

Penn, William

The 17th-century English Quaker William Penn, a religious leader, social philosopher, and colonial proprietor, was one of history's greatest advocates of religious toleration. In the American Commonwealth of Pennsylvania, which he founded, he sought to embody his political and religious principles.

Early life and education. He was born in London, October 14, 1644. His father, Adm. Sir William Penn, was forging a successful career in the navy at the time of his son's birth and could serve only as a distant model during the child's first dozen years because he was almost always at sea. Young William's mother, Margaret (*née* Jasper) van der Schuren Penn, was a childless widow when the elder Penn married her in 1644. In addition to William, the firstborn, their children included Margaret, born in 1652, and Richard, in 1656. The Penns lived in London until, shortly after William's serious bout with smallpox



Penn, pastel portrait by Francis Place. In the Historical Society of Pennsylvania.
By courtesy of the Historical Society of Pennsylvania

at the age of three, they moved into the Essex countryside near the village of Chigwell, the seat of a renowned grammar school. From his sixth to his 12th year, young Penn pored over Greek and Latin eight to ten hours a day, laying the foundations of his classical education. The strong sense of the past that was thus engendered may have contributed to the feeling of security in his young life.

This stable world was shaken by Admiral Penn's naval defeat in the West Indies in 1655 and brief imprisonment in the Tower of London as punishment. This caused the family to move back to London and then, because of continuing political uncertainty, to Ireland the following year. It was while he was thus uprooted in early adolescence, William Penn later recalled, that "the Lord first appeared unto me . . . about the 12th Year of my Age, Anno 1656," a divine encounter probably fostered by the Puritan influence of Chigwell. In Ireland he heard Thomas Loe, a Quaker itinerant, preach to his family at the Admiral's invitation, an experience that apparently intensified his religious feelings.

Early religious experience

Four years were spent in Ireland, with tutors tending to Penn's education. He was now in closer association than ever with his land-bound father. At the restoration of the Stuart dynasty in 1660, the Admiral was knighted and named a commissioner of the navy. William, who was academically well prepared, the same year entered Oxford, where he demonstrated his piety by rejecting the Anglican formalism of the university. He was fined for refusing to attend chapel and dismissed in his second year. Admiral Penn was outraged by this conduct. Determined to thwart his son's religiosity, he sent the 18-year-old on a grand tour of the Continent and to the Protestant college at Saumur, in France, to complete his studies. The Admiral's purpose was to prepare the young man for high public office, but William devoted most of two years to the study of theology at Saumur. Summoned back to England, he entered Lincoln's Inn and spent a year reading law before his studies were cut short in early 1666 by the London plague.

His limited legal knowledge was put to use in managing the family estate in Ireland. Barely 21, he accepted his father's decision that he follow this worldly course, but then he crossed paths again with Thomas Loe. At a meeting in Cork, Penn heard Loe preach on the text "There is a Faith that overcomes the World, and there is a Faith that is overcome by the World." He committed himself from that moment to the Quakers, religious radicals who were reviled by respectable society and subject to official persecution.

Quaker leadership and political activism. Bent on converting a reluctant world to righteousness, the first Quakers would stop at nothing to preach the Truth. Penn joined the sect in this spirit and was imprisoned four times for publicly stating his beliefs in word and print. He published 42 books and pamphlets in the seven years immediately following his conversion, including the famous *No Cross No Crown* in 1669, a defense of his beliefs written in the Tower of London. Penn's willingness to withstand the harshness of his confinements, when he might have used his father's position to obtain release, as well as his missionary efforts on the Continent, convinced other Quakers of his commitment. He was on the way to a position of leadership at a time when the nature of the movement was changing.

In reaction to official persecution, George Fox, the founder of the Society of Friends, was imposing discipline on his people to control the excesses of radical Quakers and, thus, to give the government less provocation. Penn's contribution to the transformation of the society was a tireless advocacy of liberty of conscience, the fundamental principle of his social thought and action. This liberty could be realized only through an official policy of tolerating dissent; but to seek toleration was to admit, implicitly, that there might be many truths. Both Fox and Penn were accommodating "a Faith that overcomes the World" to the necessity of social survival.

Advocacy of religious liberty and toleration

Penn's move from zeal to moderation reflected not only a pragmatic temperament but also a change in his personal life. Though his academic training was extensive, Penn's theorizing was almost always a response to particular problems. His views on toleration were expressed not in a philosophical treatise but in a number of timely tracts. An example is "The People's Ancient and Just Liberties Asserted, in the Trial of William Penn and William Meade" (1670), a vindication of the independence of the jury, which Penn defended in a court case that stands as a landmark in English legal history. In his own life, the death of Admiral Penn in 1670, of his brother Richard in 1673, and of four of his eight children in infancy after his marriage to Gulielma Springett in 1672 were sorrowful and sobering experiences. Penn reached middle age more mellow than millenarian, content to work within the system. Inevitably, this led to political action.

Again the Quaker leader turned to his pen, and he began lobbying as well. He not only presented his case to men of influence but also prevailed upon fellow Quakers to compile a record of persecutions (institutionalized in the Meeting for Sufferings) and lay petitions before Parliament. His entrance into politics coincided with the emergence of the Whigs, whose advocacy of legislative power and religious toleration was naturally appealing to him. His *England's Great Interest in the Choice of this New Parliament* (1679) was one of the first clear statements of Whig policy. Among his personal acquaintances, however, were Tory leaders, the King (Charles II), and his brother, the Duke of York (later James II).

Founding and governorship of Pennsylvania. Within the Society of Friends there was resentment to the worldly militance Penn represented; older Quakers stymied his strategy. Nor did Penn's foray into politics serve him well when the Whigs faded before a Tory resurgence. Frustrated in England, Penn turned toward America as a means of institutionalizing his political and religious principles. Friends had been preaching across the Atlantic since 1655. By 1674 two Quakers had come into possession of West New Jersey, only to engage in a conflict over ownership that Penn was asked to mediate. Convinced of the virtue of single proprietorship, he applied to the crown for a grant of land in the New World, drew upon his connections with men of rank, and received Pennsylvania through a charter that named him the supreme governor.

Penn did not lay full claim to this privilege. Torn between the advice of Whigs and liberal Quakers, who advocated a vigorous popular legislature to offset executive power, and the need to lure politically conservative financial backing, he drew a frame of government that limited republicanism. Power was concentrated in the governor and his council, but he also guaranteed liberty of conscience and promised that the freemen could alter the system if they so chose. The response to this gesture was a political fray in Pennsylvania that was wilder than Penn

Problems of colonial government and independence

could have imagined—the consequence of the Quakers' quest for legislative power.

The turbulence in Pennsylvania caught the attention of English officials, who were already suspicious of such Quaker peculiarities as pacifism and the refusal to take oaths. As proprietor, Penn divided his energies between the demands of provincial Quakers and the threats of the home government. Because the latter were more crucial to the autonomy of his colony, he was forced to spend most of his time in England, living in Pennsylvania less than four years.

Penn's first visit (1682–84) was cut short by a boundary controversy with Maryland, which he settled advantageously after a hasty return to London. Simultaneously, James II took the throne and began to accelerate the absolutist policies of his older brother. Penn was personally sympathetic to James, persuaded by his professions favouring toleration and fearful that Pennsylvania's independence would be sacrificed to the Dominion of New England. Penn thus became implicated in a reign abruptly terminated by revolution. Ironically, the establishment of parliamentary ascendancy and religious toleration by statute in 1689 was coincident with Penn's alienation from the government. Pennsylvania was briefly brought under royal rule (1692–94), primarily because Quakers would not cooperate in the colonial war effort.

By the mid-1690s, when Penn emerged from hiding and began rebuilding his political alliances, he found both the King and Parliament intent on strengthening the crown's colonial policy and, concomitantly, undermining the autonomy of the chartered and proprietary colonies. Penn, who led the fight against passage of the Navigation Act of 1696 that embodied the new imperial designs, emerged as spokesman for the continued independence of these colonies. His own was commercially and strategically the most important and, therefore, the most visible and open to attack.

The provincial Quakers, who were indifferent to Penn's situation, continued to hammer away at executive power, even after he granted the Pennsylvania assembly pre-eminence in the Charter of Privileges, the climax of his second visit (1699–1701) to the colony. Discouraged by this lack of gratitude, shocked by the discovery that his personal steward had defrauded him of thousands of pounds, and disappointed by the un-Quakerly behaviour of his eldest son, William Penn, Jr., he initiated negotiations in 1702 to sell the "holy experiment" (while simultaneously defending it). On the verge of concluding an agreement of sale in 1712, Penn suffered a crippling stroke, lingering on pathetically until his death six years later, on July 30, 1718. The province, in trusteeship because of his indebtedness, passed to Hannah Callowhill Penn, whom Penn had married in 1696, two years after the death of his first wife. In the 18th century, Pennsylvania was distinguished by its unicameral legislature and its cultural diversity, late testimony to Penn's principles of republicanism and toleration.

Last years

BIBLIOGRAPHY. There is a short bibliographical essay on Penn biographies in JOSEPH E. ILLICK, *William Penn the Politician: His Relations with the English Government* (1965); and a bibliography of his writings in CATHERINE O. PEARE'S readable *William Penn: A Biography* (1957). WILLIAM I. HULL, *William Penn: A Topical Biography* (1937); and MARY M. DUNN, *William Penn: Politics and Conscience* (1967), are authoritative sources, the latter focussing on the Quaker leader's ideas.

(J.E.I.)

Pennsylvania

One of the original 13 American colonies, the Commonwealth of Pennsylvania, as it styles itself, is the third most populous state in the nation. Since its founding in 1681, its geographical position has made it the Keystone State, integrating first the older colonies of the Northeast and the South and later the East and the developing territories and states of the Middle West. The role remains in evidence, with the state polarized by two great metropolitan areas: Philadelphia lies athwart the vast population belt stretching along the seaboard from Maine to Virginia, and Pittsburgh is the beginning of the booming industrial belt reaching westward across the Great Lakes plains to Chicago and Milwaukee. Its capital, Harrisburg, nestles in the foothills of the Appalachian Mountains.

An overview of the state

The nearly 11,800,000 Pennsylvanians counted in the 1970 census can look back over almost 300 years during which their state has played a pivotal role in the nation's history. An oasis of religious liberty as a colony, it played host in Philadelphia first to the meetings that led to the Revolution and to the formation of the new American government and then, from 1790 to 1800, to the government itself. Gettysburg saw one of the most crucial battles of the Civil War, and Titusville the world's first oil well. While Philadelphia's commercial and financial institutions long were providing the backbone of the nation's economy, the state's enormous coal beds were feeding the mills of Pittsburgh, Bethlehem, and other cities, which were to become major suppliers of iron and steel to American industry in peace and war. While Philadelphia was enjoying a dominant role in America's intellectual and artistic life until well into the 19th century, the great ethnic mix of Pennsylvanians was evolving one of the richest and most diversified folk cultures in the nation.

The 45,333 square miles (117,412 square kilometres) of Pennsylvania are bounded on the north by Lake Erie and New York; on the east by New York and New Jersey; on the south by Delaware, Maryland, and West Virginia; and on the west by the Panhandle of West Virginia and by Ohio. Although it is classified as a Middle Atlantic state, it does not touch the Atlantic Ocean at any point. Water, nonetheless, has been nearly as crucial in the state's growth as has the wealth of its earth. The boundary with New Jersey is formed by the Delaware River, on which Philadelphia is the major element in the Pennsylvania–New Jersey–Delaware harbour complex known as Ameriport, the world's third busiest shipping centre. In the northwest, a small panhandle bumps out to separate Ohio and New York, to form a 40-mile waterfront on Lake Erie and to give the state access to the iron-ore barges and other commerce of the Great Lakes. Located at the point where the Allegheny and Monongahela rivers meet to form the Ohio River, Pittsburgh has become one of the nation's busiest inland river ports. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; CIVIL WAR, U.S.; PHILADELPHIA; NORTH AMERICA; and APPALACHIAN MOUNTAINS.)

THE HISTORY OF PENNSYLVANIA

Swedes were the first European settlers in the area that was to become Pennsylvania. Travelling up the Delaware from a settlement at the present site of Wilmington, Delaware, Gov. Johan Printz of the colony of New Sweden established his capital on Tinicum Island in 1643, within the boundaries of modern Pennsylvania. Other Europeans, principally the Dutch, established trading posts within Pennsylvania as early as 1647, although the Swedes remained at Tinicum until 1655. In that year, rivalry and fighting between the Dutch and the Swedes led Peter Stuyvesant, governor of New Netherland, to seize New Sweden. Dutch control of the region ended in 1664, when the English seized all of New Netherland in the name of the Duke of York.

Swedish and Dutch settlements

The Quaker colony. In March 1681 King Charles II of England signed a charter giving the region to William Penn in payment of a debt of £16,000 owed by him to Penn's father, Adm. Sir William Penn. The charter, officially proclaimed on April 2, 1681, named the territory for Admiral Penn and included also the term *sylvania* ("woodlands"), as the younger Penn requested.

William Penn intended that the colony should provide a haven of religious tolerance for his fellow Quakers. While still in England, he drew up the first of his "frames of government" and dispatched his cousin, William Markham, to establish claim to the land and also to establish the boundaries of what became the city of Philadelphia. Penn arrived in 1682 and called a General Assembly to discuss the first Frame of Government and to adopt the Great Law, which guaranteed freedom of conscience in the colony. Under Penn's influence, fair treat-

ment was accorded the Indians, who responded with friendship in return. When Penn returned to England in 1684, the new Quaker province had a firmly established government based on religious tolerance and government by popular will.

Colonial growth. The next century was a period of great expansion and of turmoil for Pennsylvania. Its interior included land claimed by the French, and as time went on the Indians became increasingly hostile to the expansion of settlements to the west and north. Much of the fighting of the French and Indian War (1754–63) took place in Pennsylvania. Here the young George Washington began his trip into the Ohio Valley to warn the French to leave; later, it was in Pennsylvania that the English general Edward Braddock suffered defeat at the hands of the French and their Indian allies.

Site of battles and Revolutionary ferment

For many Pennsylvanians, the period following these conflicts marked growing dissatisfaction with British rule. Official limitations on westward expansion, especially as established by proclamation in 1763, were imposed to pacify the Indians, but Pennsylvanians nonetheless pressed westward over the Alleghenies. Outposts such as Ft. Pitt (Fort Duquesne under the French and now Pittsburgh) became settlements vital to the growing flow of trade from the opening lands to the west.

By the eve of the American Revolution, Pennsylvania had become a keystone state geographically and a centre of military, economic, and political activity. The first (1774) and second (1775–83) continental congresses met in Philadelphia, the Declaration of Independence was signed there, and after the war the city became the capital of the short-lived confederation and of the fledgling United States government.

Early years as a state. In 1790 a new state constitution replaced the unicameral legislature of the Revolutionary period with a bicameral one and a fairly strong governor. In the next 70 years, farm equipment was mechanized, roads were improved and extended, canals were built, and railroads spanned the state, all combining with the economic strength of thrifty Philadelphians to make Pennsylvania a major commercial power. Beginning in 1820, important mining companies were formed to exploit Pennsylvania's deposits of hard and soft coal, and in 1859 Edwin L. Drake drilled the nation's first successful oil well at Titusville. During this same period the state became a leading producer of textiles, ships, lumber, tobacco, and, most important, iron and steel.

The Civil War and its effects

The Pennsylvania Emancipation Act of 1781 had pledged the gradual abolition of slavery in the state. Once the Civil War broke out, Pennsylvania became once again a centre of military and political activity. The southern boundary of Pennsylvania, ratified in 1769, was the Mason and Dixon Line. It became the dividing line between the slave and the free states during the Civil War. At Gettysburg the Union Army won one of the most decisive victories of the war against a Confederate force led by Gen. Robert E. Lee.

Emergence of the modern state. With the end of the Civil War came a period of great economic, industrial, and population expansion in Pennsylvania. Until well into the 20th century, Pennsylvania was the second most populous state in the nation. In 1873 the state passed its fourth constitution; with amendments, the document has survived to the present day. In 1898 construction of a new, modern, state capitol building (replacing the earlier structure burned down the previous year) was begun at Harrisburg, which had functioned as the capital since 1812. The new building was completed in 1908.

In the 20th century, Pennsylvania has retained its position as one of the nation's wealthiest and most powerful states. Its mineral wealth was more fully exploited as the nation's industry became increasingly dependent first on coal and then on petroleum. In both world wars, Pennsylvania's heavy industries were major suppliers of iron and steel, arms, and machinery.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Topography and river systems.* The mountains and rivers dominating Pennsylvania's natural landscape have had a major influence in determining its use by man. The man-made boundaries that made it nearly rectangular in shape are totally decimated by the Allegheny section of the Appalachian Mountain system. Running across the state from southwest to northeast, it leaves only a triangle of Piedmont and Coastal Plain in the southeast, with Philadelphia and environs at its extreme point. The Allegheny Front and Mountains form an up-and-down diagonal spine across the centre of the state, while northward and westward a rugged plateau region falls almost at lakefront into the Lake Erie Lowland.

Surface features

In the eastern part of the state the Delaware is fed mainly by the Lehigh and Schuylkill. The Susquehanna, draining the largest section of the state, is a wide, shallow stream that meanders finally into Maryland and Chesapeake Bay. The Ohio system in the west has several tributaries within the state during its 980-mile course by way of the Mississippi into the Gulf of Mexico. Minor systems lead into Lake Erie in the northwest and the Potomac from the southwest.

Physiographic regions. From southeast to northwest, Pennsylvania takes in parts of seven physiographic regions. Inland from the fertile Coastal Plains around Philadelphia, the Piedmont is an area of rich limestone soils and prosperous farms that are the mainstay of the state's agriculture. The Reading Prong, part of the New England Upland, is a small tongue of a ridge in the northeast. In the south central region, the northernmost finger of the Blue Ridge extends into the state, its topography having played a large part in the Battle of Gettysburg.

The Appalachian Ridge and Valley region, running southwestward across the state to the west of the Piedmont, comprises a series of parallel ridges and valleys, including the long and wide Lehigh, Lebanon, and Cumberland valleys, which were important in America's early westward movement. From this rises the geologically newer Appalachian Plateau, whose rough-and-tumble aspect of small valleys and broken, ragged ridge lines covers more than one-half of the state, from the southwestern and western borders to the Delaware in the northeast and on into New York. Plateaus up to 2,000 feet high cover northern Pennsylvania, including Mt. Davis, at 3,213 feet (979 metres) the state's highest point. The home of the major coal-mining operations, its landscape is dotted with mines and mining towns. The sandstone soils of the mountains and plateaus are poor agriculturally. In the northwestern bump lie the Lake Erie Lowlands, a flat area of gravelly but fertile soils that once lay at lake or ocean bottom and now is a producer of grapes, wine, and potatoes.

Climate. The climate is far less variable across the state than is the landscape. Pennsylvania has a continental climate characterized by wide fluctuations in seasonal temperatures, with prevailing winds from the west. The frost-free period is longest in the southwest, the Ohio Valley, and the Erie Lowlands. Higher areas have from three to five months a year without frost. Coupled with some 42 inches (1,067 millimetres) of rain annually, this provides adequately for the cultivation of temperate-zone crops and vegetation. In the north, average temperature ranges from about 22° F (−6° C) in January to about 68° F (20° C) in July. In the southeast and western portions of the state, averages are several degrees higher.

Vegetation and animal life. Pennsylvania marks the transition zone between the northern and southern forests of North America. Over one-half of the state remains wooded, although only small areas are virgin forest. Hickory, locust, maple, cherry, beech, black walnut, elm, chestnut, poplar, ash, sycamore, willow, linden, white and yellow pine, and other conifers are all common in various parts of the state. The southeast has some trees common to the Mississippi Valley.

Pennsylvania's abundant wildlife makes it a leading state for hunting. Unprotected predatory species include red and gray foxes, wildcats, and weasels. Most large animals gradually have been eliminated by man. Species once common—*e.g.*, the black bear and the Canada lynx—are now rare, although skunks and racoons, pheasants and other game birds, and, in some areas, deer are still

found in abundance. Over 150 species of fish, including trout, bass, chub, pickerel, and white carp, are found in the streams.

Human imprints. The Pennsylvania landscape under human settlement is as varied as its physical geography. Although nearly all the land is forested or under cultivation, Pennsylvania is considered an urban, industrial state. Urban areas, however, comprise widely separated enclaves.

Rural and urban Pennsylvania

The rural areas of Pennsylvania, especially the farmlands, lie mostly in the southeastern counties. In the southwest and in the Susquehanna River Valley, pasturelands dot the landscape. Generally, several small villages, each consisting of a general store, a church, a gas station, and a few houses, are located within a short distance of a larger town that serves the surrounding farms. In the mining regions, the town structure is often completely dependent on the mines. Company towns in which almost everything—including the stores and the homes of the miners—is owned by the company can still be found. With such total dependence on a single company, a town may become a ghost town if a company discontinues mining operations.

For the most part, the urban centres developed along with specific industries. Reading is known as a textile and hosiery town, Hershey as a chocolate town, Pittsburgh and Bethlehem as steel producers, Erie as a Great Lakes port, and the Wilkes-Barre and Scranton areas as centres of the coal industry. In each case, the city extends its influence into the surrounding areas as a supplier of goods and as the determiner of artistic and cultural life.

THE PEOPLE OF PENNSYLVANIA

Four major Indian groups with tribal settlements occupied Pennsylvania at the time of European incursion. These included the Delaware, or Lenni Lenape, in the Delaware River Basin; the Susquehanna along the Susquehanna in Pennsylvania and Maryland; the Shawnee along the Susquehanna and near the present site of Easton; and various segments of the Five Nations of the Iroquois. With their bitter and bloody struggle to retain their land, the Indians were a very powerful force in the early history of Pennsylvania, but all were gradually pushed north or west as European settlers prospered. Fewer than 2,500 remain today in scattered locations.

The early Europeans. Although the Swedes and Dutch arrived in Pennsylvania first, the English Quakers were by far the most important early group that settled along the Delaware River. With a secure hold on the region, the Quakers quickly turned the Delaware counties—especially Penn's three original counties, Philadelphia, Chester, and Bucks—into a thriving commercial region.

Beginnings of ethnic diversification

Penn's policies of toleration and his experiments in democratic forms of government encouraged other groups to settle in Pennsylvania in large blocs. Germans from the Rhineland settled in numbers in the inland counties of Lancaster, Lehigh, Berks, and Northhampton. By the time of the Revolution, the Pennsylvania Dutch, as they had become known, constituted fully one-third of the colony's population. They turned their land into one of the richest and most productive farming areas in the world and made their customs, cooking, and artwork famous.

Small subgroups of the German community are important components of Pennsylvania's population in their own right. Largely religious, they include the Amish, the Mennonites, the Moravians, the Schwenkfelders, and the Dunkers. The Amish have attracted special attention because of their old-fashioned dress and educational methods and preference for living apart and farming without modern machinery.

The third major ethnic group to settle in Pennsylvania were the Scots-Irish. Pressing west past the areas of English and German settlement, this group settled along the western frontier during the middle of the 18th century. Along with the true Scots, with whom they intermingled, the Scots-Irish comprised about one-fourth of the entire population of the colony at the outbreak of the Revolution. In addition to the three main groups listed above, there were small numbers of French Protestants, or Huguenots, Welsh, Cornish, and Irish, and, of course, the Dutch and the Swedes. Through the early years of the 20th century, immigration to Pennsylvania continued to be largely German and Irish.

Later immigration. Until the 1860s, there was open land on the western frontier of the state. During this period the religious character of the commonwealth changed: the Quakers and the Pennsylvania Dutch lost numerical strength to predominantly Roman Catholic immigrants, and Presbyterianism and Methodism gained footholds in the west. In the 20th century, the most important European immigrants to enter the state have been of southern and eastern European origin, particularly Italians and Slavic peoples who provided manpower for the state's burgeoning industry and manufacturing. During and after World War II, in response to the needs of the economy, black Americans from the rural south began to settle in Pennsylvania in increasing numbers. For the most part, the newly arrived black Americans, like the Italians and Slavs, have tended to settle in urban, industrial areas.

Although Pennsylvania is a melting pot of many ethnic groups, the Quakers and the Pennsylvania Dutch continue to play roles in the culture and tradition of the state out of proportion to their actual numbers.

Contemporary demography. The census of 1970 recorded a 4.2 percent population increase in the preceding decade, a figure well below the national average of 13.3. Estimates for 1985 are well over 13,000,000, although the state will probably follow the national trend out of the cities into surrounding suburban areas. The greatest population density is found in Philadelphia and Allegheny counties, the latter including Pittsburgh. In 1790, when the first federal census was taken, 89.8 percent of the population was classified as rural. The turning point came in 1900, when for the first time the urban population, at 54.7 percent, outstripped the rural. In 1970, the urban figure was 71.5.

THE STATE'S ECONOMY

An economic overview

With fertile farmland, 16,000,000 acres of forest reserves, seemingly inexhaustible supplies of coal, many navigable waterways, and an economically strategic location on the Eastern Seaboard, Pennsylvania has always been one of the most prosperous states in the nation. As the exploitation of its primary wealth—soils, minerals, and forests—proceeded in the 18th and 19th centuries, Pennsylvania held a place second only to New York in wealth and population.

Pennsylvania still produces more steel than any other state, and its manufacturing industries are as diversified as those of any state in the Union. In the 20th century, however, the changing nature of modern industry, the use of fuels other than coal, and the growth of the rest of the country have left the Keystone State in third place in wealth (as measured in terms of gross personal income) and population. It is surpassed in manufacturing employment only by New York and in value added by manufacturing only by New York, California, Ohio, and Illinois. Technological changes have left their mark on the economy, and pockets of poverty and unemployment remain.

Sources of state income. *Mineral production and processing.* Nearly $1,000,000,000 worth of products come from Pennsylvania's mines, quarries, and wells each year. The largest share is from coal, in the production of which Pennsylvania ranks second only to West Virginia. Nearly 80 percent of the soft-coal resources remain. Oil is a minor resource in the west, but oil piped from Louisiana, Oklahoma, and Texas and brought by tanker from Venezuela and the Middle East is processed in the huge refineries lining the Delaware. The state ranks thirteenth nationally, however, in production of natural gas. Small deposits of iron ore have been discovered, along with limestone, silver and gold, copper, cobalt, and zinc. Salt beds recently were discovered in the west.

The steel industry. Pennsylvania's first ironworks, for the production of wrought iron, was built in Berks County in 1716. The eastern region long was the centre of pro-

duction, especially after large-scale exploitation of its anthracite fields began. The Pittsburgh area became increasingly important as bituminous coal and coke replaced anthracite and as the Great Lakes became the route for barges from the rich iron-ore beds around Lake Superior. Only in recent years has the complex of mills at the southern tip of Lake Michigan, in northwestern Indiana and northeastern Illinois, begun to exceed the Pittsburgh area in total steel production.

Manufacturing. Pennsylvania has always been known for the variety of its manufactures. In 1969 its industrial plants, with some 1,500,000 workers, produced more than $40,000,000,000 worth of goods. In addition to steel, Pennsylvania leads the nation in the production of flat glass, clay refractories, truck trailers, nonferrous-metal castings, and inorganic pigments. On the basis of the value of the manufactured products and the number of employees, the leading industries are metal and metal products and textiles. Other major industries include chemicals and allied products, leather and rubber goods, lumber and its remanufacture, and tobacco and its products. Industry is unevenly distributed, its greatest concentration being in Philadelphia, Allegheny, and Montgomery counties.

Retailing. The retail-sales industry is also important. John Wanamaker opened the nation's first department store in Philadelphia in the 1870s. Frank Woolworth opened his first successful five-and-ten-cent store in Lancaster in the same period. Other famous Pennsylvanians in the history of retailing are R.H. Kress, S.S. Kresge, C.G. Murphy, Newberry, and W.T. Grant, all of whom founded chains bearing their names.

Agriculture. Pennsylvania still has a sizable farm population. About one-third of the state is under cultivation in some 75,000 farms. The greatest farm acreage remains in Lancaster, York, and Berks counties in the southeast, which have the state's richest soils. Dairy products account for almost 50 percent of agricultural income, with milk the main product, followed by eggs and dairy cattle. Pennsylvania is also the leading American producer of mushrooms, cigar-leaf tobacco, sausage, scrapple, pretzels, potato chips, plantation-grown Christmas trees, state-graded apples, and quality-graded eggs. It is the second-largest national producer of ice cream and frozen dairy products and ranks among the top five states in the production of milk, eggs, sour cherries, apples, grapes, and maple syrup.

Major farm products and crops

Economic management. In 1956 the Pennsylvania Industrial Development Authority was set up to assist in the redevelopment of areas plagued by chronic unemployment. The greater part of the available funds is lent to "growth" industries. Two development corporations, the Pennsylvania Development Credit Corporation and the Southeastern Pennsylvania Economic Development Corporation, provide funds for equipment or working capital to business in the Philadelphia area.

Pennsylvania's Department of Forests and Waters is engaged in the protection and restoration of over 1,500,000 acres of forests under its control. The soil- and water-conservation division in the Department of Agriculture plays an advisory role for farmers and city dwellers. Soil-conservation districts have been established in 64 of the 67 counties.

Transportation. Pennsylvania's three major ports—Philadelphia, Pittsburgh, and Erie—are supplemented by others along the Delaware, the Susquehanna, and the channellized Inland Waterway System, which includes the Allegheny and Monongahela as well as the Ohio. Two-thirds of Philadelphia's imports are petroleum, but it has facilities for handling ores, coal, molasses, and general cargo as well. Part of Erie's facilities have been converted to the building of ore carriers for Great Lakes service.

Pennsylvania's highway system, the fourth largest in the nation, includes the Pennsylvania Turnpike, a four-lane toll road joining New Jersey and Ohio that was a model for the nation in modern superhighway construction. It is paralleled to the north by Interstate 80, also a superhighway spanning the state.

More than 8,000 miles of rail lines link the state's industrial communities. The Penn Central is the nation's largest railroad, and other lines are vital economic links throughout the Middle Atlantic region. Numerous smaller lines serve smaller areas. Philadelphia and Pittsburgh have major commercial airports, though 14 other cities are served by scheduled flights.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *The state level.* Under the constitution of 1873, the executive branch of state government comprises the governor and lieutenant governor, the attorney general, the auditor general, and the members of the Cabinet. The governor is elected for a four-year term and can succeed himself for one additional term. Other elected officials include the auditor general, the state treasurer, and the secretary of internal affairs. The governor appoints the attorney general and the superintendent of public instruction, subject to Senate approval. Among the main powers of the governor are his rights to return bills for reconsideration and to veto portions of appropriation bills. The General Assembly is made up of a Senate and a House of Representatives. Senators are elected for four-year terms and representatives for two-year terms.

Functions of the executive, legislature, and judiciary

A unified judicial system comprises the Supreme Court, Superior Court, Commonwealth Court, and courts of common pleas, community courts, municipal and traffic courts in Philadelphia, and justices of the peace and such other courts as may be established by the General Assembly. The seven justices of the Supreme Court and judges of the Superior Court are elected for 10-year terms; justices of the peace and of the municipal and traffic courts of Philadelphia for six-year terms.

Local government. Under the constitution, local government is provided for by the General Assembly, which classifies cities, counties, boroughs, school districts, townships, and special authorities by population. All Assembly provisions must be uniform throughout each class, and no legislation may be directed toward one locality. As the only first-class city, however, Philadelphia becomes the target of legislation aimed at this class. It, with Pittsburgh and Scranton, second-class cities, has a strong-mayor government. Smaller cities have several governmental forms, while boroughs have elected councils and weak mayors.

Pennsylvania has two classes of townships and four classes of school districts. Local authorities are special units that were set up in 1933 to circumvent constitutional restrictions preventing local units from incurring debt. The cities, boroughs, and townships, as all-purpose governmental units, carry out all usual public service. A constitutional amendment in 1968 allowed all local governments to opt for home rule.

Political life. Between the Civil War and the 1920s, state government was, with a few breaks, in the hands of a succession of Republican administrations that made the statehouse a seat of "boss rule." A similar dynasty persisted in Philadelphia until the late 1940s. Pennsylvanians still tend toward the Republican Party in state elections, although the Democratic majorities in the cities exercise considerable influence in Harrisburg and, as in 1970, may carry the state. Pennsylvania generally exhibits the characteristic American pattern of rural conservatism and urban liberalism, although locally such labels may bear no necessary relation to particular parties.

The social milieu. The social and economic divisions in Pennsylvania are wide, from the affluence of Philadelphia's suburban Main Line and Pittsburgh's Oakland section to the black and Puerto Rican ghettos ringing the central city, from the lush and prosperous farms of Lancaster County to the pathetic ghost towns of Appalachia. Philadelphia and other large cities face all the grave problems of contemporary cities: the blighting of human life and the physical environment, inadequate finances, social instability and racial hostility, and a struggling educational system. The day of the isolated farm or mining community is over, but unemployment in mining regions is chronic. The federal Appalachian Regional Commission, established in 1965, is attempting to encourage economic redevelopment throughout the 12 states involved.

Statistically, however, Pennsylvanians live well, with a work force of nearly 5,000,000 and unemployment generally below the national average. Incomes and costs of living vary widely both from city to city and also from city to rural areas. Income in Pittsburgh is generally the highest, due largely to the high percentage of union employment, whereas the unemployed or unemployable of the mining regions and the city slums may live well below the poverty line.

Education. Consolidation of public-school districts in recent decades has reduced their number, the eventual goal to be fewer than 500. Districts provide physical facilities, teachers, and textbooks and levy taxes and issue bonds. The state Department of Education oversees statewide standards for teacher certification, curricula, and approximately half-cost subsidies to districts.

Institutions of higher learning

Pennsylvania is the home of more than 125 colleges and universities in addition to a state system of two-year community colleges. Pennsylvania State University, in University Park, is the publicly supported land-grant institution; it has 21 branch campuses throughout the state. The University of Pennsylvania, in Philadelphia, was originally a charity school at its founding in 1740 but was transformed by Benjamin Franklin and others into an academy. In 1765 it opened the first medical school in North America. Today, among its most notable divisions are the Wharton School of Finance and Commerce and the University Museum, a leading sponsor of archaeological expeditions and research throughout the world. Though privately endowed, it receives considerable state aid.

State aid is given also to Temple University (1884), in Philadelphia, whose programs are geared heavily to the urban setting, and to the University of Pittsburgh (1787). Philadelphia is still a major centre of medical education, while in Pittsburgh the Carnegie–Mellon University, formed in 1967 by the merger of the Carnegie Institute of Technology (1900) and the Mellon Institute (1913), makes that city a centre of scientific studies. Other schools with major reputations include Bryn Mawr College, for women (1880); Haverford College (1833) and Swarthmore College (1864), both Quaker schools; and Villanova University (1842), a Roman Catholic institution—all in the Philadelphia vicinity. Dating from the 18th century are Dickinson College (1773), in Carlisle; Franklin and Marshall College (1787), in Lancaster; and Washington and Jefferson College (1787), in Washington.

Health and welfare. More than one-half of Pennsylvania's budget is allotted to education, welfare, and highways. Hospitals and centres throughout the state deal with problems of chronic illness, old age, mental retardation, and the like, and local governments are assisted in welfare payments in a broad range of categories. Urban-renewal programs also receive state and federal funding.

CULTURAL LIFE AND INSTITUTIONS

In colonial times, Philadelphia was the focus of the nation's intellectual, cultural, and political life. As Pennsylvania grew and prospered, Pittsburgh and other, smaller cities also became centres of the arts.

The arts. Two of the nation's major symphony orchestras are located in Pennsylvania. The Philadelphia Orchestra, under conductors such as Leopold Stokowski and Eugene Ormandy, has become world famous, as has the Pittsburgh Symphony Orchestra, under such conductors as Victor Herbert, Fritz Reiner, and William Steinberg. Philadelphia's Academy of Music provides a home and concert hall for its orchestra, and the world-famous Curtis Institute of Music, founded in 1924, is one of the world's leading conservatories. In addition, many community orchestras perform throughout the state. Noted all over the world is the religious music of the Moravians. At their cultural centre, Bethlehem, the Bach Choir's Bach Festival attracts music lovers every May from many states.

Philadelphia is the home of one of the world's finest art museums, the Philadelphia Museum of Art, and of the Rodin Museum. The Pennsylvania Academy of the Fine Arts offers a base for teaching and study as well as for display. The Carnegie Institute's museum and library in Pittsburgh also houses many famous paintings. Among famous artists whose homes were in Pennsylvania are Charles Willson Peale and his two sons, Raphaelle and Rembrandt, Benjamin West, Mary Cassatt, Thomas Eakins, and the Wyeths, N.C. and his son Andrew. The Calder family of Philadelphia is famous in sculpture. Alexander Milne Calder's giant statue of William Penn stands atop City Hall, and his grandson, Alexander Calder, gained renown for his free-form mobile sculptures.

The visual arts: painting and sculpture

Philadelphia long vied with New York as the theatrical capital of the young nation, though today it is basically a tryout town for shows on their way to Broadway. Summer theatres have proliferated across the state, especially in the many resort areas. Temple and Carnegie–Mellon universities offer major programs in theatre. The Pennsylvania Ballet offers entertainment both within and outside the state. Among well-known writers of the 20th century with Pennsylvania origins are Pearl S. Buck, Rachel Carson, James Michener, Christopher Morley, John O'Hara, Mary Roberts Rinehart, and John Updike.

Museums. In addition to its art museums, Pennsylvania enjoys many other museums, both general and specialized. The Franklin Institute in Philadelphia is a museum of applied science. The William Penn Memorial Museum and Archives Building in Harrisburg, the Pennsylvania Farm Museum of Landis Valley near Lancaster, the Annie S. Kemerer Museum in Bethlehem, and the Mercer Museum of the Bucks County Historical Society (crafts and craftsmen's tools) in Doylestown house fine collections. There are also small museums in Allentown, Reading, Scranton, and Greensburg and historical society museums in Lancaster, York, Reading, Wilkes-Barre, and West Chester, to mention a few. The Pennsylvania Historical Commission, established in 1913 and known since 1945 as the Pennsylvania Historical and Museum Commission, is active in conservation of the state's historical heritage; the Philadelphia Historical Commission oversees that city's many historic shrines.

Folk culture. Pennsylvania has retained strong elements of folk culture among its diverse ethnic groups. The "plain people"—the Amish, Mennonites, and other small sects—have kept their old customs and dress and often resist the use of machinery. In Lancaster County, where the Amish still farm with horses and oxen, their wagons are a familiar sight on the country roads. The folk art and the cooking of the Pennsylvania Dutch are famous, and their brightly painted hex signs adorn their large barns. At Czestochowa, in Bucks County, the largest Polish church in America is evidence of the area's Polish population.

Communications. Public figures such as Benjamin Franklin were responsible for the emergence of newspapers, magazines, and libraries in the colony. Later, the Lippincott and the Curtis publishing families became giants in the publishing world, the Lippincotts as book publishers and the Curtis family as publishers, until very recently, of such magazines as *The Saturday Evening Post*, *Ladies Home Journal*, and *Country Gentleman*. The *Farm Journal*, a Philadelphia publication, is one of the nation's largest farm magazines. Several other major publishers are also headquartered here. In the 1970s a new magazine, *Philadelphia*, was increasing in popularity.

Historic eminence in publishing

Founded in 1870, Philadelphia's *Evening Bulletin* is one of the nation's leading evening newspapers, with its pervasive slogan that, in Philadelphia, "nearly everybody reads the *Bulletin*." The state has also pioneered in the publication of many foreign-language newspapers and in black newspapers.

All of Pennsylvania's cities and larger towns have newspapers, either daily or weekly. The major cities have their own television stations, and most local communities have their own radio stations. Station KDKA, the first commercial broadcasting station in the world, opened in Pittsburgh in 1920.

In the age of television and rapid transportation, the various cultural elements in Pennsylvania see the same movies and television programs, read the same newspapers, and share in a wider national culture.

PROSPECTS

From the ghost towns of Appalachia to the skyscrapers of Philadelphia's centre city, Pennsylvania is a land of sharp contrasts. The traditional and old contrasts sharply with the new; the pastoral countryside adjoins the industrial towns; the old families spread out in suburban comfort, while the newest of the newcomers struggle in the black ghettos.

Today, Pennsylvania has its share of financial, political, and social problems. Political struggles between the state and city governments are not uncommon. And the urgent need of the cities, particularly Philadelphia, for slum clearance, urban renewal, and crime control is reflected in their demands upon the state.

BIBLIOGRAPHY. The best current source material about Pennsylvania may be found in the many publications issued by the state in Harrisburg; these are available at little or no charge. The *Pennsylvania Manual* (biennial) has outlines of the history of the state, its government, and its economic development. This manual, a direct offspring of THOMAS JEFFERSON, *A Manual of Parliamentary Practice* (1801), provides invaluable and voluminous reference material. The *Pennsylvania Statistical Abstract* (annual) is also rich in detail, with graphs and maps to illustrate the statistical tables; as is E. WILLARD MILLER, *An Economic Atlas of Pennsylvania* (1964). For the history of Pennsylvania, one of the most interesting and detailed one-volume studies is SYLVESTER K. STEVENS, *Pennsylvania: Birthplace of a Nation* (1964). Other works of Stevens include *Pennsylvania: The Keystone State*, 2 vol. (1956); *Pennsylvania: Titan of Industry*, 3 vol. (1948), with more detailed economic history; *Portrait of Pennsylvania* (1970), a small volume of photographs with a lively text; and a paperback, *Pennsylvania History in Outline* (1968). A more detailed work on Pennsylvania's geography is RAYMOND E. and MARION MURPHY, *Pennsylvania: A Regional Geography* (1937). Pennsylvania's rivers are featured in three volumes of the "Rivers of America" series: RICHARD E. BANTA, *The Ohio* (1949); FREDERICK WAY, *The Allegheny* (1942); and RICHARD P. BISSELL, *The Monongahela* (1952). Newer books describing travel in Pennsylvania include JAMES and M. CAWLEY, *Along the Old York Road* (1965); GOTTLIEB MITTELBERGER, *Journey to Pennsylvania* (1960); and ERWIN L. PETERSON, *Penn's Woods West* (1958). On Pennsylvania's agriculture, the PENNSYLVANIA HISTORICAL AND MUSEUM COMMISSION has published *Pennsylvania Agriculture and Country Life, 1648–1840* (1950). On government, an authoritative small handbook is *A Citizen's Guide to Local Government*, published by the commonwealth in 1967. On the people of Pennsylvania there are many interesting studies, among them: SOLON J. and ELIZABETH H. BUCK, *The Planting of Civilization in Western Pennsylvania* (1939); JOHN A. HOSTETTLER, *Amish Society* (1963); FREDRIC KLEES, *The Pennsylvania Dutch* (1950); WILLIAM I. SCHREIBER, *Our Amish Neighbors* (1962); the studies of Quakers and Quakerism by FREDERICK B. TOLLES, including *Quakers and the Atlantic Culture* (1960); and PAUL A.W. WALLACE, *Indians in Pennsylvania* (1961).

(C.L.T.)

Pentecostal Churches

Pentecostal churches, a group of Protestant churches originating in the United States in the 19th century and subsequently spreading throughout the world, teach that all Christians should seek a postconversion religious experience called the baptism with the Holy Spirit. They believe that this experience corresponds to the first outpouring or descent of the Holy Spirit upon the 12 disciples of Christ assembled in Jerusalem on the day of Pentecost (seventh Sunday after Easter, or the Resurrection of Christ), as recorded in Acts 1:12–2:4, and that this experience is accompanied by the same sign: the ability to speak in a language entirely unknown and foreign to the speaker, or to "speak in tongues." Pentecostals also hold that a Spirit-baptized believer may receive at least one of the other supernatural gifts that were known to have been in existence in the early church—the ability to prophesy, to heal, or to interpret what is said when someone speaks in unknown tongues. Beyond these unique distinctions, Pentecostal churches reflect those patterns of faith and practice that are characteristic of the Fundamentalist–Holiness branches of Protestantism that also originated in 19th-century America, with their emphases of biblical literalism, conversion, and moral rigour (see also FUNDAMENTALIST AND EVANGELICAL CHURCHES; HOLINESS CHURCHES).

The significance of baptism with the Holy Spirit

Nature and significance. During the early years of the 20th century, many observers prophesied the rapid demise of the Pentecostal movement. On the contrary, however, it has burgeoned and has become one of the fastest growing religious movements in the Western Hemisphere. Though no one knows exactly how many Pentecostals there are in the world, some authorities within the movement estimate that the total membership is between 8,000,000 and 12,000,000.

Despite a common belief in certain doctrines, such as baptism of the Spirit, speaking in tongues, and healing, Pentecostals have not united in a singular denomination. The taproots of this charismatic (spiritually inspired) movement reach back into too many differing personalities, cultures, and religious heritages to allow a normative Pentecostalism to become established. Consequently, it consists of scores of groups, differing in size, church government, ordinances, modes of worship, and a veritable spate of things most other churches consider to be adiaphoristic (nonessential or peripheral to matters of faith): eating pork, wearing jewelry, using tobacco, the grounds for remarriage and divorce, taking medicines, and other such matters. Despite the tendency of Pentecostals to differ and to polarize into groups that, on the one hand, favour ardour, spontaneity, legalism, or hierarchical control and, on the other, to favour order, ritual, tolerance, or independence, the movement as a whole has grown phenomenally.

The sources of Pentecostalism's appeal. *Social appeals.* The initial success of the Pentecostal movement during the early decades of the 20th century, as well as its widespread appeal since then, had its roots in the post-Civil War period. That period was a time of mounting indifference to the claims of the biblical faith. Denominations that had been noted for their revivalistic fervour suddenly changed; they became subdued. Emotional modes of religious expression—enthusiastic congregational singing, spontaneous testimonies, prayer in unison, and extemporaneous sermons on simple biblical themes by lay preachers—gave way to ordered, formal worship services that were conducted by "reverends" who had been trained in homiletics (preaching skills) and influenced by higher biblical criticism. Chautauqua lecture centres (religious, general education, and entertainment programs that originated in Chautauqua, New York) and elegantly appointed sanctuaries replaced camp meeting sites and crude, frame tabernacles.

Changes in American Protestantism in the post-Civil War Period

As the large popular denominations—Methodist, Baptist, Disciples of Christ—that had once been known as the poor man's churches became transformed into churches of the upper middle class, people of limited means began to feel more and more out of place. They yearned to return to a "heart religion" that would satisfy not only their craving for religious certainty but also their emotional, psychological, and physical needs. The newly emerging Holiness and Pentecostal movements fulfilled these needs for many—churchgoers and unchurched alike.

Psychological appeals. Pentecostals radiated a certain assurance. The tentativeness that characterized so much of the belief within the larger historic denominations was absent in Pentecostal churches. Pastors and parishioners of Pentecostal churches had an intense faith in God. They believed in the reality of the supernatural because they had been filled with the Holy Spirit. Because of various charismatic experiences, Pentecostals possessed a certainty of faith, an assurance of salvation, and an ability to live joyously and triumphantly.

Moreover, in an era of industrial expansion with its attendant evils—*i.e.*, urban loneliness and increasing alienation—Pentecostal churches continued to minister to all levels of society, especially to the economically and culturally neglected and socially isolated. To the disinherited, Pentecostalism appears to have had a perennial appeal. According to many observers, Pentecostal churches have been able to provide more opportunity for personal fellowship than is usually found in the traditional church-

es, because their congregations are generally small. A visitor is called "brother" or "sister" and is welcomed warmly.

Worship experiences in storefront churches

Many rural-to-urban migrants found it possible to worship personally and meaningfully in a Pentecostal tabernacle or storefront mission. They had many opportunities for emotional release, which undoubtedly fulfilled a genuine psychological function. Although otherwise inhibited, people who experienced the Spirit's "moving" during a worship service felt very free to participate in it. They could accompany the exuberant gospel songs by rhythmically clapping their hands or tapping their feet; they could respond to the preacher's sermon whenever so moved by shouting "Amens" and "Hallelujahs"; and they also could speak with other tongues or could admonish listeners by prophesying. According to some psychologists, all of these actions had therapeutic effects on the participants.

Faith healing. Another probable source of Pentecostalism's appeal is the emphasis placed on the doctrine of healing. From the very beginning of the Pentecostal revival, its evangelists—Mrs. Mary B. Woodworth-Etter, Charles Price, Aimee Semple McPherson, and, more recently, Oral Roberts—have taught that deliverance from sickness was provided for in Christ's atonement and is, therefore, the privilege of all who have the requisite faith. This effort to minister to the physical needs of people most likely has been responsible for winning many followers to Pentecostalism.

Origins of Pentecostalism. *Speaking in tongues.* Glossolalia, the phenomenon of speaking in tongues, in the apostolic church is known primarily from the writings of St. Luke and St. Paul. After an initial outburst of speaking in tongues, in Jerusalem on the day of Pentecost, the phenomenon appeared later at Caesarea, Ephesus, and Corinth, a city in which prophets, healers, and speakers in tongues were very active. Other post-apostolic instances of glossolalia have been recorded throughout the history of the Christian church: among the Montanists (heretical followers of the 2nd-century Phrygian prophet Montanus), among radical Anabaptists (extreme left-wing Protestants in 16th-century Germany), and among the Camisards (a radical Protestant peasant group from southern France) and Jansenists (a Catholic reform movement) in 17th- and 18th-century France. The gift of tongues was also prolific among the Shakers, a celibate communal religious sect in the mid-18th century United States. During the 19th century, an outburst of glossolalia and some instances of healing occurred, notably in the Scottish preacher Edward Irving's church in London's Regent Square, among the Mormons, and among various groups of Holiness people in the mountains of Tennessee and North Carolina.

20th-century Pentecostal origins. The roots of the modern Pentecostal movements go back to a small religious school called Bethel Bible College in Topeka, Kansas. The director of the institution, Charles Fox Parham, one of many ministers who had been influenced by the Holiness movement, believed that what he saw as a complacent, prosperity-ridden, and coldly formalistic church of the era needed to be revived and that such a revival could only be achieved by another outpouring of the Holy Spirit. He therefore instructed his students, many of whom already held ministerial credentials with Methodist, Friends, or Holiness churches, to pray, fast, study the Scriptures, and await an endowment of Pentecostal blessing and power that would be very much like the events recounted in the book of Acts.

Influence of Charles Fox Parham

On January 1, 1901, a student, Agnes N. Ozman, was the first of Parham's many students to experience a Spirit baptism, and she began speaking in an unknown tongue. Henceforth, most Pentecostals usually declared that glossolalia was the "initial evidence" that one had been truly Spirit baptized.

That recurrence of Pentecost in Topeka also had some prophetic overtones, for those who had been Spirit filled were familiar with the prophecies of the Old Testament Book of Joel ("in those days, I will pour out my spirit"). They had experienced what they saw as a supernatural outpouring of the so-called Latter Rain; they therefore believed that they must be living in the last days. Imbued with this sense of urgency, Parham and his students set out to evangelize the U.S. Southwest before the end came. Their efforts were so successful that by 1905, according to Parham, approximately 25,000 persons in Texas alone had embraced Pentecostal faith and practice. The embryonic charismatic movement was at first significant only regionally: in the tri-state region of Kansas, Missouri, and Texas, and in Alabama and western Florida. The national and international expansion occurred as a consequence of what is recorded as the Azusa Street revival of 1906.

The Apostolic Faith Gospel Mission at 312 Azusa Street in the industrial section of Los Angeles was unpretentious. Its leader, William Seymour, a one-eyed Negro former Holiness preacher, had been exposed to Pentecostal teaching at a Bible school that Charles Parham had established in Houston, Texas. Under his guidance, the old frame building, which formerly had housed a Methodist congregation, became a Pentecostal mecca (or pilgrimage centre). From April 1906 onward, rich and poor, blacks and whites, native Americans and those from abroad, and seekers and skeptics flocked to the Azusa Street Mission to receive spiritual help.

Spiritually energized and convinced that they had been charismatically endowed, scores of men and women spread out to tell of their experiences, extolling the reality and purposiveness of speaking in tongues. Soon storefront missions, modest tabernacles located in sparsely populated rural areas, and upper-story lofts in squalid urban neighbourhoods housing lively groups of Pentecostal believers sprang up in many parts of the world. These congregations represented the effects of a novel revival that was known variously as the Latter Rain, Apostolic Faith, or Pentecostal movement.

Initial unification of Pentecostal Christians

At first, the ultimate in organizational development among Pentecostal Christians was an amorphous sort of "spiritual union." Little if any thought was given to forming a separate "Pentecostal" branch of the Christian Church. Despite the fact that scores of members of the historic Protestant churches embraced the teachings and practices of tongue speaking, healing, and prophesying, they initially had no intention of withdrawing from their own churches to form another denomination. They merely wanted to be agents of reform and revival, helping to rid their churches of formalism in worship, modernism in belief, and worldliness in practice and striving to transform them into vibrant, spirit-filled communities similar to those described in the New Testament book of Acts. They fully expected the "showers of blessing," the prophetically promised Latter Rain of which they had received a few droplets, to fall copiously upon their respective denominations, which would make them wholly Pentecostal.

But the triumphant conquest of the Protestant religious world by Pentecostal ideas during those early years never materialized. In fact, the embryonic movement became the object of rather widespread opposition. Pastors who endorsed Pentecostal practices were relieved of their pulpits by their churches; missionaries who were sympathetic toward the charismatic movement were notified by their mission boards that their financial support was being stopped; and parishioners who practiced speaking in tongues found their names being stricken from membership rolls. Resolutions were passed and anathemas (rigorous ecclesiastical denouncements) were pronounced against those exhibiting Pentecostal practices. Those who desired to retain their charismatic religious way of life found it increasingly difficult to do so within the institutional framework of conventional Protestantism, with the result that many holding such views withdrew from their churches to form new and distinctly Pentecostal groups.

At the outset, most Pentecostals shunned anything that reminded them of the religious bodies from which they had separated. They were particularly wary of the multiplication of, or the concentration of, power in administrative bureaus. They were unwilling to subject themselves to external ecclesiastical control or to be bound by

common practices in matters of faith or polity. The result of such fears was a heterogeneity that was totally ineffectual. What became viewed as false doctrines spread unchecked; self-styled evangelists preyed on congregations that found it difficult to discriminate between sincere and insincere preachers; and financial accounts and records went unaudited. Leading Pentecostals, recognizing that the problems created by unbridled individualism and autonomy far surpassed the evils of organization and bureaucracy, attempted to establish a denomination that was unmistakably charismatic. A unitive Pentecostal movement, however, never developed. Divisions occurred as groups of Pentecostals tended to gravitate toward gifted leaders in their locales or because some demanded universal obedience to their unique tenets of faith or practice. Instead of a singular movement, a variety of Pentecostal groups evolved. They differed—often radically—in matters of belief, liturgy, and organization. Some favoured ardour, spontaneity, legalism, and independency; others preferred order, ritual, tolerance, and hierarchical control.

It was imperative, then, that Pentecostals organize, and they did, although not into one massive body. Most interpreters of Pentecostalism have overlooked this fact, namely, that Pentecostalism is not a singular movement, that normative Pentecostalism is virtually nonexistent, and that, therefore, it consists of a variety of groups that differ—often radically—on many matters of faith and practice.

Varieties of American Pentecostalism. *Churches of God.* In 1896, before either the appearance of Pentecostal phenomena at Topeka, Kansas, or the massive outburst at Los Angeles, at least 100 North Carolinians had received the baptism of the Holy Spirit and had spoken in tongues. They were members of an organization called the Christian Union, the name of which was later changed to the Holiness Church (1902) and, finally, to Church of God (1907). Doctrinally, that denomination is similar to the other groups that emerged from the Holiness movement in its emphasis upon sanctification and personal holiness. Moreover, it is perhaps the only large Pentecostal body that observes foot washing as an ordinance.

Holiness Church origins

The Church of God is administered by a General Assembly, the highest governing body. Composed of ministers and laymen from all recognized churches, it meets biennially to deal with matters of polity or doctrine and to select its chief administrators: four overseers, a general secretary-treasurer, and a director of world missions. Those six executives, who are elected to four-year terms, together with a council of 12 elders, who are elected every two years, constitute the Supreme Council. This is the denomination's highest administrative body and is responsible for handling policy matters in the periods between General Assemblies.

Since its origin, the Church of God (Cleveland, Tennessee), so designated to distinguish it from the many other groups bearing the same name, has experienced a number of schisms. Most of them occurred after A.J. Tomlinson, general overseer from 1909 to 1923, was impeached and removed from office. Tomlinson withdrew but remained in Cleveland and founded a Church of God bearing his name. After his death in 1943, the Tomlinson Church of God split into the Church of God of Prophecy (headed by one son, Milton) and the Church of God, World Headquarters (under the leadership of another son, Homer).

Despite this fragmentation, the Church of God (Cleveland, Tennessee) has experienced substantial growth. It has erected a multi-million-dollar printing plant and a world-headquarters office in Cleveland, Tennessee, and has developed Lee College into a four-year institution of higher learning. In 1951 it effected an amalgamation with the Full Gospel Church in South Africa, thus increasing its membership by 30,000 and its number of churches by nearly 200. Thus, in little more than half a century, the Church of God (Cleveland, Tennessee) has grown from the five-church denomination of 150 members that convened at Union Grove, Tenneseee, in 1907 to a rather respectable size: 2,917 ordained ministers, 3,933 churches, and an inclusive membership of 257,995.

Assemblies of God. With 11,677 ministers, 8,619 churches, and 625,027 members, the Assemblies of God is probably the largest Pentecostal denomination in the United States. Until its formation at Hot Springs, Arkansas, in 1914, Pentecostals in America were governed either by a small association with a dynamic leader at its head, such as Charles Fox Parham or Florence Lewis Crawford of the Apostolic Faith, or they were governed by bodies such as the Church of God and the Pentecostal Holiness Church, which were largely regional. No national organization existed to serve the thousands of Pentecostals who were either not from the South or, if they were, were not from the Holiness groups within American Protestantism. Some Pentecostal leaders were aware of this—especially Eudorus Bell, Howard A. Goss, Arch P. Collins, and Daniel C. Opperman—and they provided the impetus for establishing a denomination that would fuse the heterogeneous Pentecostal assemblies, missions, and tabernacles that were scattered throughout the United States. They called a "General Council of Pentecostal saints" to convene at Hot Springs (April 2–12, 1941). About 120 pastors and evangelists from 20 states and several foreign countries responded. Their primary task was to establish an acceptable system of organization and to formulate a doctrinal statement to which the delegates, coming from such diverse religious traditions, could subscribe. They adopted a simple kind of polity, a curious admixture of Congregational and Presbyterian elements. The Council elected an Executive Presbytery to serve as the central administrative group; this organ was empowered to execute the mandates given it by the General Council and to act for the council in all matters that affected its interest when it was not in session.

Establishment of the Assemblies of God

Except for a pronouncement that "the Holy inspired Scriptures are the all-sufficient rule for faith and practice . . . and we shall not add to or take from them," that first General Council postponed action on the matter of a definitive doctrinal statement. Subsequently, however, a *Statement of Fundamental Truths* was adopted. The document demonstrated that the Assemblies of God are Trinitarian (believing in God as Father, Son, and Holy Spirit) and Arminian (accepting the doctrines of grace and free will as espoused by the 16th–17th-century Dutch theologian Arminius). They also subscribed to two ordinances (Baptism by total immersion in water and the Lord's Supper), held a view of sanctification (becoming holy) that may be described as "progressive," or gradual, rather than "instantaneous" in regard to moral purity, and, finally, were strongly pre-millennial (believing in the doctrine of Christ's Second Advent before the 1,000-year reign of Christ and his saints).

From the outset, the Assemblies of God has been intensely missions-conscious. Its members contribute millions of dollars annually toward the support of more than 800 missionaries who serve in about 70 countries. In addition, the denomination conducts a diversified program of home missions among foreign language groups in America's urban centres, on Indian reservations, in prisons, and among the deaf and the blind.

Its printing plant, the Gospel Publishing House, located in Springfield, Missouri, prints more than ten tons of religious literature each day. It operates two fully accredited colleges of arts and science—Southern California College (Costa Mesa) and Evangel College (Springfield, Missouri)—and a number of regionally located Bible institutes.

Church of God in Christ. Of the several Pentecostal groups that have been established by gifted personalities, the largest probably is the Church of God in Christ. This large Pentecostal denomination, the membership of which is principally within the black community, is largely the result of the dynamic leadership of one individual, Charles H. Mason. At the turn of the 20th century, Bishop Mason was pastor of the Church of God in Christ in Memphis, Tennessee, and he was associated with the Holiness movement. When news of the outpouring of Pentecostal blessing at the Apostolic Faith Gospel Mission

Influence of Charles H. Mason

reached him, he decided to go to Los Angeles. During his five-week stay, he received Spirit baptism and spoke in tongues. He returned to his congregation and began to preach the need for a charismatic experience. Tensions developed, and a non-Pentecostal faction withdrew. Mason proceeded to convene a General Assembly and to declare that the Church of God in Christ was henceforth Pentecostal. At that first Assembly, which met in Memphis, Tennessee, *c.* 1907, Mason was elected general overseer of about 12 churches located principally in Tennessee, Arkansas, Mississippi, and Oklahoma.

For decades, Mason almost single-handedly supervised the activities of a denomination that in the early 1970s claims more than 4,000 churches and a worldwide membership in excess of 2,000,000. In 1933, however, after his administrative responsibilities had apparently become too burdensome for him, Mason, who had the titles "Chief Apostle," "Bishop," and "Elder," relinquished some of his authority to four assistant bishops and later to state overseers. This form of church government, with its adherence to Episcopal authority, is not characteristic of most contemporary Pentecostal groups, which combine elements of Congregational and Presbyterian polity. In doctrinal matters, however, there is little difference between the Church of God in Christ and other charismatic groups, especially the Churches of God and the Pentecostal Holiness Church, that stress perfectionism.

Pentecostal Holiness Church, Inc. At the turn of the 20th century, a number of churches that had been affiliated with the Holiness movement were converted to Pentecostal beliefs and practices and withdrew from their parent bodies. Gradually, three such groups—The Fire-Baptized Holiness Church, the Pentecostal Holiness Church, and the Tabernacle Pentecostal Church—consolidated and formed the Pentecostal Holiness Church, Inc. Unlike most other Pentecostal churches, this denomination permits candidates for Baptism to "have the right of choice between the various modes as practised by the several evangelical denominations." It even permits the Baptism of infants. In matters of polity, it is decidedly less democratic than other Pentecostal bodies. Reflecting its Methodist Episcopal heritage, the denomination is divided into conferences: general, annual, district, and missionary. The General Board of Education oversees the operations of three church-related institutions of higher learning: Emmanuel College, a fully accredited academy and junior college in Franklin Springs, Georgia, close to the denominational headquarters; Holmes Theological Seminary in Greenville, South Carolina; and Southwestern Pentecostal Holiness College in Oklahoma City, Oklahoma, which Oral Roberts, a well-known faith healer and evangelist, helped to establish. Although the expansion of the church has not been spectacular, it has been consistently steady. In 1916 it reported having 192 churches and 8,096 members; by the 1970s it had grown to 1,355 churches and 66,790 members.

Pentecostal Church of God of America, Inc. When the General Council of the Assemblies of God was formed in 1914, many ministers could not understand how a local church could affiliate with a national organization and not lose its sovereignty. Many of them, therefore, refused to unite with the Assemblies of God. When a subsequent General Council adopted a *Statement of Fundamental Truths,* this merely confirmed the suspicions of these independents, who regarded such an action as a regression to the conventional ecclesiasticism from which many Pentecostals had just recently emerged. Consequently, they united to form the Pentecostal Church of God of America, Inc., which demonstrated its independence by establishing rather liberal policies on ordination, church membership, and the issue of divorce and remarriage of the clergy. Though not very active in the area of foreign missions, the Pentecostal Church of God has conducted an extensive ministry among the American Indians. Headquarters for the church, which had 975 churches and 115,000 members by the 1970s, are in Joplin, Missouri, where new facilities were constructed in 1950.

United Pentecostal Church, Inc. During the second decade of the 20th century, the Pentecostal movement was convulsed by a doctrinal controversy that originated on the West Coast of the United States and swept through the heartland of America. The "New Issue," "Oneness," or "Jesus Only" churches, as they were variously called, contended that true Baptism must be "in the name of Jesus" only, rather than in employing the traditional Trinitarian formula. Furthermore, the proponents of this view denied that there were three Persons in the Godhead, asserting instead that there were three manifestations of one Person, namely, Jesus. This teaching, for a time, threatened to decimate the newly organized Assemblies of God, because many of its leaders submitted themselves to rebaptism in "Jesus' name." After the fourth General Council in St. Louis, Missouri, in 1916, when the Assemblies of God articulated a position that was staunchly Trinitarian, the "unitarian" faction withdrew to form the Pentecostal Assemblies of the World, Inc., locating their headquarters in Indianapolis, Indiana, where G.T. Haywood, a black evangelist, had a strong "Jesus Only" congregation. Originally, all unitarian Pentecostals were united in this one interracial body. In 1924, however, the white members asserted that integration hindered their evangelistic efforts and withdrew to form what became the Pentecostal Church, Inc. In 1945 it united with another body of unitarian Pentecostals, the Pentecostal Assemblies of Jesus Christ, to form the United Pentecostal Church, Inc., the largest body of "Jesus Only" Pentecostals. In 1968 it numbered 2,500 churches and reported having 225,000 members. Besides espousing a somewhat unorthodox doctrinal position, the United Pentecostal Church, Inc., takes a more rigorous stand than most other Pentecostal groups concerning such practices as the use of cosmetics, wearing revealing clothing, watching television, or listening to radios. These practices are considered to be worldly and not in accord with the way of life of the believer who is filled with the Spirit.

The rise of the "Jesus Only" movement

International Church of the Foursquare Gospel. Aimee Semple McPherson, a popular and colourful American revivalist preacher, was conducting a revival campaign in Oakland, California, in 1921. According to her own testimony, she saw a vision of four living creatures reminiscent of the one recorded in the Old Testament Book of Ezekiel (1:5–10). To "Sister" Aimee the faces of these four creatures—a lion, a man, an ox, and an eagle—symbolized perfection, and she equated a perfect, foursquare, full gospel with one that would stress four themes: salvation, Holy Spirit baptism, healing, and the Second Coming of Christ. These four themes became the hallmark of her preaching, and, in a short time, hundreds flocked to her meetings; in two years she was able to dedicate Angelus Temple in Los Angeles as the "mother church" of the Foursquare Gospel Association. From 1923 onward, her organization, which initially centred on her and her multi-million-dollar sanctuary (Angelus Temple), grew to national and international importance.

The rise of Aimee Semple McPherson

Students who graduated from her Evangelistic and Missionary Training Institute had launched into the work of establishing home and foreign missions. Branch churches sprang up in the environs of Los Angeles; by 1925, 32 had been established in southern California, and 50 more were appealing to Angelus Temple for permission to affiliate. By 1970 the Foursquare Church had become one of the largest Pentecostal denominations: 1,796 churches and an inclusive membership of 186,782.

The episcopal character of its church government bears the stamp of the founder, around whom everything centred during her lifetime. Aimee Semple McPherson was president of the church and the ruling power and voice of the organization. She held the power of veto; she appointed all officers; she hired all personnel and set their salaries. Her son and successor, Rolf K. McPherson, carries on this tradition. Doctrinally, however, the church is similar to the Assemblies of God, on the ministerial rolls of which Mrs. McPherson's name appeared for several years during her early evangelistic career.

Regional, ethnic, small, and cultic varieties of Pentecostalism. The forenamed Pentecostal denominations have expanded, nationally and internationally, and rep-

resent the majority of Pentecostals in the United States. There are, however, about 24 other groups—many rather obscure—that have had an impact only upon certain sections of the United States, that are racially or ethnically oriented, that have never become numerically impressive, or that remain on the outer fringes because of a preoccupation with individual practices.

Most of the racially or ethnically oriented Pentecostal groups are in general agreement in matters of faith and practice. Perhaps the two most widely known are the United House of Prayer for All People, the handiwork of Bishop Charles Emmanuel "Sweet Daddy" Grace, and the Church of God with Signs Following, the so-called "Mother Church of snake-handling" in the South. They are, to be sure, by no means representative of the Pentecostal movement.

Pentecostalism outside the United States. Though the phenomenal growth of U.S. Pentecostalism during the last 50 years is well documented, the expansion of Pentecostalism in other parts of the world has been equally remarkable, though less well known. In Mexico, the 146,000 Pentecostals constitute the largest segment of Evangelical Protestantism. The growth of Pentecostalism in Latin America, especially Argentina, Brazil, and Chile, has been quite impressive. The memberships of the Congregação Cristã and Assembléias de Deus do Brasil, the two largest Brazilian Pentecostal bodies, are in excess of 4,000,000. Chile probably is the most Pentecostal of all the South American countries: two-thirds of its Protestant inhabitants are Pentecostal.

Latin American and European Pentecostalism

In Sweden as well as Norway, the Filadelfia Pentecostal Churches are numerically second only to the state church (Lutheran). Most astonishing, perhaps, is the growth of Pentecostalism in the Soviet Union. Despite the fact that the movement has no official recognition and is frequently the object of scurrilous attacks both in the Soviet press as well as in government circles, the Pentecostal movement apparently has experienced significant growth in the two decades following World War II. According to a British Pentecostal leader, J. Nelson Parr, there were, in the early 1960s, approximately 250,000 Pentecostal believers in the U.S.S.R., especially in the Ukraine and in Siberia. Because many Pentecostals have been intransigent, refusing to affiliate with the All-Union Council of Evangelical Christians and Baptists, which is approved by Soviet authorities, and are hostile toward the Orthodox Church or toward religious groups that are politically docile, they have often been vilified by Soviet authorities for engaging in activities that are characterized as "hostile to humanity."

As the Pentecostal movement has grown, it has experienced change. Though continuing to minister to those on the social periphery, Pentecostal churches are no longer meeting only the needs of the dispossessed or disinherited. Allegiance to the storefront, second-floor mission, or frame tabernacle on the wrong side of the tracks has given way to interest in more permanent structures. Pentecostals in Grand Falls, Newfoundland, have constructed an ultramodern building that cost $1,000,000; the Apostolic Church in England recently erected the largest church building in Wales.

Recent developments. Also, Pentecostals were once militantly anti-intellectual. Many of them held the view that those who were filled with the Spirit needed no one to instruct them. Since 1950, however, nearly a score of Bible colleges and colleges of arts and science have been established by various Pentecostal denominations. Pentecostal services have changed as well. The lack of any order of worship and the emotional and charismatic outbursts that were allowed free course have largely disappeared. The spontaneous outbursts of glossolalia that frequently punctuated the services of worship have become controlled by the minister.

In the early days of the movement, Pentecostals were sometimes charged with being aloof, critical, or suspicious of non-Pentecostals. This possibly was a reaction to the vilification that was directed at them by many of the established denominations. Whatever the reason, the attitude persisted until the 1950s, after which the suspicion and criticism of others subsided and was replaced by a probing self-criticism.

A growing number of Pentecostals have been aroused to concern themselves with social problems. Unlike many of their predecessors, they have not been as prone to ignore the various social, economic, and political problems that concern their contemporaries. Some, for the first time, have recognized an obligation to attempt to relate the principles of the gospel to these problems. As a result of this concern, some have become involved in political activities; others have become counsellors to prison inmates; still others have been working with drug addicts, neighbourhood gangs, prostitutes, and other persons overpowered by urban problems.

Transformation of Pentecostalism

A large segment of the Pentecostal movement has undergone a transformation. While this has been going on and the larger groups have been gaining adherents and becoming increasingly successful, some others, remembering the simplicity of the early days, have become uncomfortable and have started to withdraw and form new factions. Such new factions—as the early Pentecostals before them—advocate little or no organization, stress "leadings of the Spirit" in place of a fixed order of worship and a structured organization, and prefer isolation to compromise. At the end of World War II, a veritable spate of these independent Pentecostal groups appeared: the New Order of the Latter Rain, the World Church, Wings of Healing, and the Full Gospel Fellowship of Ministers and Churches, International.

Teachings, organization, and practice. *Teachings.* In matters of doctrines, Pentecostals can be described as Evangelicals whose theology is akin to Fundamentalism. Almost universally, they believe in the infallibility of the Bible, the Trinity, the virgin birth and complete deity of Christ, the atoning sacrifice of his death for the sins of the world, the Resurrection, and the return of Christ in bodily form to earth. In other matters of faith and practice, they resemble revivalistic and Holiness movements that flourished during the late 19th century. They believe that one must pray unceasingly, expressing godly sorrow for his sins until he feels an inward assurance of salvation. They also declare that a person can attain holiness if he abandons his own efforts to achieve perfection and permits the Holy Spirit to live within him the Christ life.

But the teaching unique to Pentecostalism, setting it apart from other types of Protestant Christianity, is the belief that a baptism in the Holy Spirit, accompanied by speaking in tongues, should be the normative religious experience of all Christians. Pentecostals base their belief in a Spirit baptism and glossolalia on instances recorded in Acts. To them, speaking in tongues is a supernatural utterance in a language never learned by the speaker, not understood by him, and, furthermore, rarely comprehended by the listener. Pentecostals value the purposiveness of the tongue-speaking experience. They believe that tongues often strengthen the individual Christian in his witness; that, when accompanied by interpretation, glossolalia edifies the church; and that this activity is a vehicle of the loftiest kind of spiritual praise. What is frequently overlooked, however, is that, for most aware Pentecostals, glossolalia is not intended to be the culminating event in a catalog of religious experiences but rather the commencement of a new life in the Spirit that is distinguished by certain charismatic gifts, such as the ability to speak wisely or prophetically, the power to heal or work miracles, and the ability to discern error or control demons.

Pentecostal normative experiences

Worship. Pentecostal worship presents a variegated pattern. In the United States, for example, some churches are still presided over by lay preachers who encourage demonstrative meetings, in which opportunities for handclapping, crying, shouting, and the exercise of the charismata (spiritual powers) abound. On the other hand, in churches served by trained ministers—trained either in college, seminary, or Bible institute—worship is more orderly; disrupting manifestations of emotionalism have largely disappeared; ardour has been counterbalanced by a growing respect for order; and a sort of modified liturgical tradition has emerged. Even in the more sedate

churches, however, Pentecostal worship is still joyful, emotional, and sometimes unpredictable: strong men weep, and quiet women shout "Hallelujahs."

The ordinary Pentecostal worship service may include many, if not all, of the following elements: (1) spirited singing of gospel hymns by the entire congregation, often accompanied by a choir and orchestra; (2) extemporaneous prayers, generally by one person but frequently by the entire gathering; (3) personal testimonies by church members who will witness publicly to some spiritual triumph or answered prayer, who will confess to some act of sin, or who will request the prayers of fellow believers for a particular physical, spiritual, or material need; (4) biblical, evangelistic preaching that concentrates on one or more of the following topics—salvation, Spirit baptism, divine healing, or the Second Coming; and (5) the altar service, during which time those who wish to be "saved, sanctified, filled with the Holy Spirit, or healed" are invited to go forward to the altar, where they are prayed for by the minister and many of the members.

Pentecostal ordinances

Like most Protestants, Pentecostals practice two ordinances: Baptism, generally of adults and by immersion; and Communion (the Lord's Supper). A few groups, such as the Churches of God, advocate another—foot washing, explaining that, because Jesus instructed his disciples to wash another's feet and because the early church practiced this rite, it, too, ought to be continued as a Pentecostal ordinance.

Organization. During the early years of the Pentecostal revival, many of its adherents were anti-denominational, aloof toward established denominations, and fragmented among themselves over a variety of issues.

In the 1940s, however, changes began to occur. Pentecostals decided to abandon their isolationism. They joined the National Association of Evangelicals (1943); they united to form national as well as international fellowships: *i.e.*, the Pentecostal Fellowship of North America (1948) and the Pentecostal World Conference (1947). Pentecostal groups from Latin America have applied for and have been admitted to membership in the World Council of Churches.

Recent estimate of Pentecostalism by other Christian Churches. Through organizations such as the Full Gospel Businessmen's Fellowship International, the efforts of Pentecostal ecumenists and the impact of the salvation-healing revivals, conducted by outstanding television personalities such as Oral Roberts, the charismatic movement has affected not only Protestants of nearly every denomination but many Roman Catholics as well. This has forced many critics who had earlier rejected Pentecostalism as a suspect religious movement to revise their appraisals of it.

Pentecostals are frequently lauded for their dynamic missionary concern, for re-emphasizing the central place that ought to be accorded the Person and work of the Holy Spirit in the Christian Church, for helping nominal Christians and casual churchgoers become vital believers and zealous witnesses after experiencing their personal Pentecost, and for ministering to the heterogeneous needs of the dispossessed—*i.e.*, the economically disinherited, the educationally deprived, the geographically uprooted, the physically afflicted, and the psychologically insecure.

Other estimates of Pentecostalism have been less positive. The movement has been criticized for channelling too much of its resources and services into evangelistic missionary enterprises and for remaining indifferent to the manifold social problems that afflict contemporary society.

BIBLIOGRAPHY

General surveys of Pentecostalism: WALTER J. HOLLENWEGER, *Handbuch der Pfingstbewegung,* 10 vol. (1965–67), a comprehensive handbook of Pentecostal history, teachings, organizations, and practices, containing ample quotations from original sources and an exhaustive bibliography; JOHN THOMAS NICHOL, *Pentecostalism* (1966), the only single-volume history in English that describes the worldwide scope of Pentecostalism and attempts to classify the scores of varieties of Pentecostalism in the United States; GORDON F. ATTER, *The Third Force* (1962), a brief handbook prepared by a Pentecostal minister and Bible College instructor in Canada for his students; N. BLOCH-HOELL, *Pinsebevegelsen* (1956; Eng. trans., *The Pentecostal Movement,* 1964), a work important for an understanding of Pentecostalism in Norway and other Scandinavian countries, including an excellent section on the early developments.

Pentecostalism in Europe and Great Britain: WOLFGANG SCHMIDT, *Die Pfingstbewegung in Finnland* (1935), the only source generally available about Finnish Pentecostals; PAUL A.G. FLEISCH, *Die Pfingstbewegung in Deutschland* (1957), the most exact history of German Pentecostalism; DONALD GEE, *The Pentecostal Movement* (1941), an important historical sketch of Pentecostalism in the British Isles; LEONHARD STEINER, *Mit folgenden Zeichen* (1954), a survey of Pentecostalism in Switzerland and elsewhere by a leader of Swiss Pentecostalism.

Pentecostalism in the United States: CARL BRUMBACK, *Suddenly from Heaven* (1961), an authoritative history of the Assemblies of God; CHARLES W. CONN, *Like a Mighty Army: Moves the Church of God, 1886–1955* (1955), a meticulously documented history of the Church of God; STANLEY FRODSHAM, *With Signs Following,* rev. ed. (1941), a valuable source document on the early days of the charismatic revival; KLAUDE KENDRICK, *The Promise Fulfilled* (1961), a detailed history of the eight most prominent American Pentecostal bodies; KEVIN and DOROTHY RANAGAN RANAGHAN, *Catholic Pentecostals* (1969), a recent report on the spread of Pentecostal movements on Catholic campuses; JOHN L. SHERRILL, *They Speak with Other Tongues* (1964), one of the best treatments of the recent revival of charismatic phenomena in non-Pentecostal circles.

(J.T.N.)

Pepin the Short

Pepin III, called Pepin the Short, was the founder of the Frankish Carolingian dynasty of medieval Europe and the father of Charlemagne. He was the first king of the Franks to be anointed by the pope upon accession to the throne.

Rise of the mayors of the palace

For years the Merovingian kings had been unable to prevent power from slipping from their hands into those of the counts and other magnates. The kings were gradually eclipsed by the mayors of the palace, whose status developed from that of officer of the household to regent or viceroy. Among the mayors, a rich family descended from Pepin of Landen (Pepin I) held a position of especial importance. When Charles Martel, the scion of that family, died in 741, he left two sons: the elder, Carloman, mayor of Austrasia, Alamannia, and Thuringia, and Pepin III, mayor of Neustria, Burgundy, and Provence. No king had ruled over all the Franks since 737, but to maintain the fiction of Merovingian sovereignty, the two mayors gave the crown to Childeric III in 743.

Charles had had a third son, however—Grifo, who had been born to him by a Bavarian woman of high rank, probably his mistress. In 741, when his two brothers were declared mayors of the Franks, Grifo rebelled. He led a number of revolts in subsequent years and was several times imprisoned. In 753 he was killed amid the Alpine passes on his way to join the Lombards, at this time enemies of the Franks as well as of the papacy.

Numerous other rebellions broke out. In 742 men of the Aquitaine and Alamannia were in revolt; in 743 Odilo, duke of Bavaria, led his men into battle; in 744 the Saxons rebelled; in 745 Aquitaine, and in 746 Alamannia, both for the second time.

In 747, when Carloman decided to enter monastic life at Rome, a step he had been considering for years, Pepin became sole ruler of the Franks. But Pepin was ambitious to govern his people as king, not merely as mayor. Like his father, he had courage and resolution; unlike his father, he had a strong desire to unite the papacy with the Frankish realm. In 750 he sent two envoys to Pope Zacharias with a letter asking: "Is it wise to have kings who hold no power of control?" The Pope answered: "It is better to have a king able to govern. By apostolic authority I bid that you be crowned King of the Franks." Childeric III was deposed and sent to a monastery, and Pepin was anointed as king at Soissons in November 751 by Archbishop Boniface and other prelates.

Pepin and Pope Stephen II

The Pope was in need of aid. Aistulf, king of the Lombards, had seized Ravenna with its lands, known as the exarchate. Before long, Lombard troops marched south, surrounded Rome, and prepared to lay siege to its walls.

This was the state of things when in 752 Zacharias died and Stephen II became pope. In November 753 Pope Stephen made his way over the stormy mountain passes to Frankish territory. He remained in France until the summer of 754, staying at the abbey of Saint-Denis, Paris. There he himself anointed Pepin and his sons, Charles and Carloman, as king and heirs of the crown.

The Pope returned to Italy accompanied by Pepin and his army. A fierce battle was fought in the Alps against Aistulf and the Lombards. The Lombard king fled back to his capital, Pavia; Pepin and his men plundered the land around Pavia until Aistulf promised to restore to papal possession Ravenna and all the Roman properties claimed by the Pope.

Aistulf broke his word. Again and again Pope Stephen wrote to Pepin of his difficulties. In 756 the Frankish king once more entered Italy. Aistulf was once more constrained to make promises, but the same year he died—of a fall from his horse—and in April 757 a new king, Desiderius, became ruler of the Lombards. That year Stephen II also died, and Paul I was elected pope. He, too, constantly wrote to Pepin asking for help.

But the King of the Franks had other concerns. He had to put down revolts in Saxony in 748 and 753 and a rising in Bavaria in 749. He was continually marching against rebellious Aquitaine. In 768 Pepin died at Saint-Denis, on his way back from one of his Aquitainian expeditions.

Pepin is remembered not only as the first of the Carolingians but also as a strong supporter of the Roman Church. The papal claims to territory in Italy originated with Pepin's campaigns against Aistulf and the latter's pledge to return the Roman territories. His letters also show him calling for archbishoprics in Frankish territory, promoting synods of clergy and layfolk, and as deeply interested in theology.

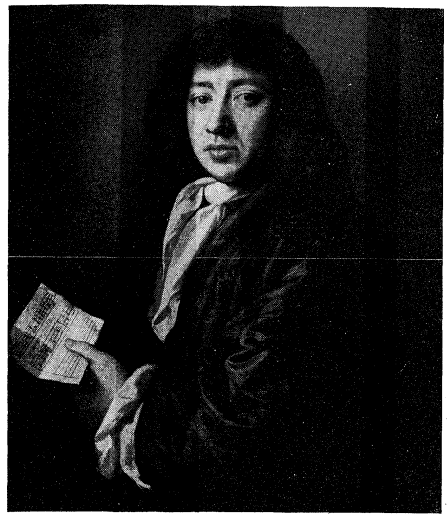
BIBLIOGRAPHY. EPHRAIM EMERTON, *Letters of Saint Boniface* (1940), an excellent translation, with a helpful preface; THOMAS HODGKIN, *Italy and Her Invaders*, vol. 7 (1899), a full account of Frankish invasions; J.M. WALLACE-HADRILL, *The Barbarian West, 400–1000*, 3rd rev. ed. (1967), history in Africa, Italy, and France; E. DUMMLER (ed.), *Monumenta Germaniae Historica, Epistolae Merovingici et Karolini Aevi*, vol. 1 (1892); F. KURZE (ed.), *Annales Regni Francorum* (1895); W. WATTENBACH (ed.), *Monumenta Germaniae Historica, Scriptores Rerum Merovingicarum*, vol. 2 (1888).

(E.S.D.)

Pepys, Samuel

English man of letters and naval administrator whose celebrated diary gives a valuable, entertaining, and intimate picture of the social life of his day, Samuel Pepys was born on February 23, 1633, in Salisbury Court near St. Bride's Church, London. He was the son of a working tailor who had come to London from Huntingdonshire, in which county, and in Cambridgeshire, his family had lived for centuries as monastic reeves, rent collectors, farmers, and, more recently, small gentry. His mother, Margaret Kite, was the sister of a Whitechapel butcher. But, though of humble parentage, Pepys rose to be one of the most important men of his day, becoming England's earliest secretary of the admiralty and serving in his time as member of Parliament, president of the Royal Society (in which office he placed his imprimatur on the title page of England's greatest scientific work, Sir Isaac Newton's *Philosophiae Naturalis Principia Mathematica*), master of Trinity House and of the Clothworkers' Company, and a baron of the Cinque Ports. He was the trusted confidant both of Charles II, from whom he took down in shorthand the account of his escape after the Battle of Worcester, and of James II, whose will he witnessed before the royal flight in 1688. The friends of his old age included Sir Christopher Wren, Sir Isaac Newton, John Evelyn, Sir Godfrey Kneller, John Dryden, and almost every great scholar of the age.

Early career. Samuel Pepys (pronounced peeps) was sent, after early schooling at Huntingdon, to St. Paul's School, London. In 1650 he was entered at Trinity Hall, Cambridge, but instead went as a sizar to Magdalene College, obtaining a scholarship on the foundation. In March 1653 he took his B.A. degree and in 1660 that of



Pepys, oil painting by John Hayls, 1666. In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

M.A. Little is known of his university career save that he was once admonished for being "scandalously overserved with drink." In later years he became a great benefactor of his college, to which he left his famous library of books and manuscripts. He was also once offered—but refused—the provostship of King's College, Cambridge. Education and marriage

In December 1655 he married a penniless beauty of 15, Elizabeth Marchant de Saint-Michel, daughter of a French Huguenot refugee. At this time he was employed as factotum in the Whitehall lodgings of his cousin Adm. Edward Montagu, later 1st earl of Sandwich, who was high in the lord protector Cromwell's favour. In his diary Pepys recalls this humble beginning, when his young wife "used to make coal fires and wash my foul clothes with her own hand for me, poor wretch! in our little room at Lord Sandwich's; for which I ought forever to love and admire her, and do." While there, on March 26, 1658, he underwent a serious abdominal operation, thereafter always celebrating the anniversary of his escape by a dinner—"This being my solemn feast for my cutting of the stone."

In 1659 Pepys accompanied Montagu on a voyage to the Sound. About the same time he was appointed to a clerkship of £50 per annum in the office of George Downing, one of the tellers of the Exchequer, after whom Downing Street was later named. It was while working in Downing's office and living in a small house in Axe Yard that on January 1, 1660, he began his diary. A few months later he sailed, as his cousin's secretary, with the fleet that brought back Charles II from exile. Appointed, through Montagu's interest at court, clerk of the acts of the navy at a salary of £350 per annum and given an official residence in the navy office in Seething Lane, he became in the next few years a justice of the peace, a commissioner for and, later, treasurer of, Tangier, and surveyor of naval victualling. When he entered upon his functions, he was ignorant of almost everything that belonged to them. His chief use of his position was to enjoy his newfound importance and the convivial companionship of his colleagues Admirals Sir William Batten and Sir William Penn. But early in 1662 there came a change. The colleagues whose bacchanalian habits and social position had made them so attractive began to prove irksome, and their insistence on their superior experience and status galled Pepys's pride. In his isolation, he sought for ways by which he could show himself their equal. He had not far to look, for his fellow officers were anything but attentive to business. "So to the office," Pepys wrote, "where I do begin to be exact in my duty there and exacting my privileges and shall continue to do so." He had found his vocation.

Naval administration. It was not in Pepys's nature to do things by halves. Having resolved to do his duty, he set out to equip himself for its performance. In the summer of 1662, he occupied his leisure moments by learning the

multiplication table, listening to lectures on shipbuilding, and studying the prices of naval stores: "into Thames Street, beyond the Bridge, and there enquired among the shops the price of tar and oil, and do find great content in it, and hope to save the King money by this practise." At the same time, he began his habit of making careful entries of all contracts and memoranda in large vellum books—beautifully ruled by Mrs. Pepys and her maids—and of keeping copies of his official letters.

The qualities of industry and devotion to duty that Pepys brought to the service of the Royal Navy became realized during the Second Dutch War of 1665–67—years in which he remained at his post throughout the Plague and saved the navy office in the Great Fire of London. Before trouble with his eyesight caused him to discontinue his diary in 1669—an event followed by the death of his wife—these qualities had won him the trust of the King and his brother James, the duke of York, the lord high admiral. In 1673, in the middle of the Third Dutch War, when York's unpopular conversion to Catholicism forced him to resign his office, Pepys was appointed secretary to the new commission of admiralty and, as such, administrative head of the navy. In order to represent it in Parliament—before whom he had conducted a masterly defense of his office some years before—he became member first for Castle Rising and, later, for Harwich. For the next six years he was engaged in stamping out the corruption that had paralyzed the activities of the navy. His greatest achievement was carrying through Parliament a program that, by laying down 30 new ships of the line, restored the balance of sea power, upset by the gigantic building programs of France and the Netherlands. In his work both at the Admiralty and in Parliament, Pepys's unbending passion for efficiency and honesty (combined with a certain childlike insistence on his own virtue and capacity for being always in the right) made for him powerful and bitter enemies. One of these was Lord Shaftesbury, who in 1678 endeavoured to strike at the succession and at the Catholic successor, the Duke of York, by implicating Pepys in the mysterious murder of the London magistrate Sir Edmund Berry Godfrey, the crime on which the full credulity of the populace in the Popish Plot depended. When Pepys produced an unanswerable alibi, his enemies endeavoured to fasten Godfrey's murder on him indirectly by accusing his confidential clerk, Samuel Atkins. Despite the third-degree methods employed against him, Pepys also proved an alibi for Atkins, who would otherwise almost certainly have perished. Six months later, his enemies brought into England a picturesque scoundrel and blackmailer called John Scott, who had begun his life of crime in what today is Long Island, New York, and whom Pepys had endeavoured to have arrested at the time of Godfrey's death on account of his mysterious activities disguised as a Jesuit. Pepys was flung into the Tower on an absurd charge of treason brought against him by Scott and supported by the Exclusionists in Parliament, as also on a minor and equally unjust charge of popery, brought against him by a dismissed butler whom he had caught in bed with his favourite maid. Had not Charles II almost immediately dissolved Parliament and prevented a new one from meeting for a further year and a half, Pepys would have paid the penalty for his loyalty, efficiency, and incorruptibility with his life. He employed his respite with such energy that by the time Parliament met again he had completely blasted the reputation of his accuser by collecting circumstantial details of his infamies from almost every country.

Member of Parliament

In 1683, when the King felt strong enough to ignore his opponents, Pepys was taken back into the public service. He had accompanied the Duke of York in the previous year on a voyage to Scotland, and he now sailed as adviser to the Earl of Dartmouth to evacuate the English garrison of Tangier—a voyage that he described in a further journal.

Return to the Admiralty

On his return, in the spring of 1684, he was recalled by Charles II to his old post. Entitled secretary of the affairs of the admiralty of England and remunerated by a salary of £500 per annum, he combined the modern offices of first lord and secretary of the admiralty, both administering the service and answering for it in Parliament. For the next four and a half years, including the whole of James II's reign, Pepys was one of the greatest men in England, controlling the largest spending department of state. With his habitual courage and industry, he set himself to rebuild the naval edifice that the inefficiency and corruption of his enemies had shattered, securing in 1686 the appointment of a special commission "for the Recovery of the Navy." When, at the beginning of 1689, after James II had been driven from the country, Pepys retired, he had created a navy strong enough to maintain a long ascendancy in the world's seas. Hitherto there had been brief spells when the emergence of a naval genius like Sir Francis Drake and Robert Blake had given England a temporary advantage over its maritime neighbours, but for long periods the English "sovereignty of the seas" had been an idle boast. When Pepys became associated with the navy in 1660, the line of battle had consisted of 30 battleships of a total burden of approximately 25,000 tons and carrying 1,730 guns. When he laid down his office, he left a battle line of 59 ships of a total burden of 66,000 tons and carrying 4,492 guns. Not only had he doubled the navy's fighting strength; he had given it what it had never possessed before and what it never again lost—a great administrative tradition of order, discipline, and service. He had made the scabbard of the sword that the great 18th-century admirals used.

Last years

"To your praises," declared the orator of Oxford University, "the whole ocean bears witness; truly, sir, you have encompassed Britain with wooden walls." Pepys's last 14 years, despite attempts by his political adversaries to molest him, were spent in honorable retirement in his riverside house in York Buildings, amassing and arranging the library that he ultimately left to Magdalene College, Cambridge, corresponding with scholars and artists, and collecting material for a history of the navy that he never lived to complete, though he published a prelude to it in 1690, describing his recent work at the Admiralty, entitled *Memoires relating to the State of the Royal Navy of England for ten years determined December 1688*. He died on May 26, 1703, at the Clapham home of his former servant and lifelong friend William Hewer. His fellow diarist John Evelyn wrote of him: "He was universally belov'd, hospitable, generous, learned in many things, skilled in music, a very greate cherisher of learned men of whom he had the conversation."

The diary. The diary by which Pepys is chiefly known was kept between his 27th and 36th years. Written in Thomas Shelton's system of shorthand, or tachygraphy, with the names in longhand, it extends to 1,250,000 words, filling six quarto volumes in the Pepys Library. It is far more than an ordinary record of its writer's thoughts and actions; it is a supreme work of art, revealing on every page the capacity for selecting the small, as well as the large, essential that conveys the sense of life; and it is probably, after the Bible and James Boswell's *Life of Samuel Johnson*, the best bedside book in the English language. One can open it on any page and lose oneself in the life of Charles II's London, and of this vigorous, curious, hardworking, pleasure-loving man. Pepys wanted to find out about everything because he found everything interesting. He never seemed to have a dull moment; he could not, indeed, understand dullness. One of the more comical entries in his diary refers to a country cousin, named Stankes, who came to stay with him in London. Pepys had been looking forward to showing him the sights of the town—

> But Lord! what a stir Stankes makes, with his being crowded in the streets, and wearied in walking in London, and would not be wooed by my wife and Ashwell to go to a play, nor to White Hall, or to see the lions, though he was carried in a coach. I never could have thought there had been upon earth a man so little curious in the world as he is.

Pepys's style

Pepys possessed the journalist's gift of summing up a scene or person in a few brilliant, arresting words. He makes us see what he sees in a flash; his Aunt James, "a poor, religious, well-meaning, good soul, talking of nothing but God Almighty, and that with so much innocence that mightily pleased me"; and his sister Pall, "a pretty,

good-bodied woman and not over thick, as I thought she would have been, but full of freckles and not handsome in the face." He could describe with wonderful vividness a great scene: as, for example, the day Gen. George Monck's soldiers unexpectedly marched into a sullen City and proclaimed there should be a free Parliament—"And Bow bells and all the bells in all the churches as we went home were a-ringing; it was past imagination, both the greatness and suddenness of it." He described, too, the Restoration and coronation; the horrors of the Plague; and the Fire of London, writing down his account—so strong was the artist in him—even as his home and its treasures were being threatened with destruction:

> We saw the fire as only one entire arch of fire from this to the other side of the bridge, and in a bow up the hill for an arch of above a mile long: it made me weep to see it. The churches, houses, and all on fire and flaming at once; and a horrid noise the flames made, and the cracking of houses at their ruine.

Above all, Pepys possessed the artist's gift of being able to select the vital moment. He makes his readers share the very life of his time: "I staid up till the bell-man came by with his bell just under my window as I was writing of this very line, and cried, 'Past one of the clock, and a cold, and frosty, windy morning.' " He tells of the guttering candle, "which makes me write thus slobberingly"; of his new watch—"But Lord! to see how much of my old folly and childishness hangs on me still that I cannot forebear carrying my watch in my hand in the coach all the afternoon and seeing what o'clock it is one hundred times"; of being awakened in the night—

> About 3 o'clock this morning I waked with the noise of the rain, having never in my life heard a more violent shower; and then the cat was locked in the chamber and kept a great mewing and leapt upon the bed, which made me I could not sleep a great while.

Pepys excluded nothing from his journal that seemed to him essential, however much it told against himself. He not only recorded his major infidelities and weaknesses; he put down all those little meannesses of thought and conduct of which all men are guilty but few admit, even to themselves. He is frank about his vanity—as, for example, in his account of the day he went to church for the first time in his new perriwig: "I found that my coming in a perriwig did not prove so strange to the world as I was afeared it would, for I thought that all the church would presently have cast their eyes upon me, but I found no such thing"; about his meannesses over money, his jealousies, and his injustices—"Home and found all well, only myself somewhat vexed at my wife's neglect in leaving her scarfe, waistcoat and night dressings in the coach today; though I confess she did give them to me to look after." For he possessed in a unique degree the quality of complete honesty. He is both Everyman and the recording angel; his diary paints not only his own infirmities but the frailty of all mankind.

BIBLIOGRAPHY

Diary: After the successful publication of John Evelyn's diary in 1818, Pepys's diary was transcribed—with great accuracy—by JOHN SMITH, later rector of Baldock, Hertfordshire, at the instance of the master and fellows of Magdalene College, Cambridge. About a quarter of it, ed. by the 3RD LORD BRAYBROOKE, was published as *Memoirs of Samuel Pepys, Esq., F.R.S.*, 2 vol. (1825). Many fuller editions followed, until the whole diary, with the exception of about 50 short passages omitted on grounds of decency, ed. by H.B. WHEATLEY, appeared in 10 vol. (1893–99, reprinted 1968). The St. Olave edition, 18 vol., was published in 1903. Later editions and selections include *Everybody's Pepys*, abridged and ed. by O.F. MORSHEAD (1926 and 1960); and *The Diary*, selections ed. by JOHN WARRINGTON, Everyman's Library, 3 vol. (1953). The definitive ed. (the first in which the entire diary is printed) in a new and complete transcription is currently being issued under the editorship of ROBERT LATHAM and WILLIAM MATTHEWS, *The Diary of Samuel Pepys*, to comprise 11 vol.: 9 vol. of text, 1 vol. of commentary, and an index volume (1970–).

Correspondence: The Life, Journals and Correspondence of Samuel Pepys, ed. by JOHN SMITH, 2 vol. (1841); *Private Correspondence and Miscellaneous Papers of Samuel Pepys, 1679–1703*, 2 vol. (1926), *Further Correspondence of Samuel Pepys, 1662–1679* (1929), and *Samuel Pepys's Naval Minutes* (1926), all ed. by J.R. TANNER; *The Tangier Papers of Samuel Pepys*, ed. by EDWIN CHAPPELL (1935); *The Letters of Samuel Pepys and His Family Circle*, ed. by H.T. HEATH (1955).

Pepys's Library: A Descriptive Catalogue of the Naval Manuscripts in the Pepysian Library at Magdalene College, Cambridge, ed. by J.R. TANNER, 4 vol. (1903–23); *Bibliotheca Pepysiana: A Descriptive Catalogue of the Library of Samuel Pepys*, 4 vol. (1914–40); S. GASELEE, *The Spanish Books in the Library of Samuel Pepys* (1921).

Biography and criticism: H.B. WHEATLEY, *Samuel Pepys and the World He Lived In* (1880), *Pepysiana* (1899); SIR J.F. BRIDGE, *Samuel Pepys: Lover of Musique* (1903); PERCY LUBBOCK, *Samuel Pepys* (1909); E.H. MOORHOUSE, *Samuel Pepys* (1909); J.R. TANNER, *Samuel Pepys and the Royal Navy* (1920, reprinted 1971), *Mr. Pepys: An Introduction to the Diary Together with a Sketch of His Later Life* (1925, reprinted 1971); EDWIN CHAPPELL, *Samuel Pepys As a Naval Administrator* (1933); SIR ARTHUR BRYANT, *Samuel Pepys: The Man in the Making* (1933; 2nd ed., 1947), *The Years of Peril* (1935; 2nd ed., 1948), and *The Saviour of the Navy* (1938; 2nd ed., 1949)—new ed., *Samuel Pepys*, 3 vol. (1967); RUDOLF KIRK, *Mr. Pepys upon the State of Christ-Hospital* (1935); CECIL EMDEN, *Pepys Himself* (1963); RICHARD BARBER, *Samuel Pepys, Esquire* (1970).

(Ar.Br.)

Perception

Perception, or perceiving, refers to the process whereby sensory stimulation is translated into organized experience. That experience, or percept, is the joint product of the stimulation and of the process itself. Relations found between various types of stimulation (*e.g.*, light waves and sound waves) and their associated percepts suggest inferences that can be made about the properties of the perceptual process; theories of perceiving then can be developed on the basis of these inferences. Because the perceptual process is not itself public or directly observable (except to the perceiver himself, whose percepts are given directly in experience), the validity of perceptual theories can be checked only indirectly. That is, predictions derived from theory are compared with appropriate empirical data, quite often through experimental research.

Historical background

Historically, systematic thought about perceiving was the province of philosophy. Indeed, perceiving remains of interest to philosophers, and many issues about the process that were originally raised by philosophers are still of current concern. As a scientific enterprise, however, the investigation of perception has especially developed as part of the larger discipline of psychology.

Philosophical interest in perception stems largely from questions about the sources and validity of what is called human knowledge (see EPISTEMOLOGY). Epistemologists ask whether a real, physical world exists independently of man's experience, and if so, how man can come to learn its properties and how the truth or accuracy of that experience can be determined. They also ask whether there are innate ideas or whether all experience originates through contact with the physical world, mediated by the sense organs. For the most part, psychology bypasses such questions in favour of problems that can be handled by its special methods. The remnants of such philosophical questions, however, do remain; researchers are still concerned, for example, with the relative contributions of innate and learned factors to the perceptual process.

Such fundamental philosophical assertions as the existence of a physical world, however, are taken for granted among most scientific students of perceiving. Typically, researchers in perception simply accept the apparent physical world particularly as it is described in those branches of physics concerned with electromagnetic energy, optics, and mechanics. The problems they consider are those relating to the process whereby percepts are formed from the interaction of what is called physical energy (for example, light) with the perceiving organism. Of further interest is the degree of correspondence between percepts and the physical objects to which they ordinarily relate. How accurately, for example, does the visually perceived size of an object match its physical size as measured (*e.g.*, with a yardstick)?

Questions of the latter sort imply that perceptual experi-

ences typically have external referents and that they are meaningfully organized, most often as objects. Meaningful objects, such as trees, faces, books, tables, and dogs, are normally seen rather than separately perceived as the dots, lines, colours, and other elements of which they are composed. In the language of Gestalt psychologists, immediate human experience is of organized wholes (*Gestalten*), not of collections of elements.

A major goal of Gestalt theory in the 20th century was to specify the brain processes that might account for the organization of perception. Gestalt theorists, chief among them the German-U.S. psychologist and philosopher, the founder of Gestalt theory, Max Wertheimer and the German-U.S. psychologists Kurt Koffka and Wolfgang Köhler, rejected the earlier assumption that perceptual organization was the product of learned relationships (associations), the constituent elements of which were called simple sensations. Although Gestaltists agreed that simple sensations logically could be understood to comprise organized percepts, they argued that percepts themselves were basic to experience. One does not perceive so many discrete dots (as simple sensations), for example; the percept is that of a dotted line.

The role of brain function

Without denying that learning can play some role in perception, many theorists took the position that perceptual organization reflects innate properties of the brain itself. Indeed, perception and brain functions were held by Gestaltists to be formally identical (or isomorphic), so much so that to study perception is to study the brain. Much contemporary research in perception is directed toward inferring specific features of brain function from such behaviour as the reports (introspections) people give of their sensory experiences. More and more such inferences are gratifyingly being matched with physiological observations of the brain itself.

Many investigators relied heavily on introspective reports, treating them as though they were objective descriptions of public events. Serious doubts were raised in the 1920s about this use of introspection by the U.S. psychologist John B. Watson and others, who argued that it yielded only subjective accounts and that percepts are inevitably private experiences and lack the objectivity commonly required of scientific disciplines. In response to objections about subjectivism, there arose an approach known as behaviourism that restricts its data to objective descriptions or measurements of the overt behaviour of organisms other than the experimenter himself. Verbal reports are not excluded from consideration as long as they are treated strictly as public (objective) behaviour and are not interpreted as literal, reliable descriptions of the speaker's private (subjective, introspective) experience. The behaviouristic approach does not rule out the scientific investigation of perception; instead, it modestly relegates perceptual events to the status of inferences. Percepts of others manifestly cannot be observed, though their properties can be inferred from observable behaviour (verbal and nonverbal).

One legacy of behaviourism in contemporary research on perception is a heavy reliance on very simple responses (often nonverbal), such as the pressing of a button or a lever. One advantage of this Spartan approach is that it can be applied to organisms other than man and to human infants (who also cannot give verbal reports). This restriction does not, however, cut off the researcher from the rich supply of hypotheses about perception that derive from his own introspections. Behaviourism does not proscribe sources of hypotheses; it simply specifies that only objective data are to be used in testing those hypotheses.

Behaviouristic methods for studying perception are apt to call minimally on the complex, subjective, so-called higher mental processes that seem characteristic of adult human beings; they thus tend to dehumanize perceptual theory and research. Thus, when attention is limited to objective stimuli and responses, parallels can readily be drawn between perceiving (by living organisms) and information processing (by such devices as electronic computers). Indeed, it is from this information-processing approach that some of the more intriguing theoretical contributions (*e.g.*, abstract models of perception) are currently being made. It is expected that such practical applications as the development of artificial "eyes" for the blind may emerge from these man–machine analogies. Computer-based machines that can discriminate among visual patterns already have been constructed, such as those that "read" the code numbers on bank checks.

CLASSICAL PROBLEMS

Sensing and perceiving. Many philosophers and psychologists have commonly accepted as fundamental a distinction made on rational grounds between sensing and perceiving (or between sensations and percepts). To demonstrate empirically that sensing and perceiving are indeed different, however, is quite another matter. It is often said, for example, that sensations are simple and that percepts are complex. Yet, only if there is offered some agreed upon (a priori) basis for separating experiences into two categories—sensations and percepts—can experimental procedures demonstrate that the items in one category are "simpler" than those in the other. Clearly, the arbitrary basis for the initial categorization itself cannot be subjected to empirical test. (See also SENSORY RECEPTION, HUMAN.)

Problems of verification aside, the simplicity–complexity distinction derives from the assumption that percepts are constructed of simple elements that have been joined through association. Presumably, the trained introspectionist can dissociate the constituent elements of a percept from one another, and in so doing, experience them as simple, raw sensations. Efforts to approach the experience of simple sensations might also be made by presenting very simple, brief, isolated stimuli; *e.g.*, flashes of light.

Perceptual learning

Another commonly offered basis for distinction is the notion that perceiving is subject to the influence of learning while sensing is not. It might be said that the sensations generated by a particular stimulus will be essentially the same from one time to the next (barring fatigue or other temporary changes in sensitivity), while the resulting percepts may vary considerably, depending on what has been learned between one occasion and the next (see LEARNING, PERCEPTUAL).

Some psychologists have characterized percepts as typically related to external objects and sensations as more nearly subjective, personal, internally localized experiences. Thus, a spontaneous pain in the finger would be called a sensation; however, if the salient feature of experience is that of a painfully sharp, pointed object, such as a pin located "out there," it would be called a percept.

The above definitional criteria all relate to properties of experience; that is, they are psychological. An alternative way of distinguishing between sensing and perceiving that has become widely accepted is physiological–anatomical rather than psychological. In this case, sensations are identified with neural events occurring immediately beyond the sense organ, whereas percepts are identified with activity farther "upstream" in the nervous system, at the level of the brain. This assignment of anatomical locations to sensory and to perceptual processes seems consistent with psychological criteria. That is, the complexity and variability of percepts (both a product of learning) are attributed to the potential for physiological modification inherent in the vastly complex neural circuitry of the brain.

Temporal (time) relations. Clearly, many subjective processes (such as problem solving) take time to run their courses. This is true even for such relatively simple activities as perceptual discriminations in the size of different objects. It is not readily apparent, however, whether percepts themselves—which, for example, might enter as elements in problem solving—take time to form. To the naïve observer, percepts probably do seem essentially instantaneous: the moment a square is shown, a square seems to be seen. Yet, experimental evidence suggests that percepts, even of simple geometric forms, follow a measurable, developmental time course. In some instances the temporal development of percepts is relatively long (on

the order of seconds), and in some it is quite brief (on the order of thousandths of a second).

Pictures that are incomplete or ambiguous provide good examples of relatively long-term temporal development of percepts. Look at Figure 1 and continue looking until

By courtesy of Bell Laboratories

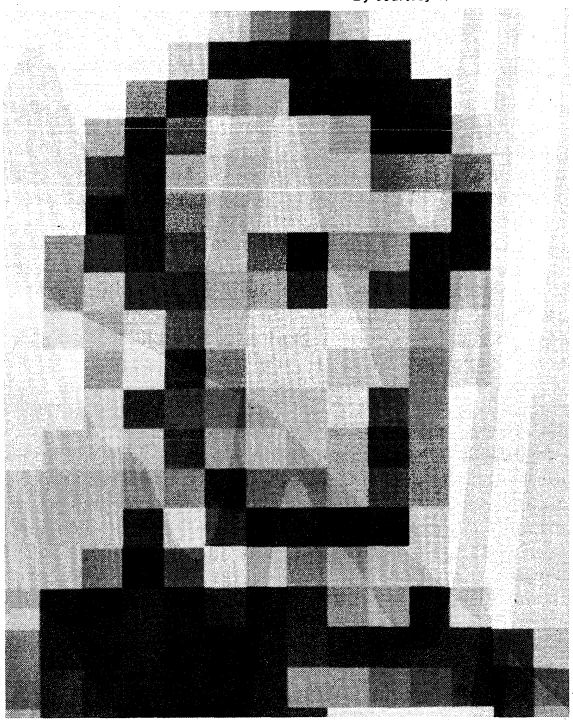


Figure 1: An ambiguous picture. Increasing viewing distance permits more precise perception (see text).

you see something more than a pattern of black, gray, and white patches. Abruptly, you probably will perceive a familiar face that, on subsequent viewing, will reappear to you without difficulty. How long it takes for such a percept to develop will vary considerably from one person to another, perhaps revealing fundamental differences among individuals in their speed of perceptual processing. It might be instructive to show Figure 1 to several people, and with the aid of a stopwatch, measure the time it takes each of them to achieve the desired percept, both initially and then on some later occasion. (Figure 1 commonly is seen by most people as the face of Abraham Lincoln.)

A somewhat different way in which percepts may change with time is illustrated in the article ILLUSIONS AND HALLUCINATIONS. On initial viewing of this type of drawing, one will probably immediately see a meaningful picture. After continued gazing at the drawing, the initial percept may abruptly be replaced by another. Thereafter, the two percepts should alternate with the passage of time. Stimuli of this sort (which can yield more than one percept) raise such questions as, for example, what determines the initial percept; why do some people first see a vase whereas others see two profiles; why does the initial percept give way to the alternate; what determines the rate of fluctuation from one percept to the other; do differences from one person to another in the rate of fluctuation of ambiguous figures indicate fundamental differences in perceptual activity? Tentative answers to such questions continue to be proposed.

Instances of slowly developing percepts require relatively simple procedures to uncover. Those percepts with a very rapid time course may be studied with the aid of instruments known as tachistoscopes that permit the durations of visual stimuli to be precisely controlled. Sophisticated electronic tachistoscopes flash reliably for periods as brief as one millisecond (one-thousandth of a second). Such precision permits study of the short-term development (microgenesis) of such percepts as those of simple geometric figures. Thus, it has been found that perception of a small black disk is disrupted (masked) by the rapidly successive tachistoscopic presentation of a second stimulus: a black ring that fits snugly around the disk. Indeed, as far as experimental subjects can tell, the disk target simply does not appear, though when flashed without being followed by the ring, it is readily detectable. Other target and mask stimuli also have been successfully employed.

Backward masking, or metacontrast

It has been theorized that it takes time for any percept to develop; that one's experience of a figure such as a disk develops from the figure's centre outward; that one's percept of the disk becomes stable only when the outer contour is appreciated; and that the ring functions in backward masking (metacontrast) because in the course of its own emergence as a percept the inner contour of the ring "absorbs" the perceptually developing contour of the disk. (Unless the viewer becomes aware of its contour, the disk theoretically cannot be perceived.) This interpretation is consistent with evidence of an optimal time interval between disk and ring onsets (about 30 to 50 milliseconds) for the best masking effect. Thus, masking is most evident at the moment developing awareness of disk contour is held to coincide in space and time with the initial perceptual growth of the ring's inner contour. If the ring is presented too soon or too late, theoretically, the contour absorption on which masking presumably depends is ineffective.

While such theories are controversial, masking experiments in general do clearly indicate that in human beings there is a brief period (100 to 200 milliseconds at most) during which a percept is highly vulnerable to disruption. Whatever its exact mechanism, the phenomenon of masking manifestly demonstrates that percepts do not emerge instantaneously and full-blown at the moment of sensory stimulation. Thus, assuming that percepts are synthesized from simpler elements, relatively complex percepts would be expected to take longest to develop and, hence, to be most vulnerable to masking. Yet empirical studies show just the opposite, indicating that the more complex the visual target, the more difficult it is to mask.

Perceiving as synthesizing. The organization apparent in percepts has been attributed by some to learning, as being built up through arbitrary associations of elements that have repeatedly occurred together in the person's experience. Other theorists (particularly Gestaltists) stress the view that perceptual organization is physiologically inborn, being inherent in innate aspects of brain functioning rather than depending on a synthesizing process of learning to combine simpler elements into more complex, integrated wholes. One way of resolving such theoretical disputes would be to deprive people from birth of all visual sensory experience and, hence, of all opportunity for visual perceptual learning. Then at the time normal sensory function was restored, they would need to be tested to determine what perceptual functions, if any, were intact. Such a strategy was proposed in a letter to the British philosopher John Locke by a fellow philosopher William Molyneux in 1690. Molyneux's suggestion waited until the 20th century to be taken seriously, after surgical methods had been found to restore the sight of people born blind because of cataract (clouded lens within the eye).

The use of cataract surgery to study perception

After removal of their cataracts, such newly sighted people are found to be normally sensitive to changes in intensity of illumination and to colour. Though they are able initially to tell when a figure is present, they cannot at first discriminate one simple shape from another, nor can they readily remember the shape of a just-exposed object. This deficiency extends to such socially important visual stimuli as people's faces. Only after a long and painstaking period of experience—perhaps of several months duration—do such seemingly primitive visual performances as discriminating a square from a triangle come easily. Until then, the person must count corners, for example, to achieve accurate discrimination.

Findings derived from cataract surgery have provided a rich source of hypotheses for further research, including posited neurophysiological mechanisms (*e.g.*, assemblies of brain cells) that might serve as the medium for the

structural changes presumed to accompany perceptual learning. This situation led to experimental attempts to show how, through repeated stimulation, the perceptual system could progress from performance of only very primitive functions to the highly complex operations (such as form identification and discrimination) that are characteristic of the mature organism.

A host of experiments using laboratory animals as subjects (*e.g.*, pigeons, rats, cats, dogs, monkeys, and chimpanzees) have been conducted to determine, under rigorous experimental control, the extent to which learning early in life contributes to later perceptual functioning. By analogy with humans born with cataract, such animals were deprived of visual experience from as close to birth (or hatching) as possible; *e.g.*, chimpanzees were reared in darkness. In another type of experiment, animals were reared in environments that provided more than the normal amount and variety of stimulation or were exposed to specific stimuli they might not ordinarily encounter. These research strategies are said to provide impoverished or enriched environments. Experiments of both sorts have consistently provided verification of the general hypothesis that early perceptual experience plays an important role in later perceptual (as well as intellectual and emotional) development, even producing changes in brain weight and biochemistry. This research also offers a strong scientific rationale for efforts to enrich the environments of so-called disadvantaged or culturally deprived children.

Studies of human infants indicate that their early perceptual experiences are not the "blooming, buzzing confusion" postulated by the U.S. psychologist William James late in the 19th and early in the 20th century. Rather, even infants one or two days old are capable of refined visual discriminations. Recording of the infants' visual fixations as their eyes move indicates them to have reliable preference for one paired stimulus over another, giving evidence of visual discrimination. Research employing this technique shows that preferences among various visual patterns or shapes generally are related to the complexity and novelty of the stimuli, for infants and for older human subjects as well. Evidence of this sort seems out of keeping with the findings of cataract surgery, which suggest that figural discrimination is not innate for the visually naïve. It may be, however, that adults who were blind with cataracts at birth suffer more than mere lack of normal visual experience; they are not quite comparable to visually naïve, but otherwise intact, infants. It may be that visual experience is necessary, not to generate pattern perception but to maintain it; that is, the infant's built-in (innate) perceptual abilities may somehow deteriorate through disuse.

The visual cliff technique

Similar research has dealt with visual depth perception in laboratory animals and human babies. One technique (the visual cliff) depends on the evident reluctance of young animals to step off the edge of what seems to be a steep cliff. The so-called visual cliff apparatus in one of its versions consists of a narrow platform on which the subject is placed and two wide platforms on either side of it. Although both flanking platforms are equally and only slightly lower than the central platform, the subject sees visual patterns designed so that one looks much deeper than the other. Typically, the subject explores the central platform and then investigates the flanks, finally stepping down onto the shallow-appearing side. By this response, the subject indicates sensitivity to visual depth cues. To discover if prior visual experience is necessary for the typical avoidance of the flanking platform that looks deepest requires subjects able to locomote well from the start (*e.g.*, chicks) or those deprived of visual experience (*e.g.*, rats) until their locomotor ability has developed. Research with the visual cliff clearly demonstrates the presence of depth perception in visually naïve subjects.

In summary, there is evidence among cataract sufferers of the necessity of early visual experience in human visual pattern discrimination; laboratory animals reared in impoverished or enriched environments demonstrate the importance (if not necessity) of early visual stimulation for perceptual development; human infants and other visually naïve animals behave as if they are innately capable of pattern and depth perception (*i.e.*, without the need for learning). These data together suggest that some basic visual functions, including pattern perception, are built in but that visual experience serves to maintain and elaborate them.

Structuralism

In a theory called structuralism, that everyday perceptual experience is structured or synthesized from "sensations," psychologists such as the English-U.S. introspectionistic psychologist Edward Bradford Titchener even devised a formal method of introspection for experimentally analyzing (or taking apart) percepts in an effort to reveal their constituent elements. The procedure required that the introspecting experimental subjects learn to avoid reporting on their experiences as they naïvely seemed. To establish this way of treating experience required careful training. One consequence of this training is that the observer's introspective reports may be contaminated by his expectations and hence may, in all honesty, reflect little more than his theoretical biases.

But the problem remains interesting: If percepts are indeed syntheses of simpler elements, can those elements be made to appear in experience? If so, what will they turn out to be? Can this problem be investigated without recourse to the structuralist's method of introspection and reliance on the reports of strongly biassed observers?

Evidence that percepts have constituent elements emerged serendipitously from research on stabilized retinal images. The image cast on the retina of the eye by a fixed object normally is continually moving because the perceiver himself is always in motion. Even when dampened by physical restraint, some residual movement will be left, attributable largely to high-frequency tremors (nystagmus) of the eyeballs. If the perceiver functioned as if he were a camera, the normal instability of the retinal image would produce a blurred percept and a concomitant impairment of visual acuity.

It is not feasible to eliminate eye movements, but it is possible to stabilize or fix the location of the retinal image by coupling the source of the image to the eyeball itself. An optical lever system can be so adjusted that when the eye moves the image source moves with it, and potential motion in the retinal image is eliminated. As expected, visual acuity is slightly enhanced when the retinal image is kept motionless. A remarkable, unexpected finding, however, was that such stabilized images rapidly seem to disappear, the perceiver losing awareness of them. It would seem that some movement in retinal image is needed to maintain perception over extended periods of time.

The optical lever system

One limitation of the optical lever system is that it permits the use of only very simple targets, such as straight, vertical lines. With a different device (in effect, a miniature projector attached to the eyeball), stabilized images of complex patterns may be presented. Complex patterns are found to produce percepts that are relatively slow to deteriorate; furthermore, they do not disappear in toto. The manner of the fragmentation is perhaps revealing of the way in which complex percepts are synthesized. Speaking metaphorically, observing how percepts "come apart" under retinal stabilization may be very much like discovering the structure of a rock by striking it with a powerful hammer blow.

Indeed, under retinal stabilization, single lines seem to disappear and reappear in a unitary (altogether) fashion. In a figure comprised of several lines (say, a square), percepts of parallel lines are likely to disappear and reappear together; proximity also affects the joint perceptual fate of pairs of lines. Retinally stabilized segments of such geometric figures as circles and triangles can seem to disappear and reappear without implicating the entire figure. In the disappearance of percepts of triangles, lines rather than angles are the functional units. (This finding is embarrassing to earlier theorizing about the crucial role of angles in the development of the neural network underlying the percept of a triangle.)

Clearly, with stabilized images, the constituent perceptual elements of complex geometric forms are lines, straight or curved; and lines with the same orientation are likely to have similar perceptual fates, as though forming

a higher-order component of complex patterns than do individual lines. These conclusions are remarkably similar to those drawn from studies of the effect of visual stimuli on the electrical activity of single neurons in the cerebral cortex. A finding of major theoretical significance is the failure of percepts of circles, squares, and triangles to act as units. Such percepts are treated in classical Gestalt theory, however, as though they are basic and unitary and not readily decomposable.

PRIMARY TENDENCIES IN PERCEPTUAL ORGANIZATION

Gestalt principles. Gestalt theory was meant to have general applicability; its main tenets, however, were induced almost exclusively from observations on visual perception. Whatever their ultimate theoretical significance, these observations have been raised to the level of general principles. It is conventional to refer to them as Gestalt principles of perceptual organization.

The overriding theme of the theory is that stimulation is perceived in organized or configurational terms (*Gestalt* in German means "configuration"). Patterns take precedence over elements and have properties that are not inherent in the elements themselves. One does not merely perceive dots; he perceives a dotted line. This notion is captured in a phrase often used to characterize Gestalt theory: "The whole is more than the sum of its parts."

The notion of "good" configurations

Of the many principles of organization that have been enunciated by various Gestalt theorists, the most general is referred to as *Prägnanz*. In effect, according to the principle of *Prägnanz*, the particular perceptual configuration achieved, out of a myriad of potential configurations, will be as good as prevailing conditions permit. What constitutes a "good" configuration, or a poor one, is unfortunately not clearly specified, though several properties of good configurations can be listed, chief among them being simplicity, stability, regularity, symmetry, continuity, and unity. What happens when these properties of figures come into conflict is not specified, but should be possible to determine empirically.

The principle of closure often operates in the service of *Prägnanz*; for example, a circular figure with small gaps in it will be seen as a complete or closed circle. Similarly, if a portion of the image of a figure falls on the blind spot of the retina, a complete figure often will still be perceived. Some distortions from good configuration may be so large as to preclude closure; in those cases, the figures may be a source of tension for the observer.

Prägnanz may also be achieved through good continuation; this principle describes a tendency for smooth continuity of contour to be dominant over discrete, irregular, abruptly changing contours. Thus, a figure composed of the overlapping outlines of an ellipse and a rectangle will probably be seen as such rather than as three figures, each with irregular, noncontinuous borders.

Closure and good continuation represent two of the factors that are held to determine what percepts will emerge from a complex stimulus. Implicit in them (and in the general principle of *Prägnanz*) is the assumption that whenever possible some figure will be perceived; more specifically, that the visual field will be articulated into figures and patterns of figures. It is understood that such emerging patterns are not *in* the stimulus. Although they are *permitted by* the stimulus, they are created by the perceptual system; that is, by the perceiver himself.

In the illustrations in Figure 2, in the panel on the left, the vertical distance between elements is less than the horizontal distance. By virtue of this differential proximity, the elements become perceptually organized into columns. In the right-hand panel, similarity, another principle of organization, is operative. Here, by virtue of similarity in brightness, the visual field tends to be perceptually articulated into alternating sets of black and gray rows.

It is not at all obvious why organization by similarity should occur; physical stimulation allows but does not demand it. Clearly in that case the articulation of the visual field into columns reflects a tendency in the perceptual system itself. Organization by proximity may not seem to reveal anything more than a close correspondence between perception and stimulation. (Though as

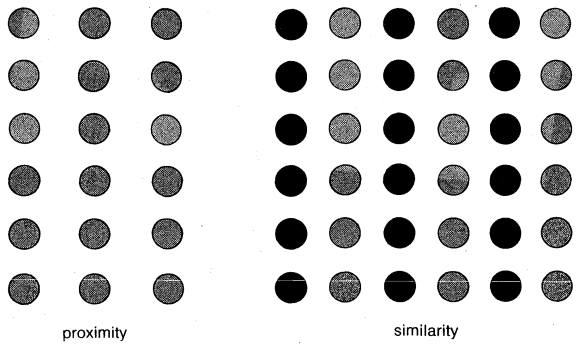


Figure 2: *Examples of Gestalt principles of organization.* (Left) Horizontal distance between dots is greater than vertical distance. (Right) equal distance between horizontal and vertical.

argued by the Gestalt theorist Kurt Koffka, it is not an adequate explanation to say that "things look as they do because they are what they are.") Yet, when a proximity pattern like the one shown in Figure 2 was briefly presented and subjects were asked (under guise of another task) to reproduce what they saw, many people failed to indicate a differentiated percept of columns. Instead, they reproduced a homogeneous matrix of elements. After repeated exposures, some of those subjects began to draw proximity-based columns of elements. Organization according to the principle of proximity seems to be neither universal nor, for those who achieve it, immediate.

Proximity patterns

In the latter experiment, people who failed to obtain the differentiated percept of columns scored significantly lower on a test of verbal intelligence than did those who succeeded at some point in the experiment. Perhaps the Gestalt principles of organization apply to perceivers (such as Gestalt theorists) whose intellectual development has reached a high degree of maturity. When organized percepts are easy to come by, gradations in intelligence do not seem to matter; when some barrier to organization is imposed (as by brief stimulus exposure), however, then the effect on perception of such differences among individuals may show up.

If people see a pattern of columns in the left panel of Figure 2 because that is how the stimulus is constructed, then why do some people not see it that way? Both achieving and failing to achieve an organized percept must be explained. Surely, part of the explanation must lie in the nature of the perceptual process itself. Thus, the experimental results indicate that perceptual organization is not universal and immediate; rather, they support the major tenet of Gestalt theory that things look as they do because of the organization imposed by the perceptual process (*i.e.*, by the perceiver).

One Gestalt principle, that of common fate, depends on movement and is quite striking when observed. According to the principle of common fate, stimulus elements are likely to be perceived as a unit if they move together. An illustration of this principle is provided by a well-camouflaged object, such as a military vehicle; when stationary, the elements of the vehicle are integrated, through proximity, similarity, and so on, into patterns of background elements, and the object is difficult to detect. But it is easy to see it once it starts moving; with all of its elements moving in unison, the vehicle is readily perceived as a unitary figure, clearly segregated from its background.

Gestalt views of movement

Movement is also at the heart of a set of observations of considerable significance in the historical development of Gestalt theory. These observations concern circumstances in which people perceive movement in the absence of actual physical motion of the stimulus. One familiar instance of this class of events is referred to as the phi phenomenon. In simplest form, the phi phenomenon can be demonstrated by successively turning two adjacent lights on and off. Given appropriate temporal and spatial relations between the two lights, an observer will perceive the first light as if it were moving from its location to that of the second light. The phi phenomenon is basic to the eye-catching displays used on theatre mar-

quees and to cinematic and television presentations. The motion-picture screen, for example, presents a series of briefly flashed, still images; the movement people see is a creation of their own perceptual systems (see also PERCEPTION OF MOVEMENT).

It is the lack of one-to-one correspondence between stimulation and perception, as dramatically illustrated in the phi phenomenon, that underscores the Gestaltists' dissatisfaction with stimulus-bound models of perception and their insistence on the priority of patterns and relations. What people perceive is determined not only by what is present at the point under direct observation but also by what is occurring in the total stimulus context or display.

Context effects. One of the simplest instances of relational (or context) effects in perception is that of brightness contrast. Thus, the apparent brightness of a stimulus depends not only on its own luminance but also on that of the surrounding stimulation. The same gray square looks whiter against a dark background and blacker when placed in a bright surround. Similarly, a white or gray patch will take on an apparent hue that is complementary to the colour of the surround (*e.g.*, the patch will seem tinged with yellow when it is placed against a blue background).

Analogous context effects

Analogous context effects are evident in many commonplace experiences. A man of average height seems to be a runt when he is on a basketball court with much taller players; yet the same man looms like a giant when refereeing a game played by little boys. It is known that a typical winter's day seems delightfully balmy when temperatures rise after a week of subfreezing weather.

To the Gestaltist, contrast effects dramatize the relational nature of perception. They also play a significant role in a more recently developed adaptation-level theory, which also provides a general perceptual model. At the core of the model is the notion that the manner in which a stimulus is perceived depends not only on its own physical characteristics but also on those of surrounding stimuli and of stimuli previously experienced by the observer. In other words, the perceiver is said to be perceptually adapted to past sensory stimuli; his adaptation level forms a kind of zero point against which any new stimulus is perceived. An example is provided by the almost overwhelming silence one experiences when the sound of an air conditioner (to which he has adapted) suddenly ceases.

Gestalt theorists also attached significance to the observer's history of stimulation; indeed, some of them interpreted so-called figural aftereffects within a Gestaltist model of brain functioning. Figural aftereffects refer to changes in the perceived shape or location of a figure following its inspection; for example, a curved line will appear to get straighter after prolonged inspection. Or the distance between two parallel lines seems to change as an aftereffect of previous inspection.

In a typical experiment one looks at a point adjacent to a dark vertical bar (the inspection figure) on a screen. Following this inspection period, the dark bar is replaced by two identical pairs of vertical lines, one pair on either side of the region where the bar had been, the second pair alongside in a region not previously exposed to the inspection figure. The subject again fixates the same point. A figural aftereffect shows up as a greater apparent distance between the pair of lines surrounding the region of the inspection figure even though the other pair is actually identical. This distortion is not simply a generalized contrast effect because it occurs only in the small area along the borders of the inspection figure; that is, the effect is localized and restricted.

Localized satiation in the brain

It thus has been speculated that visual exposure to a figure induces in the brain a condition of localized satiation. The passage of electrical activity is assumed to be impeded in satiated areas of the brain. Moreover, it is postulated that the perceived distance between two borders of a figure is directly related to the time it takes for electrical currents to pass between them. Thus, it is held that one effect of satiation is to increase the apparent distance between the borders of a figure that straddles a satiated region. Whatever the merits of such physiological speculations, they have stimulated a vast amount of research on figural aftereffects. Good evidence for similar effects in other senses, such as touch, also has been obtained. Clearly, perception can be influenced not only by the context of current background but also by the residues (aftereffects) of previous stimulation.

Concurrent visual stimulation may modify one's acuity in detecting auditory stimuli. Similar interactions are claimed to occur for other combinations of senses. Some dentists report success in using audioanalgesia, in which stimulation with sound waves is said to reduce the experience of pain in the mouth. The high specificity of some of the reported sensory interactions seems to preclude an explanation that concurrent stimulation works by changing the subject's general level of alertness. However these intersensory effects might be mediated, they do suggest that the brain does not function as a collection of entirely independent sensory channels. As a physical system, the brain follows physical principles; thus overlapping and spreading or waning fields of neural excitation in the brain have been theorized to underlie such phenomena as closure and audioanalgesia. Köhler referred to these models of neural analogues of perceptual phenomena as physical Gestalten; unfortunately, there is little direct physiological evidence for them.

An alternative to field effects in brain functioning is the assumption that local stimulation gives rise, in one-to-one fashion, to a mosaic of local responses. Implicit in the mosaic hypothesis is a kind of telephone switchboard model of the brain as a machine in which the electrical activity is strictly confined to separate pathways of neurons that are well insulated (isolated) from one another. The Gestaltists rejected this model because in its early formulations it did not explain intersensory and intrasensory perceptual phenomena. A more sophisticated machine model, however, provides for fieldlike effects through the operation of complex networks of neural elements. It is held that electrical activity remains confined to discrete pathways, but that these pathways do not simply travel straight through the system; that they also interconnect, with both excitatory and inhibitory consequences. Supporting evidence comes from records of the electrical activity in single neurons in the cat brain; when the cat's eye is probed by a small spot of light, a specific area on the retina can be found that serves to excite a given brain neuron.

Further mapping of the cat retina often uncovers inhibitory areas adjacent to the one that is excitatory; that is, when light strikes those retinal areas the activity in the brain neuron being monitored is depressed. The excitatory and inhibitory areas thus comprise the brain neuron's retinal receptive field. Analogous inhibitory effects have also been found in research on the eye of the crab, *Limulus*. Such context effects as brightness contrast could be based on these simple inhibitory mechanisms. It remains to be seen, however, just how many perceptual phenomena that fit Gestalt field theory also can be handled by sophisticated variants of the machine or mosaic model.

Perceptual constancies. Even though the retinal image of a receding automobile shrinks in size, the normal, experienced person perceives the size of the object to remain constant. Indeed, one of the most impressive features of perceiving is the tendency of objects to appear stable in the face of their continually changing stimulus features. Though a dinner plate itself does not change, its image on the retina undergoes considerable changes in shape and size as the perceiver and plate move. What is noteworthy is stability in perception despite gross instability in stimulation. Such matches between the object as it is perceived and the object as it is understood to actually exist (regardless of transformations in the energy of stimulation) are called perceptual constancies.

Dimensions of visual experience that exhibit constancy include size, shape, brightness, and colour. Perceptual constancy tends to prevail for these dimensions as long as the observer has appropriate contextual cues; for example, perception of size constancy depends on cues that allow one a valid assessment of his distance from the

object. With distance accurately perceived, the apparent size of an object tends to remain remarkably stable, especially for highly familiar objects that have a standard size. Thus, people's heads all tend to look the same size regardless of distance; similarly, an object identified as a lump of coal tends to look black even when intensely illuminated.

Loss of perceptual constancy

The experience of constancy may break down under extreme conditions. If distance is sufficiently great, for example, the perceived size of objects will decrease; thus, viewed from an airplane in flight, there seem to be "toy" houses, cars, and people below. To the extent that they prevail, the constancies lend the perceiver's experience and behaviour relative stability. Imagine an alternative, kaleidoscopic perceptual world in which everything seems to change, solid objects apparently swelling, shrinking, and warping with every movement. Breakdown in perceptual constancy seems to complicate the course of some psychiatric disorders in which the perceptual boundary between the sufferer and the external world is weakened. Normal constancies also can be intentionally overcome, as in paintings of flabby watches and distorted people that apparently depict the unique perceptual world of the artist.

INDIVIDUAL DIFFERENCES IN PERCEIVING

Theoretical assertions about perceiving are often made as though they apply indiscriminately to all organisms, or at least to all people. Perhaps perceptual principles of such great generality eventually will be uncovered. In the meantime it is evident that there are clear differences in perceptual functioning among individuals, among classes of individuals, and within the same individual from one occasion to another.

Age. That perceptual functioning should change with the perceiver's age is expected on the grounds that psychological development stems from maturation and learning. Indeed, empirical evidence for age-related changes in perceiving is substantial. There are, for example, reliable data that perceptual constancies are enhanced with the person's increasing age, improvement leveling off at about age ten. Similarly, there is a great deal of evidence for both decreased and increased susceptibility to various optical illusions with increasing age (see ILLUSIONS AND HALLUCINATIONS). Those illusions that become less pronounced with increasing age probably depend on the subject's changes in scanning and on his increased ability to segregate parts of a pattern from one another; illusions that become more pronounced probably reflect the operation of expectancies that develop through experience. Anatomical and physiological changes in the eye itself also may account for some age-related perceptual changes.

Optical illusions among older persons

Sex. It is difficult to assess the degree to which differences related to the sex of the perceiver are biologically based or are the cultural product of traditional differences in sex role. Biological sex and sex role thus far have been hopelessly confounded in experiments with human subjects.

Sex differences in perceiving, whatever their basis, can be illustrated in research on differences in the style with which people perceive. This stylistic difference emerges in extremes of response to context. If a person perceives the world as highly differentiated, he tends to resist contextual influences and is said to be field independent; the person who perceives in an extremely diffuse style, the field-dependent individual, tends to be highly susceptible to contextual effects. Thus, field-independent people are superior in locating a simple visual figure (*e.g.*, a triangle) embedded in a complex pattern; similarly, field-independent subjects can better adjust a rod in a tilted frame to the true vertical when no other visual cues to verticality are present.

Both age and sex are found to be implicated in these differences in perceptual style. Specifically, field dependence declines with increasing age, as does the closely related susceptibility to optical illusions. In North American studies, female subjects tend to be more field dependent than are males, especially after puberty. Perhaps these results are distinctive of cultures in which females are at least implicity trained to be passive and perceptually diffuse, and in which males are encouraged to assume an active, perceptually articulated stance. This hypothesis has received some support in studies of the parent–child interactions characteristic of the early years of the two types of subject.

Culture. Beyond sex differences in perceiving that seem to be culturally imposed, there is evidence for more general cultural influences on perception. The burden of much research is to show that the type of physical environment people construct for themselves or choose to inhabit can influence their style of perceiving. There are African groups (*e.g.*, Zulu and Bushmen), for example, whose environments are virtually lacking in rectangular forms, by contrast with the carpentered, right-angled world of people in Western cultures. People in these African groups also fail to make use in their art work of two-dimensional representations of three-dimensional objects. Such differences in visual environments show up in tests of susceptibility to illusions. Zulu and Bushmen subjects are relatively resistant to those visual illusions that depend for their effectiveness on the subjects' treating the lines comprising the pictures as borders of three-dimensional, rectangular objects. Analogous effects with different classes of illusion have been shown for other peoples who live in a perceptually unique environment.

Expectancies, motives, and conflicts. The inherent ambiguity of much sensory stimulation is coupled with a strong tendency of people to impose order and meaning on their experience. It may thus be predicted that a person's perceptions may be influenced by such factors as his expectancies, needs, values, and conflicts. The cross-cultural differences between, say, Europeans and Zulus in susceptibility to illusions probably arise from differences in expectancies. The same can be said for the ease with which one becomes aware of Lincoln's face in Figure 1 once it has been perceived or even after simply being told what the picture is supposed to be.

Psychoanalytic theory explicitly calls for motivational influences on such functions as memory, thinking, and perceiving (see MOTIVATION). In particular, the theory is concerned with unconscious motives and conflicts and with unconscious defenses (such as repression) used to control them. According to the psychoanalytic hypothesis, there should be wide perceptual variation among individuals in response to stimuli that have motivational significance. At any rate, a host of experiments have been designed to show that perceiving is indeed subject to unconscious influences.

Studies using obscenities

In some studies, for example, it seemed that so-called obscene words flashed on a screen had to be exposed longer than apparently neutral (control) words before their meaning could be perceived. In the other studies, children of poor families have been found to overestimate the size of coins as compared with the judgments of children of richer families. One major problem with such research lies in finding or creating appropriate experimental and control stimuli. Considering differences in the use of language, for example, it is most unlikely that what once were widely called obscene words would currently evoke the conflicts and defenses of more than a few subjects.

Assuming suitable stimuli can be found, an even more serious problem arises around the interpretation of the subjects' behaviour; for example, do people really find it more difficult to recognize obscene words or are they simply reluctant to admit recognition? Problems of this sort have plagued researchers, and unambiguously interpretable experiments in this field are most difficult to produce. The hypothesis of such individual influences as motivation on perception remains appealing and viable, but unproved.

BIBLIOGRAPHY. For a clear presentation of Gestalt theory and its case against structuralism and behaviourism, see W. KOHLER, *Gestalt Psychology* (1959). A comprehensive historical overview is available in E.G. BORING, *Sensation and Perception in the History of Experimental Psychology* (1942). Illustrations and discussions of Gestalt principles of organiza-

tion along with material on illusions, context effects, and related phenomena are provided in W.N. DEMBER, *The Psychology of Perception* (1960); J.J. GIBSON, *The Perception of the Visual World* (1950); and J.E. HOCHBERG, *Perception* (1964). Implications of research on early experience for perceptual and intellectual development are spelled out in J.M. HUNT, *Intelligence and Experience* (1961). Two excellent collections of technical articles on perception, covering a wide range of topics, have been compiled and edited by R.N. HABER: *Contemporary Theory and Research in Visual Perception* (1968), and *Information-Processing Approaches to Visual Perception* (1969). For a scholarly discussion of depth perception and a lucid description of an elegant series of experiments, see B. JULESZ, *Foundations of Cyclopean Perception* (1971). The potential contributions of neurophysiological research to an understanding of perception are suggested in N. WEISSTEIN, "What the Frog's Eye Tells the Human Brain: Single Cell Analyzers in the Human Visual System," *Psychol. Bull.* 72:157–176 (1969).

(W.N.D.)

Perception of Movement

Most animals (including people) move in search of food that itself often moves; they move to avoid predators and to mate. Animals must perceive their own movements to balance themselves and to move effectively; without such perceptual functions the chances for survival would be sharply reduced.

VISUAL CUES

The eye is by far the most effective organ for sensing movement. Some animals are especially sensitive to visual stimuli that move in specific ways. For instance, electrical patterns from the eye of a frog show that some elements in the organ respond only when the stimulus is about the size of a fly moving in the insect's range of speed. Generally the eyes of lower animals seem to respond selectively to what is of importance to survival. In these animals the eye's retina does much of the visual processing. This is an economical arrangement since the animal tends to respond only to essential stimuli, the brain having little to do but relay signals to the motor system. It is an inflexible mechanism, however; higher animals process visual information in more elaborate ways, the brain being more heavily involved. Thus, some cells in the visual area of the cat's brain respond only to moving stimuli, sets of movement-detector cells functioning specifically for each direction across the field of view. Features of human visual experience also suggest that movement detectors exist in the human brain.

Eye and brain

Each retina in most higher animals has a central (foveal) zone for detailed colour and pattern vision and a surrounding peripheral zone that effectively is sensitive only to the grosser features of the outer visual field. The peripheral retina is especially sensitive to movement (often a signal of danger), which induces a reflex turning of the eyes to project the image on the fovea and permit the moving object to be recognized.

Visual stability. Mechanisms have evolved that yield stable, clear visual input despite swaying and other blurring factors. In a reflex mechanism called optokinetic nystagmus, the eyes pursue a moving scene to keep the image stationary on the retina. When they can move no farther, they snap back and pursue the scene again in a to-and-fro alternation of slow pursuit and quick return. These eye movements are readily observed in people who are looking at a moving pattern of stripes or turning their heads, this response being inhibited only when something stationary is visually fixated.

Similar nystagmic movements are triggered by impulses arising in the inner ear when the head moves. These persist even when the eyes are closed and may be felt by pressing the eyelids lightly as one rotates the whole body.

Nystagmus; saccades

In a related stabilizing activity the eyes scan in quick jerks (saccades) with short fixations; *e.g.*, in reading. Normally the eyes cannot move steadily over a stationary scene but make a series of stationary images (like still photographs); visual function tends to be suppressed when there is saccadic blurring. Yet the eyes can follow a steadily moving object smoothly.

When one looks from one point to another, movements of the retinal image are the same as those produced by a moving scene on a stationary eye. It might be thought that the sensory structures found in the eye muscles would provide the cues for judging whether it is the eye or the scene that has moved. Yet we see the scene as stationary only when we move our eyes voluntarily and not when they are moved passively by the finger. This suggests that motor-nerve signals inform us whether our eyes are moving, rather than the sensory structures in the eye muscles. When the eye is moved by pushing it with the finger there is no normal motor discharge to inform the brain, and changes in retinal image are perceived as movement of the scene. Indeed, people with paralyzed eye muscles experience the scene as moving when they try to move their eyes. When the motor discharge thus generated is not accompanied by the expected image motion, the person falsely perceives the scene and the eye to be moving together.

Relative visual movement. A visual field containing familiar objects provides a stable framework against which relative motion may be judged. People often report that an isolated point of light in a dark room is moving when it is not; the experience is known as autokinetic movement. It was observed in 1799 by Alexander von Humboldt while he was watching a star through a telescope, and he attributed it to movement of the star itself. Not until about 60 years later was the effect shown to be subjective, apparently arising from instability in the sense of eye position without a visual frame of reference.

Autokinesis; induced movement

Similarly, if a small object is presented in a frame with nothing else in view, movement usually is attributed to the object even when only the frame moves. This induced movement effect reflects our tendency to use the larger surround as a stable frame of reference. Recall the illusion that your train is moving when it is really the moving train alongside that, seen through the window, is falsely accepted as the frame of reference.

People cannot perceive very slow movement; below a minimum speed (about that of the minute hand on a watch) movements become imperceptible and can only be inferred (as in remembering the previous position of the hour hand). At high speeds, one perceives a blurred streak rather than a definite object in motion.

Movement aftereffect. When a parade is interrupted after some minutes, the pavement may seem to move in the opposite direction to the marchers who have passed. Phenomena similar to this movement aftereffect occur in other senses. For instance, after disembarking, a sailor feels the land to be rolling like a ship as the result of kinesthetic and vestibular aftereffects. The visual movement aftereffect probably arises when movement detectors in the brain that respond to the original direction of motion become fatigued, leaving predominant those detectors that respond to contrary movement.

Apparent movement. Motion-picture film is a strip of discrete, still pictures but produces the visual impression of continuous movement. Stationary light bulbs coming on one after the other over the theatre entrance also produce an impression of steady movement. In part, such effects of apparent movement (called the visual phi phenomenon) depend on persistence of vision: visual response outlasts a stimulus by a fraction of a second. When the interval between successive flashes of a stationary light is less than this visual-persistence time, the flicker will appear to fuse into a continuous light. The flicker frequency at which this occurs is called the perceiver's flicker-fusion frequency (or critical flicker frequency) and represents the temporal resolving power of his visual system at the time. Another process on which apparent movement depends is a tendency (called visual closure or phi) to fill in the spaces between adjacent visual objects. This means that the movement detectors of the visual system are triggered as effectively by a closely spaced pair of lights alternately going on and off as by a single light moving back and forth. It would seem that two aspects of visual function (flicker fusion and phi) make the motion-picture industry possible.

Flicker fusion; phi

Stroboscopic effect. When a rotating electric fan is illuminated by a flashing light source (called a strobo-

scope) so that a flash arrives whenever a fan blade passes a fixed position, the blades will seem to stand still. This is a useful way of observing fast-moving objects such as machinery or insect wings. If the flashes occur less frequently, the object will seem to move slowly in its actual direction; when the flashes arrive more frequently, the object will seem to move backward, as stagecoach wheels may do in the cinema.

Visual movement in depth. An object moving directly away from an observer provides fewer visual cues of movement than it would by moving across the field of view. However, changes in retinal-image size are produced that give a clue to its movement. Thus a stationary, but shrinking, luminous object in the dark is seen as if it were receding. Other clues to movement in depth are changes in the convergence angle of the two eyes, in the focussing mechanism, and in the haziness and brightness of the object.

NONVISUAL CUES TO MOVEMENT

Auditory. Sound waves travel well in water, and fish rely heavily on acoustic cues to detect moving objects. Land animals, although typically more visually oriented, also make some use of such cues, including changes in intensity (loudness) and small differences in the time at which the wave reaches each ear. Some animals (*e.g.*, rabbit, horse) have mobile external ears that track moving sound sources. Bats vocalize high-frequency waves, detecting objects by sonic reflection (similar to sonar).

Kinesthetic. Kinesthesis here refers to experiences that arise during movement from sense organs in the membranes lining the joints and from the sense of effort in voluntary movement; receptors in muscles seem to have little role in the perception of bodily movements. Depending on speed of motion and the joint involved, blindfolded people can detect a passive joint movement as small as a quarter of a degree. People vary widely in the accuracy with which they can actively produce movement of a given extent; this ability also varies with direction of movement and the opposing friction, mass, and springiness.

Phantom limb

Kinesthetic perception may persist for a limb that has been amputated, giving rise to a hallucinatory experience known as the phantom limb. The patient may experience vividly the "movement" of the absent part; a recent amputee may attempt to stand on his missing leg or to grasp with his missing hand.

Vestibular system. Vestibular structures, enclosed in a fluid-filled cavity in the region of each inner ear, include the utricle, a small sac containing minute sensitive hairs associated with tiny sandlike granules called otoliths. The utricle functions as a linear accelerometer. When the head tilts relative to gravity or is accelerated, the relatively dense otoliths deflect the hair cells and nerve impulses are transmitted to the brain. At constant velocity the otoliths become stable, stimulation ceases, and a person must rely on other cues (*e.g.*, by observing the passing scene) to detect his motion.

Equilibrium

Vestibular structures for each ear also include three fluid-filled semicircular canals, each in a different plane. Each canal has a swelling (ampulla) that contains the cupula, a cluster of sensitive hairs embedded in a jellylike mound. As the head moves in the plane of a given canal, motions of the fluid deflect the cupula to produce nerve impulses. These travel through the brainstem to other brain and spinal centres that mediate equilibrium or balance and that generate nystagmic eye movements.

Taken together, the semicircular canals serve as a rotary accelerometer. If a person is rotated at constant velocity and then is suddenly stopped, the cupula is redeflected to give a feeling of rotation in the opposite direction; this also gives rise to dizziness and postrotational nystagmus. Dancers and skaters learn to overcome such effects by concentrating on some fixed, visible object; with their eyes closed they are as likely to fall as anyone.

Overstimulation of the vestibular system (*e.g.*, on a ship or airplane) may induce motion sickness. A person with vestibular function totally destroyed is not subject to motion sickness; but if the vestibular mechanism is impaired only on one side, each movement of the head can be nauseating. Such a patient takes a long time to compensate for this imbalance.

In outer space there is no gravitation to mediate feelings of up or down, although these still may arise from visual cues. The utricles and vestibular canals still respond to movements of the head, however, and serve orientation within the spacecraft.

BIBLIOGRAPHY. J.J. GIBSON, *The Senses Considered as Perceptual Systems* (1966), a readable book presenting a major theoretical viewpoint; R.L. GREGORY, *Eye and Brain: The Psychology of Seeing* (1966), a well-produced popular account, see esp. ch. 7; I.P. HOWARD and W.B. TEMPLETON, *Human Spatial Orientation* (1966), a comprehensive review with a detailed key to the scientific literature; I.M. SPIGEL (ed.), *Readings in the Study of Visually Perceived Movement* (1965), a very useful collection of articles from scientific journals, understandable to the layman.

(I.P.H.)

Perciformes

The order Perciformes, the perchlike fishes, is the largest group of fishes in the world today, comprising over 6,000 species that have been classified into about 150 families. Perciform fishes occur in abundance in both marine and freshwater areas of the world, ranging from shallow freshwater ponds to depths of over 2,300 metres (7,500 feet) in the oceans. Most perciforms are marine fishes, generally found along coastal areas of tropical and temperate regions of the world. The order includes many of the world's most important food and game fishes, such as tunas, mackerels, bonitos, and skipjacks (family Scombridae), billfishes and marlins (Istiophoridae), swordfish (Xiphiidae), sea basses (Serranidae), and carangids (Carangidae), a large family that includes pompanos, jacks, cavallas, and scads. The freshwater food and sport fishes of the perciform order include the sunfishes (Centrarchidae) and the perches and walleyes (Percidae). Many perciforms are popular aquarium fishes.

GENERAL FEATURES

Size range. Perciform fishes comprise a wide variety of sizes, ranging from the tiny freshwater goby *Pandaka pygmaea* (Gobiidae) of the Philippines, which is fully grown at about 1.2 centimetres (less than one-half inch) in length, to the black marlin (*Makaira indica*), swordfish (*Xiphias gladius*), and the bluefin tuna (*Thunnus thynnus*), which attain lengths of about 3.3 metres (11 feet). The bluefin tuna and the Indo-Pacific black marlin have been known to exceed 680 kilograms (1,500 pounds) in body weight. Generally, most percoid fishes fall within the range of 30 to 250 centimetres in length.

Distribution. Perciform fishes occur worldwide and are clearly a highly successful group. Coral reefs of tropical seas abound with colourful perciforms, such as wrasses, butterfly fishes, gobies, damselfishes, blennies, and cardinal fishes. The perciform order comprises a large part of the fauna of the Indo-West Pacific region, which is probably the world's richest in the variety of its fish fauna. Of the Antarctic fish fauna, approximately 75 percent belong to the Perciformes. These include the icefishes (Chaenichthyidae), known for their "bloodless" appearance, which results from the lack or near lack of red blood cells and blood pigments. Freshwater perciforms include the cichlids (Cichlidae), which occur naturally in India, Africa, South America, and parts of southern North America; these fishes also have been introduced elsewhere. The perch and sunfish families are found in North America and Europe, and the European perch (*Perca fluviatilis*) occurring well north in Siberia.

IMPORTANCE TO MAN

Use as food. Since early times, the rivers and oceans have provided man with food; fishing was one of man's earliest means for securing food. Archaeological findings among shell mounds of Scotland include the sea bream (Sparidae) as part of the diet of early man. The Nile perches (Latidae) have been found as mummies in ancient tombs in Egypt. The goatfishes (Mullidae) appear in ancient Roman archives as one of the most highly

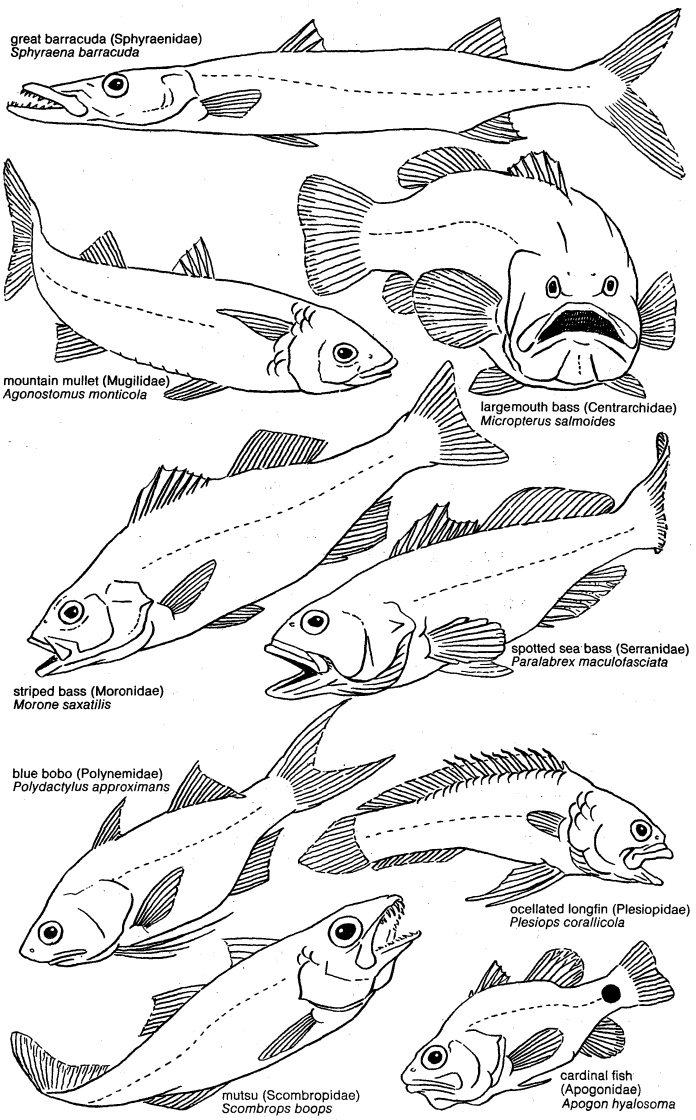


Figure 1: Representative perciforms of the families Sphyraenidae, Mugilidae, Polynemidae, Scombropidae, Apogonidae, Moronidae, Serranidae, Plesiopidae, and Centrarchidae.
Drawing by G. Christman

valued food fishes, and in Japan the goatfish holds a good market and is eaten raw as *sashimi* or as dried fish cakes known as *kamaboko*. In Japanese art through the ages, the fish god is shown with the "king of sea fish" under one arm; this highly valued food fish is the porgy *Chrysophrys major* (Sparidae). A Japanese New Year's dinner usually includes *buriko*, the eggs of the sandfish *Arctoscopus japonicus* (Trichodontidae).

The perciform fishes play an important part in commercial fisheries all over the world. The fishery statistics of the Food and Agricultural Organization of the United Nations show that the catch for tunas, bonitos, billfishes, and various tuna-like fishes exceeds 1,500,000 metric tons; a similar catch was recorded for chub mackerel (Scombridae) alone during 1970. Catches for sea breams and grunts (Pomadasyidae) exceed 100,000 metric tons, and the catches for freshwater perciforms were around 4,000 metric tons for cichlids and 2,000 metric tons for the yellow perch, *Perca flavescens*. Other important food fishes include groupers (Serranidae), snappers (Lutjanidae), carangids, and dolphins (Coryphaenidae). Isinglass, which is used in the production of jellies and also in the process of clarification of wine and beer, is obtained from fishes that include the drums (Sciaenidae) and the threadfins (Polynemidae). The skin of the wolffishes (Anarhichadidae) provides a leather of fair quality. The guanin present in the skin of the Japanese cutlass fish (*Trichiurus*; Trichiuridae), is used in the manufacture of artificial pearls in Japan.

Breeding and cultivation of perciforms has been successful in many parts of the world. The African mouthbreeder (*Tilapia macrocephala*, Cichlidae) has been successfully introduced in many areas and is valued for its rapid rate of reproduction and growth, providing a source of low cost protein.

Aquarium fishes. Colourful and interesting perciforms are kept for aesthetic reasons by aquarists, augmenting an industry partially supported by fishes of other orders. Popular aquarium fishes of the perciform order include cichlids, butterfly fishes (Chaetodontidae), angelfishes (Pomacanthidae), labyrinth fishes (suborder Anabantoidei) such as the Siamese fighting fish (*Betta splendens*) and the kissing gourami (*Helostoma temmincki*), and various gobies (Gobiidae), blennies, and blenny-like fishes of the suborder Blennioidei.

The freshwater angelfish *Pterophyllum scalare* and the discus (*Symphysodon discus*) are among the most popular aquarium fishes for breeding because of their remarkable means of feeding their young on the mucous secretions of their bodies.

Poisonous perciforms

Tropical food poisoning, or ciguatera, has been attributed to some perciforms that are otherwise considered to be excellent food fishes. Among these are certain carangids, snappers, barracudas (Sphyraenidae), surgeonfishes (Acanthuridae), groupers, and porgies. A species completely edible in one area may be poisonous in an area just a few hundred miles away. This curious phenomenon has not yet been fully explained, although it has been suggested that the source of poisoning may be a toxic form of blue-green alga passed up the food chain and thus present in the food of toxic species.

Danger to man. A few of the perciforms are known to be harmful to man. Swimmers have been attacked by the barracuda (*Sphyraena*), which is a voracious fish reaching nearly two metres (six feet) in length. Perciforms possessing venom glands are also considered dangerous fishes. The dorsal spine of the weever fishes (Trachinidae) has a grooved structure containing a venom gland; in addition, there is also a stinger located on the opercular (gill cover) structure. Both the stinger and the dorsal spine can be extremely painful if stepped on in shallow waters. Similar venom-bearing structures are found in the dragonets (Callionymidae) and surgeonfishes. The venomous spines in the surgeonfish are located on either side of the caudal peduncle (the narrow stalk just in front of the tail). Especially well armed are the electric stargazers (*Astroscopus*, Uranoscopidae), which are capable of discharging up to 50 volts of electricity from the modified muscle tissue just posterior to the eyes; in addition, they possess a venom spine just above the pectoral fins. The venom from uranoscopids has been known to cause death in man.

NATURAL HISTORY

Life history. Many perciforms live out their whole lives in small areas, but others, especially open-ocean (pelagic) species, perform extensive migrations, about which much remains to be learned. Some marine serranids, however, are anadromous (*i.e.*, entering fresh or brackish water to spawn); some freshwater perciforms, such as certain species of gobies, enter the sea to spawn (catadromous). Tuna (Scombridae) may travel across the entire Pacific Ocean from the California coast to Japan, or the reverse. Spawning in perciforms generally takes place in shallow coastal areas or in rivers and ponds among rocks, seaweeds, and aquatic plants. *Paraclinus marmoratus*, a clinid blenny, is known to lay eggs at times in the lumen (cavity) of a living sponge.

Breeding behaviour among perciform fishes is diverse. Pairing of male and female is common, although a single female may pair with more than one male, as among certain serranids, perches, and cichlids. The sexes are usually distinct, but hermaphroditism (presence of functional male and female organs in a single individual) normally occurs among certain sea basses and porgies.

The young of the black sea bass (*Centropristis striatus*) are mostly females with normal egg-laying functions; after five years, however, some of these females transform into functional males. About 11 species of sparids have been found to display hermaphroditism at one time or throughout their lifetime.

Differences in appearance between sexes

Characteristic differences usually exist, especially during the breeding season, between sexes regarding colour, size, markings, or structure. The male is generally smaller in size (some exceptions are found in sunfishes, gobies, and blennies) and has brighter coloration of the fins and body. Black, white, green, red, blue, and silver are colours characteristic of the brightly coloured males of damselfishes (Pomacentridae), wrasses (Labridae), labyrinth fishes, and cichlids. Structural differences between the sexes vary from easily observed characteristics, such as the presence of a longer dorsal fin in the male dragonets, to less obvious characteristics, such as larger canines in the dentition of male blennies (Blenniidae) and gobies. Bands, blotches, and tinged markings may also be characteristic of the brighter coloured males, as in some wrasses. During the spawning season, males of certain species of cichlids, parrotfishes (Scaridae), sparids, and wrasses develop a swelling on the forehead that may persist throughout life. These characteristics presumably enhance the male in its breeding and courtship activities.

Drawing by G. Christman

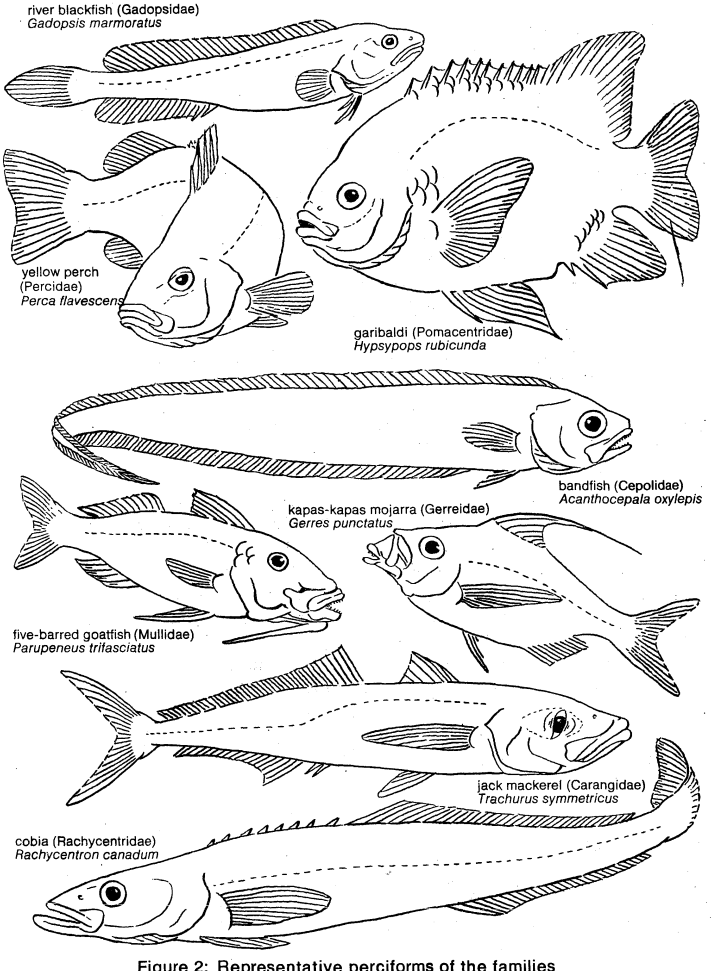


Figure 2: Representative perciforms of the families Pomacentridae, Percidae, Cepolidae, Mullidae, Gerreidae, Carangidae, Rachycentridae, and Gadopsidae.

Little is known of the courtship activities in most marine perciforms. The male dragonet, with extended fins and gill covers, performs a display of colours while swimming repeatedly around the female. A similar courtship activity is also carried out by the fighting fishes (*Betta*).

Fertilization is usually external, although internal fertilization, in which the eggs are fertilized within the body of the female, is well known in such groups as surfperches (Embiotocidae); and the male in such cases generally possesses an intromittent organ, which functions in the transfer of sperm to the female. Internal fertilization also occurs among some of the gobies, clinid blennies (Clinidae), apogonids, and eelpouts. Perciforms that undergo internal fertilization mostly are viviparous; *i.e.*, they give birth to live young. The number of young in viviparous perciforms varies from three to 50 in the surfperches. The larger viviparous forms, however, have been found to produce a greater number of young.

Most perciform fishes are oviparous—*i.e.*, they lay eggs that are fertilized externally. The number of eggs laid varies from a few hundred to more than 3,000,000 in a 32-pound (15 kilogram) yellowtail (*Seriola dorsalis*, Carangidae). Often, the eggs are released to float freely, but many species have evolved elaborate nest-building behaviour. Nest construction frequently consists merely of clearing away of a small area under rocks, which may be on the open bottom or even inside empty animal shells. The sunfishes and darters (Percidae) use their fins or body to dig a circular depression for use as a nest. Wrasses construct a nest out of stones, shells, and seaweed. Males generally undertake the task of building the nest, but in many cases both male and female share the labour.

Nest-building behaviour

Most labyrinth fishes build bubble nests, the procedure being similar among members of the family. The male Siamese fighting fish takes a bubble of air into his mouth, coats it with a mucous secretion, then blows the coated bubble to the surface; this process is repeated until a bubble nest is formed. After pairing, the female allows the fertilized eggs to drop to the bottom, where the male picks them up in his mouth and blows them into the bubble nest. Many marine perciforms produce pelagic eggs; *i.e.*, that float on or near the surface of the open sea. Almost all of the freshwater perciforms produce demersal eggs (*i.e.*, that sink to the bottom). A certain amount of adhesiveness in the demersal eggs keeps the eggs together in clusters; the elongated shape of the clusters of perch eggs help in securing the clusters to aquatic plants and rock bottoms. Not all eggs become attached to aquatic plants and other objects; the mature male humphead (Kurtidae) possesses a hooked structure on his forehead, to which the cluster of eggs is attached as soon as the female produces them. Oral incubation, in which the eggs are held in the mouth of one of the parents, is found in certain species of cardinal fishes (Apogonidae), jawfishes (Opisthognathidae), labyrinth fishes, and cichlids. The male, female, or both may incubate the eggs orally until they hatch, after which the young may be mouthbreeders. Similarly, guarding of the nest sites may be undertaken by the male, female, or both parents; however, males of some sunfishes, darters (*Etheostoma*), and the Siamese fighting fish defend their nest against intruders. In addition to guarding the nest, certain perciforms also aerate the eggs by directing a flow of water into the nest with fanning movements of the fins.

Although there is no evidence of parental care in perciforms that produce pelagic eggs, a strong protective behaviour is shown by most perciforms that build nests or carry their eggs around with them. The male dwarf cichlid may help in the care of the young, but it is the female that looks after the eggs, removing dead eggs from the clutch. In certain other cichlids (*Apistogramma species*, for example) the female may help free the young from the eggs by gently chewing off the egg shells. The young of many cichlids follow their mother around and quickly enter her mouth should danger threaten. The discus and the freshwater angelfishes of the cichlid family feed their young on mucous secretions of their own bodies. The male Betta guards the young until they can swim away freely on their own.

Behaviour. *Territorial activity.* Territorial behaviour is found in many perciforms, especially during the breeding season, when the male, and in some cases the female, displays territorial behaviour in guarding the nest of eggs or the young; such fishes include certain cichlids, sunfishes, and darters. The young tigerfish (Theraponidae) protects a restricted area around a small hole dug by using its body; such territorial behaviour disappears when the ti-

gerfish grows beyond a length of about nine centimetres (3.5 inches). An intruder approaching the burrow of a jawfish is usually greeted by a threatening pose of flared gill covers and erected fins. Gobies and blennies are also known for their marked territorial display; peck order may be present among gobies holding territories, with the highest degree of competition between male gobies of the same size. The characteristic threats of gobies and blennies include flaring gill covers, gaping jaws, puffing of throats, head raising, and shaking of bodies. When threat displays fail to settle a territorial dispute, male gobies fight, biting and chasing each other.

Threat displays of gobies and blennies

Sound production. The significance of sound production among perciform fishes is not well known, but most acoustic activity seems to be related to feeding and spawning periods. The level of sound production in croakers (Sciaenidae) increases considerably in the spawning season during the hours of late evening. There is also a difference in level between day and night; this may result from their feeding time. Damselfishes produce clicking sounds during feeding time, grinding the pharyngeal teeth. Another type of sound produced during feeding can be heard when parrotfishes feed on plant material covering reefs, biting on coral with their powerful platelike teeth. Grunts produce sounds by grinding their upper and lower pharyngeal teeth; the sounds are in turn amplified by the air (swim) bladder. Croakers, however, produce sounds by vibrating muscles of the abdomen that are attached to the sides of the air bladder, amplifying the vibrations of other muscles. The tigerfishes, or grunters (Theraponidae), have a similar system for sound production.

Drawing by G. Christman

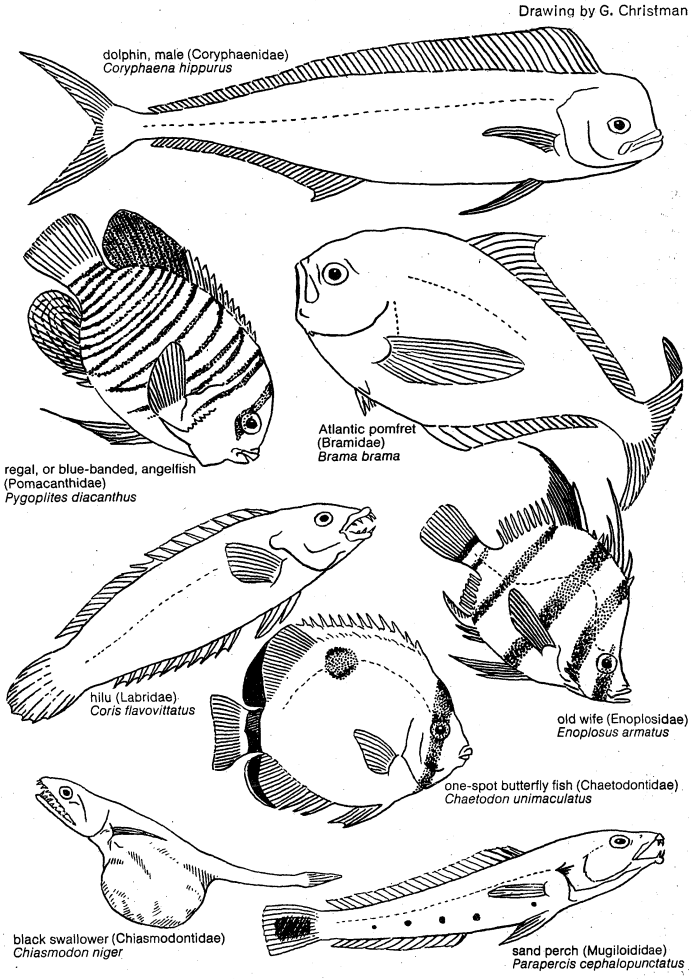


Figure 3: Representative perciforms of the families Coryphaenidae, Bramidae, Pomacanthidae, Enoplosidae, Chaetodontidae, Labridae, Mugiloididae, and Chiasmodontidae.

Feeding behaviour. Perciforms include both predator and prey species and are thus of great importance within the ecological food chains. The diverse adaptations for feeding are partly responsible for the success of this abundant order. Many of the colourful perciforms that occur around coral reefs are herbivorous fishes, the food of which consists mainly of plankton, algae on corals, and other reef vegetation; such fishes include parrotfishes, damselfishes, butterfly fishes, rabbitfishes and surgeonfishes. Among freshwater perciforms, certain species of *Tilapia* depend on aquatic plants for food. Most freshwater perciforms, however, are carnivorous, taking mosquitoes, insect larvae, and small insects. The larger predatory perciforms, in both freshwater and saltwater, feed on smaller fishes and even on birds and small mammals. They occupy a higher position within the food chain; examples include barracudas, groupers, tunas, and billfishes. The dolphins (*Coryphaena*) use their speed to catch fast prey such as flying fishes (Exocoetidae). The bluefish (*Pomatomus saltatrix*, Pomatomidae) is known for its voracious feeding behaviour; it feeds on open-water schooling fishes and, for unknown reasons, will continue to kill food fishes after its hunger is satiated.

Feeding behaviour of archer fishes

An interesting means of securing food is seen in archer fish (*Toxotes*, Toxotidae). The structure of the mouth in the archer fish is modified to form a groove along the roof of the mouth, against which the tongue fits to form a tube. The fish is able to direct a drop of water with remarkable accuracy at insects clinging to vegetation above the water surface. Thus bombarded, the insects fall into the water where they are quickly seized. Similar but less powerful squirting behaviour is also found in the butterfly fish. Another interesting type of feeding behaviour is seen in an African cichlid, which practices lepidophagy, the eating of scales plucked from other fishes.

Some predators lie in wait for their prey instead of pursuing it. An outgrowth of the mouth of the stargazer (*uranoscopus scaber*) acts as a lure for prey. Groupers are also known to lie in wait for prey among rocks.

Adaptations of the mouth and jaw structure are seen in many of the perciforms. The piscivorous nandids (Nandidae) and the leaf fishes (Polycentridae) have large protrusible mouths capable of taking prey two-thirds their size, and the deeply cleft mouth of the swallowers (Chiasmodontidae) permits them to pass prey larger than themselves into their highly distensible stomachs.

Interspecific relationships. Mutual relationships among species are found in many perciform fishes. The cleaner fishes of the wrasse genus *Labroides* (Labridae) are well-known for their role in the removal of parasites from larger carnivorous fishes. The larger fishes recognize the cleaner fish and will not devour it. They allow free passage into their cavernous mouths and gill chambers, in which the cleaner fish feeds upon leftovers and parasites. Each *Labroides* maintains a "cleaning station," which is visited regularly by larger fishes such as groupers, eels, jacks, and snappers. A relationship of a protective nature exists in the butterfishes (Stromateidae), the young of which are often found among the tentacles of jellyfishes; the fishes are immune to the stings of the jellyfishes. Fry of horse mackerel and tuna (Scombridae) have been also found among the tentacles of jellyfishes. A similar relationship is seen in the clown anemone fish (*Amphiprion percula*), which is found among the tentacles of sea anemones. The mucous substances secreted by the anemone fish protect it from the stinging cells of the sea anemone. Some anemone fishes seek out only one type of sea anemone; others do not show any species preference. The sleepers of the genus *Vireosa* (Eleotridae) are usually found close to rock oysters and clams, into which they quickly disappear when danger threatens. A similar relationship exists between certain sea cucumbers (sac-shaped echinoderms of the class Holothuroidea) and cucumber fishes (Carapidae). These fishes are found among starfishes, clams, and sea urchins, as well as sea cucumbers. Some are host specific and may even parasitize the host, as in the Florida cucumber fish (*Carapus bermudensis*), which seeks out a specific sea cucumber of the genus *Actinopyga*, within which the cucumber fish makes it home. At times *Cara-*

pus also feeds on the internal organs of the sea cucumber; this does not really harm the host because it regenerates the lost parts.

The blind goby, *Typhlogobius californiensis,* depends entirely upon holes dug by the ghost shrimp (*Callianassa*) for a home, and is unable to live without its help. Other gobies are known to share holes with burrowing worms, pea crabs, and snapping shrimps.

Imitative resemblance among perciforms

Certain perciform fishes depend upon imitative resemblance for survival. Immature tripletails (Lobotidae) will turn on their sides and float on the surface of the water, resembling dead leaves; similar behaviour is found in the leaf fish *Monocirrhus polycanthus* (Nandidae). Some wrasses (Labridae) resemble green algae because of their body coloration, a mixture of white, green, and brown. A remarkable mimic is seen in the case of the sabre-toothed blenny (*Aspidonotus taeniatus*), which mimics the cleaner fish *Labroides.* By resembling a cleaner fish, the blenny is able to approach other fishes and surprise them by rushing in to bite off a piece of fin (see MIMICRY). Similar mimicry also occurs in an East Indies species of blenny that mimics a wrasse, apparently for food and protection.

FORM AND FUNCTION

The nature and diversity of the perciforms make a general definition of the group difficult; the most common characters are found in the large families of sea basses, mackerels, perches, sunfishes and others. Perciform fishes usually have spines present on their dorsal, anal, and pelvic fins. The dorsal fin is usually divided into two parts, with the first part supported by one or more spiny rays; these are believed to have evolved for defense purposes. The pelvic fins are usually present, directly below or a little ahead of the pectoral fins, and they are supported by one spine and five or fewer soft rays. This position of the pelvic fins gives the perciforms an advantage in manoeuvring over short distances. The pelvic fins are lacking in some perciforms; in others, such as gobies, they are united to form a cuplike sucker; and, in the gouramis, the pelvic fin may be drawn out into long filaments.

Drawing by G. Christman

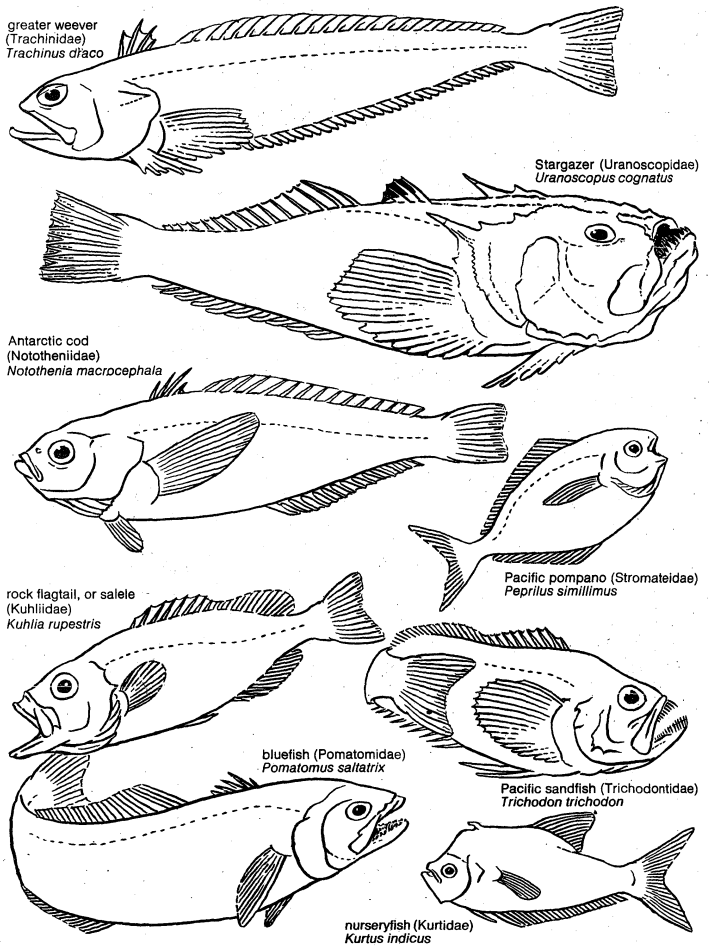


Figure 4: Representative perciforms of the families Kuhliidae, Pomatomidae, Trichodontidae, Trachinidae, Uranoscopidae, Notothenidae, Stromateidae, and Kurtidae.

A diversity of mouth and jaw structure occurs in the perciforms; most of it is brought about by the various types of feeding behaviour. Perciforms usually have protrusible jaws; and in the leaf fishes and swallowers, the jaws are easily distensible. The protrusible jaw may have thick lips, as in the wrasses, or may possess fleshy projections, as in certain species of African cichlids. Weever fishes (*Trachinus*) and stargazers (*Uranoscopus*) possess jaws that are directed upward; the jaws help when capturing prey as they lie buried in the sand. The upper jaws are greatly prolonged in the swordfishes and billfishes; the significance of this feature is rather uncertain. Many of the perciform species that inhabit coral reefs have modifications of the snout and jaws; the butterfly fishes have a straight tubelike mouth for reaching food among coral crevices.

Other structures of the perciforms have also undergone modification according to the various types of feeding behaviour. Most of the piscivores possess numerous short, fine, and pointed teeth; *e.g.,* the perches and sea basses. Barracudas have long piercing canine teeth for holding and stabbing prey, and certain gobies and blennies characteristically have long, curved canines in the lower jaw only. Perciforms that are either herbivorous or consumers of small invertebrates in addition to vegetation possess incisors, which are chisel-like teeth, as in certain sea breams; incisors may become fused into a beaklike structure, as in the parrotfishes. Enlarged pharyngeal (throat) teeth are present in some perciforms and are used for grinding and crushing food such as clams and snails. Tooth structure has undergone various modifications in many of the African cichlids. Herbivorous cichlids possess chisel-like teeth to feed on plants and algae; the piscivorous ones have strong, pointed teeth. Cichlids that feed on the eggs and young of other species possess a highly distensible mouth with reduced teeth embedded in the gums. (M.W.L.)

CLASSIFICATION

Annotated classification. The classification presented here is mainly that of British ichthyologist P.H. Greenwood and colleagues, published in 1966; it in turn is fundamentally that established by the British ichthyologist Carl C. Tate Regan in 1929. Certain changes have been incorporated, based mostly on characters from the nervous system.

All of the features distinguishing the order Perciformes are found in the fishes of the most generalized group, the suborder Percoidei, which contains the sea basses, sunfishes, perches, and fishes of many other families. As the subordinal name implies, the fishes composing it are "percoid," or perchlike in appearance. The fishes in the other suborders have presumably evolved from a percoid-like ancestor, but some have changed so much as to hardly resemble a percoid fish externally.

ORDER PERCIFORMES

Swim bladder not connected by an open duct to the throat; dorsal, anal, and pelvic fins usually with spines; dorsal fin usually with the first or anterior part supported by spiny rays and the rest by soft (articulated) rays, the spinous- and soft-rayed portions often separated from each other so as to constitute two distinct dorsal fins, or there may be a notch in the profile of the dorsal fin that indicates the two joined portions. A fin is considered long- or short-based on the basis of the length of its attachment to the body. Pelvic fins with 1 spine and 5 or fewer soft rays, or pelvic fins absent; pelvic fins thoracic in position (*i.e.,* placed ventrally below the base of the pectoral fins), pelvic fins sometimes ahead of pectoral fins (that is, jugular in position), or, occasionally, pelvic fins posterior to pectoral fins; pelvic girdle usually directly attached to base of pectoral girdle; caudal fin with not more than 17 principal (branched) rays supporting it; skull lacking orbitosphenoid bone; shoulder girdle lacking meso-

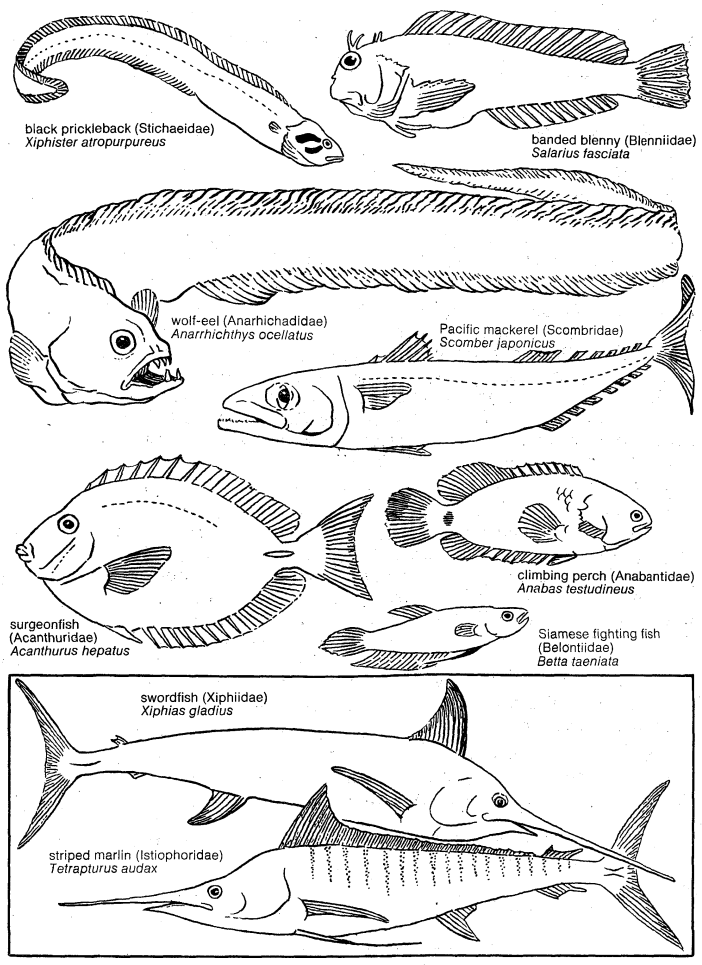


Figure 5: Representative perciforms of the families Blenniidae, Stichaeidae, Anarhichadidae, Acanthuridae, Scombridae, Xiphiidae, Istiophoridae, Anabantidae, and Belontiidae.
Drawing by G. Christman

coracoid bone; jaws typically protrusible; premaxillary bone of upper jaw excluding the maxilla from the gape of the mouth. Scales usually rough-edged (ctenoid), provided with small teeth (ctenii) along their posterior edge, sometimes round and smooth (cycloid). About 6,000 species, marine and freshwater; most species along shorelines in tropics and temperate zones, and in freshwater, the number of species dropping off drastically in higher latitudes. Fossil remains from the Upper Cretaceous Period (more than 75,000,000 years ago).

Suborder Mugiloidei

Spiny-rayed dorsal fin rather widely separated from soft-rayed dorsal; pelvic fins of 1 spine and 5 rays, not thoracic but located more posteriorly on the abdomen; scaled lateral line with nerve pattern resembling that of some lower nonperciform fishes; taste nerves of trunk like many percoids; olfactory bulbs far forward, unlike other perciforms.

Family Mugilidae (mullets). Lower Oligocene to present; cigar-shaped, roundish body, short snout; large cycloid scales; usually numerous small movable teeth; muscular gizzard-like stomach in many species. Moderately large schooling fishes, 30 to 90 cm (1 to 3 ft) long. Less than 100 species; tropical and temperate waters, some in brackish water and freshwater.

Suborder Sphyraenoidei

Two dorsal fins, both short-based and widely separated from each other; pelvic fins some distance posterior to pectorals; pattern of trunk lateral line nerves in a rudimentary percoid pattern resembling pre-percoid patterns found in atheriniform fishes; pattern of taste nerves on trunk resembles that in atheriniform fishes; large teeth set in deep sockets.

Family Sphyraenidae (barracudas). Eocene to present; large, elongated, pikelike, with long, pointed jaws and big teeth; piscivorous; probably not over 120 cm (4 ft) long; all warm seas; about 20 species; fine game fishes.

Suborder Polynemoidei

Pelvic fins thoracic; pectorals low on side of body, divided into an upper normal part and an unusual lower part, consisting of a number of fin rays grown out into long sensory filaments, reaching to anal fin and far beyond in some species.

Family Polynemidae (threadfins). Upper Miocene to present; resembling mullets in body shape and widely separated two dorsal fins but like anchovies (see CLUPEIFORMES) in ventral mouth with projecting snout, rather deeply cleft mouth, and adipose eyelids; in most warm seas, often abundant at river mouths and over sandy bottoms; about 24 species, most 30–69 cm (12–24 in.) long, but 1 giant species reaches 180 cm (6 ft).

Suborder Percoidei

The largest suborder in both numbers of families and of species; fishes typically of a perch or bass appearance; jaws protrusible; dorsal fin usually conspicuously spinous, often with the spinous and soft portions separated or nearly so or with a notch between the 2; anal fin with 2 or more spines at anterior end, occasionally with 1 spine, sometimes with more than 3 spines; pelvic with 1 spine and usually 5 soft rays; body often somewhat deep rather than elongated; territorial, bottom-oriented, investigative shorefishes with great close-quarters swimming manoeuvrability, swimming backward and forward short distances with much use of pectoral fins; great variety in adaptive design and operation of jaws; generalized predators of fishes and crustaceans in all warm seas near shores, especially in tropics; in freshwater. Almost 4,000 species, some of large size; about 90 families.

Superfamily Percoidea (basses, perches, sunfishes, cichlids, damselfishes, and many others)

About 60 families grouped together because they show no great morphological specialization away from the general bass, grouper, or perch kind of fish taken as a model. Most inhabit shores of tropical and temperate seas or lakes. Two aspects most obviously contributing to the success of this group are the adaptability of the protractile mouth and the detailed variety of specializations for swimming manoeuvrability in restricted areas.

Family Scombropidae. Pliocene to present; rare, deep water marine (down to 600–800 m, or 2,000–2,600 ft); this and the next several families retain some features that may have been those of the most generalized ancestors of present-day percoids such as: 2 dorsal fins separate, anal fin with 2 spines, primitive pattern of taste nerves on the body and remains of a special system of lateral line nerves on the head. About 7 genera and 20 species.

Family Acropomatidae. Rare, deepwater marine species rather similar to scombropids; anus located anteriorly from normal position at front of anal fin. Light organs present; midwater depths of 300–500 m (1,000–1,650 ft); 2 species; Indo-Pacific.

Family Apogonidae (cardinal fishes). Eocene to present; 2 well-separated dorsal fins; 2 spines in anal fins; midsection of body often short, with head, eyes, and caudal peduncle proportionately larger; mouth large. Males orally incubate eggs. Mostly marine species, often reddish; live around coral reefs in tropics and subtropics; nocturnal; a few species from deep oceanic midwaters 1,000–1,200 m (3,300–4,000 ft); most with special lateral line system of nerves and free organs well developed on head (among the few percoids with the system). About 200 species.

Family Moronidae (temperate basses). Eocene to present. Two dorsal fins connected at their bases. Most species slim-looking basses; well-known food and game fishes such as striped bass and white basses of the genus *Morone*. Some species anadromous. Weight to 50 kg (about 110 lb) in striped bass. About 12 species, marine and freshwaters of North America, Europe, Africa, Australia, and the Orient.

Families Pomadasyidae and Banjosidae (grunts). Eocene to present. Spinous and soft dorsal fins continuous, often notched. Resemble snappers (Lutjanidae), but teeth weaker, canines absent; sound produced by grinding pharyngeal teeth together and amplified by adjacent swim bladder. Second anal spine often much enlarged; about 75 species, tropical and subtropical shorefishes, many entering estuaries; good food and game fishes.

Family Centropomidae (snooks or robalos). Eocene to present. Elongated, basslike fishes; head long and sloping; sizable mouth, projecting lower jaw; two separate short-based dorsal fins; short-based anal fin. Oceans and estuaries of Pacific and Atlantic coasts of tropical Americas; about 8 species, 45–150 cm (about 1½ to 5 ft); good food and game species.

Family Dinolestiidae. One species, resembling, but not related to, the barracudas (Sphyraenidae, above). Marine; Australia and Tasmania; length to 50 cm (20 in.).

Family Latidae (Nile perches). Eocene to present. Closely similar to Centropomidae. Piscivorous; freshwaters of Africa; at river mouths and along coasts in Southeast Asia and northern Australia; up to about 180 cm (6 ft) long and 140 kg (300 lb); excellent food fishes.

Family Percichthyidae (perch trouts). Eocene to present. Dull-coloured, small, perchlike freshwater and marine fishes of Chile and Argentina. Dorsal fin deeply notched. About 9 species.

Family Ambassidae (glass perches). Small fishes similar in body form to Apogonidae; body short, rather deep. Spinous and soft-rayed parts of dorsal fin nearly separated by deep notch. Good predators on mosquitoes; large schools in sea around river mouths and in freshwater; Indo-Pacific; about 24 species.

Drawing by G. Christman

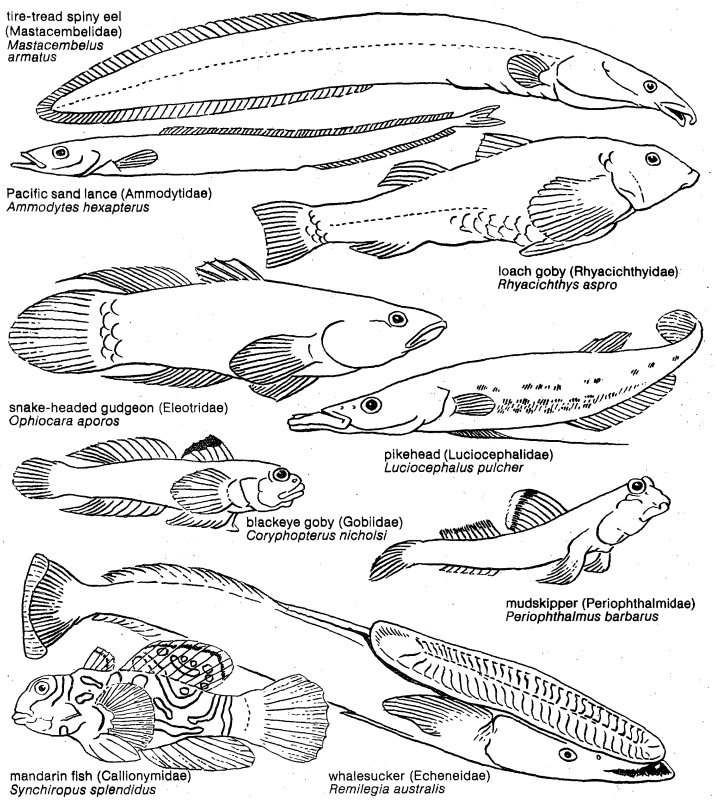


Figure 6: Representative perciforms of the families Luciocephalidae, Mastacembelidae, Echeneidae, Ammodytidae, Eleotridae, Gobiidae, Periophthalmidae, Rhyacichthyidae, and Callionymidae.

Family Serranidae (sea basses, groupers, and many others). Paleocene to present. Structurally diverse; variously similar to many of the percoid families mentioned above in their general spiny-rayed, perchlike appearance. Dorsal fin continuous, but may be deeply notched; spinous portion of dorsal fin with longer base than soft dorsal portion; anal usually with 3 spines and short based; no scaly sheath along base of dorsal and anal fins; mouth large; pectorals broadly rounded; caudal fin usually truncate or rounded, sometimes moderately forked. About 400 species in tropical, subtropical and warm temperate seas; some in freshwater; good food and game fishes; weight to about 320 kg (700 lb).

Families Pseudochromidae, Grammidae, Plesiopidae, Pseudoplesiopidae, and Acanthoclinidae. Quite similar, small, darkly colourful, rather secretive coral-reef basslike fishes of tropical Indo-Pacific and Caribbean seas. An interesting specialization of numerous species is the presence of multiple horizontal, interrupted lateral lines on trunk: one along the back, one along the side, and one along the bottom of each side of the body. Dorsal and anal fins vary from few or no spines up to 24; long dorsal fin, sometimes deeply notched between spines with a little banner or flag of fin membrane extending up and out from the end of the spine. Together, about 40 species.

Families Glaucosomidae (pearl perches), *Lobotidae* (triple tails), *and Banjosidae.* Deep-bodied perchlike fishes, all found in eastern Pacific Ocean, except *Lobotes*, which also occurs elsewhere in tropical salt and freshwater; 5 or 6 species in Glaucosomidae; 1 in each of the other families.

Family Priacanthidae (big eyes or catalufas). Eocene to present. Deep-bodied, reddish, serranid-like dwellers of deeper offshore waters, toward the bottom. Jaws almost vertically hinged; carnivorous. Tropical Indo-Pacific and Atlantic; few species; 30–45 cm (12–18 in.) long.

Family Centrarchidae (sunfishes and basses). Eocene to present. Moderately deep-bodied; spinous and soft dorsal fins continuous, not separate as in Percidae; more than 3 anal spines. Freshwaters of North America; only 1 species, *Archoplites interruptus*, native west of the Rocky Mountains; various species widely introduced elsewhere; prefer quieter waters, such as ponds, lakes, swamps; excellent game fishes; size from 30 gm (1 oz) to about 10 kg (22 lb); 30 species.

Family Embiotocidae (surfperches). Miocene to present. Laterally compressed, ovate, smooth-scaled, fairly small, with small head. One long dorsal fin, depressible into a scaled sheath alongside the fin. Give birth to actively swimming young; 23 species from central Baja California, to Japan, absent from Aleutian chain; most species (20) occur in California; one freshwater species (*Hysterocarpus traski*) in central California; 12–30 cm (5–12 in.).

Families Nandidae and Polycentridae (near leaf fishes and leaf fishes). Small, mostly piscivorous fishes with large to huge protrusible mouths; consume prey up to ⅔ their own length. Bodies moderate to deep, laterally compressed; long spinous dorsal fin and 3 to 13 spines in anal fin; soft dorsal and anal short-based and colourless, used together with colourless pectoral fin in *Polycentrus* and *Monocirrhus* in swimming imperceptibly toward prey without any evident signs of fin or body movement. Six species in freshwaters (and brackish for *Nandus*) from India to Malaysia; for Polycentridae, West Africa, Central America and east coast of South America. Six species.

Family Cichlidae (cichlids). Eocene to present. Small freshwater (a few brackish water) percoids, resembling damselfishes (below) and North American sunfishes. One pair of nostrils instead of the 2 of most fishes; spinous dorsal fin long-based; anal fin short-based, 3-spined dorsal and anal fins often pointed at posterior ends; caudal fin usually rounded; first pelvic rays of males often elongated; lateral line interrupted. Native to Texas, Central America, South America, and West Indies; Africa; Palestine, India, and Sri Lanka (formerly Ceylon). Complicated courtship, nest-building and mouth-brooding (several genera). Over 700 species, many important aquarium fishes; almost 600 native to 3 big rift lakes in Africa (especially Tanganyika, Malawi, and Victoria); a few species up to 30 cm (12 in.); 1 up to 9 kg (20 lb; genus *Tilapia*).

Family Pristolepidae. A deep-bodied, laterally compressed, small-mouthed percoid, resembling a cichlid or pomacentrid but having a patch of blunt, molariform teeth on base of skull opposite similar teeth on rear floor of mouth. Long spinous dorsal fin; oval 3-spined; interrupted lateral line. One species only, of unsolved percoid relationship but evidently not allied to Nandidae as once thought. Freshwaters; Burma to Indo-China and Malaysia.

Family Pomacentridae (damselfishes and anemone fishes). Eocene to present. Abundant, conspicuous, active little fishes, often brightly coloured, of shores and coral reefs, mainly in tropical, a few in subtropical seas. Resemble cichlids in general appearance and, like cichlids, also have only 1 pair of nostrils. Lateral line continuous or interrupted; long spinous dorsal fin; anal with 2 spines, sometimes 3; soft dorsal and anal fins similar; caudal fin usually forked; all fins may have pointed filamentous ends; head scaled; scales large, ctenoid; bases of unpaired fins scaled; first ray of pelvic fin somewhat elongated; floor of mouth with one triangular, fused tooth plate in pharynx (pharyngeal plate). Territorial, pugnacious; about 250 species.

Family Percidae (perches, walleyes, darters). Eocene to present. Spinous and soft dorsal fins usually well separated; anal with 1 or 2 spines and short-based; scales ctenoid; bodies rather elongated. All freshwater, temperate species; perches and pike perches Holarctic with a few brackish-water species and a marine species of pike perch in parts of Black and Caspian seas; darters are native only to North America. Perches prefer quiet waters, darters running waters; pike perches occur in either and are semi-migratory. Many species build nests and show parental care; size up to about 90 cm (3 ft) and 25 lb for walleyes; darters from 2.5 to 7.5 cm (1 to 3 in.). About 125 species, of which about 100 are darters.

Family Sciaenidae (drums or croakers). Upper Cretaceous to present. Some species resemble cods, others resemble sea basses; most have lower jaw short or underslung, with upper

jaw and snout extending beyond lower jaw; often 1 or more barbels (fleshy filaments) at tip of lower jaw. Spinous and soft dorsal fins separate; soft dorsal fairly long based; anal fin small, with 2 spines (most percoids have 3 or more); lateral line continues out to posterior end of caudal fin (unusual for most percoids); air bladder often with intricate outpocketings and with muscles attached to it that operate to make resonating sounds in air bladder, hence name of drum; surface of head may be cavernous through expansion of lateral line canal system. Most species occur on slopes of continental shelf, a few around islands; most in tropics, a few in temperate waters; a few freshwater. About 160 species; size from about 100 gm to 100 kg (a few ounces to 225 lb); many important food fishes.

Family Odacidae (rock whitings). Teeth incompletely fused together making a parrot-like beak as in parrot fishes (Scaridae), but odacids are evidently not related to wrasses (Labridae), parrotfishes, or whitings (Sillaginidae), all of which they also resemble. Seven species; southern Australian and New Zealand seas.

Family Labracoglossidae. Resemble grunts (Pomadasyidae) to which they are allied. Five species. Easter Island to New Zealand, Australia, Japan.

Family Sillaginidae (whitings). Oligocene to present. Elongated fishes with long, conical snout, small mouth; moderately long dorsal and anal fins; anal with 2 weak spines. About 6 species of small marine fishes of shallow water; Indo-Pacific, often in estuaries and river mouths; dig in bottom with long snouts for food.

Family Branchiostegidae (tilefishes). Pliocene to present. Body elongated; large, oblique mouth with strong canines; body deep through chest region; eyes high on head at top of steep sloping forehead; single, rather long dorsal fin; fin spines weak. Moderate to large body size; about 20 species, most in shallow seas of tropics and temperate zone.

Family Lactariidae (milk trevally). Miocene to present. Moderately deep-bodied, laterally compressed; mouth large, oblique; eyes large; pectorals pointed; two dorsal fins separated; anal fin long-based. One or two species marine in Indo-Pacific.

Families Owstoniidae and Cepolidae (bandfishes). Eocene to present. Owstoniids are marine, deep-water fishes, basslike, but large mouth is oblique, eye large, and dorsal and anal fins long, continuous, and high; caudal fin with long rays; body tapers noticeably. Cepolids similar, but have a long tapering body and are shallow- and deep-water fishes. Cepolids occur from Europe through Mediterranean to India, China, Japan, and Philippines; owstoniids only in Far East; few species.

Family Mullidae (goatfishes). Miocene to present. Resemble minnows (Cyprinidae); have a long pair of chin barbels that usually lie flat against chin, except when in use as sense organs, probing the bottom for food. Spinous dorsal fin well separated from soft dorsal. Fifty to 60 species at reefs and shallow sandy or muddy bottoms near shore in tropics and warmer temperate seas.

Family Lutjanidae (snappers). Miocene to present. Resemble sea basses (Serranidae), but when mouth is closed jaw slips under bony cover over preorbital area (between eye and jaw); enlarged canine teeth in jaws; spinous dorsal fin longer than soft and joined to it; anal fin short-based; caudal fin usually truncate. About 250 species; marine and brackish water in all warm seas.

Families Nemipteridae, Scolopsidae, Lethrinidae, Pentapodidae. Resemble Lutjanidae; some with wider preorbital area under which upper jaw slips; others (Nemipteridae) with molar teeth in sides of jaws and incisors or canine-like teeth at front end of jaws. About 50 species collectively; marine, Atlantic and Indo-Pacific, especially around coral reefs; some species up to 90 cm (3 ft) long; good eating.

Family Sparidae (breams, porgies). Eocene to present; resembling fishes of families Nemipteridae and Lutjanidae; mouth small but with a powerful dentition of incisors or canines across the front of the jaws and molar-like teeth to the sides, the teeth enlarging toward the rear of the mouth; used in crushing crustaceans, mollusks, and small fishes. About 100 species in shallow seas of tropical, subtropical, and temperate zones; most species less than 30 cm (1 ft) long, a few up to 120 cm (4 ft); important food and game fishes.

Family Gerreidae (mojarras). Small perchlike fishes with compressed, rather deep, silvery bodies; they are set off from most other percoids by their highly protrusible mouths used in probing soft, sandy bottoms to catch food organisms; about 40–50 species, mostly tropical marine, a few in freshwater; abundant in American tropics and East Indian region. Resemble fishes of family Leiognathidae, but the similarity is thought to be the result of parallel evolution.

Family Pempheridae (sweepers). Perchlike fishes with compressed body and a very short-based dorsal fin but long-based, low anal fin; big eyes. About 20 species in tropics of Atlantic, Indian, and Pacific oceans.

Family Bathyclupeidae. Resemble sweepers but apparently not related; compressed body; prominent lower jaw; single short-based dorsal fin; anal fin long-based; eye large; mouth large. Few species; deep-sea midwaters, 400–500 m depth.

Family Leptobramidae (salmon trout). A slender carangid-like species with large mouth, rather long-based anal fin, and with a single dorsal fin placed behind the beginning of anal fin; resembles Pempheridae but apparently is not related to it; a single species reaching 43 cm (17 in.) and about 2 kg (4 lb); off coasts of western Australia and New Guinea.

Family Carangidae (jacks, scads, and pompanos). Eocene to present. Lateral line usually strongly arched anteriorly and the posterior part usually armed with enlarged, keeled, scutelike bony scales; second dorsal and anal fins elevated and falcate anteriorly and sometimes followed by short-based free finlets; soft-rayed portions of dorsal and anal fins preceded by separated or nearly separated free, short spines; swift-looking fishes with thin, streamlined bodies, very constricted caudal peduncle, deeply forked caudal fin, and long, sickle-shaped pectoral fins; eyes often covered with clear eyelids; worldwide in tropics and warm temperate areas; over 200 species; most are good food and game fishes.

Family Rachycentridae (cobias). Long and slender, somewhat resembling mackerels and shark suckers (remoras); may be related to shark suckers (family Echeneidae); soft portion of dorsal fin preceded by a row of short, separate dorsal spines. One species, worldwide in warm seas but absent in eastern Pacific; size up to 90 cm (3 ft) and over 70 kg (150 lb); pelagic in habit.

Family Coryphaenidae (dolphins). Sleek, fast-swimming fishes of tropic and temperate open seas; speeds up to 65 km (40 mi) per hour; dorsal and anal fins long-based and low; dorsal fin begins on head; deeply forked caudal fin; long, slender pelvic fins fitting into groove on belly; young males of one species change into adults with almost vertical forehead. Two species, 1 large, up to 150 cm (5 ft) and 25 kg (55 lb).

Family Formionidae (black pomfrets). Body deep, strongly compressed; dorsal and anal fins long-based and elevated at anterior end, both fins with spines rudimentary in adult fish; a few enlarged scutes at end of lateral line, a characteristic of fishes of jack family (Carangidae). One species, up to about 60 cm (2 ft); Indian Ocean to Australia and China.

Family Menidae (moonfishes). Eocene to present. Body strongly compressed, triangular, very deep behind head, with edge of chest razor-sharp; anal fin very long and low, reaching forward to pelvic fins; first ray of pelvic fin long and filamentous; mouth protrusible into an upward extended tube; one species, less than 30 cm (12 in.) long, in Indo-Pacific seas.

Family Leiognathidae (slip mouths). Oligocene to present. Small fishes; body greatly compressed, ovate; mouth small but highly protrusible into a tube pointing up or down; long dorsal and anal fins elevated anteriorly, folding down into scaly sheath lying alongside bases of the fins; exude large amounts of mucus after capture. About 25 species, mostly marine shallow-water shorefishes, some in estuaries of Indo-Pacific area.

Family Bramidae (pomfrets). Body deep, strongly compressed; head broadly rounded in profile; pectorals long and curved; long-based dorsal and anal fins, low, sloping back to narrow caudal peduncle and deeply forked caudal fin; mouth small, oblique; about 23 species of darkly coloured fishes in rather deep waters of open ocean; size up to 90 cm (3 ft). One species, *Pteraclis velifera*, with enormously high and long fanlike dorsal and anal fins.

Family Caristiidae (manefishes). Rare black pomfret-like fish from midwater depth of 1,000 m (3,300 ft) over much deeper bottoms; dorsal fin begins far forward over end of cranium, high and like a mane; pelvic fins very long; 1 or 2 oceanic species.

Family Monodactylidae (fingerfishes). Includes family Psettidae. Small, silvery, ovate, small-mouthed fishes of salt and brackish water, temporarily in freshwater. Pelvic fins minute; dorsal and anal fins long; anterior spines of dorsal fin short and detached. Five species, coasts of Africa to India and Southeast Asia.

Family Toxotidae (archer fishes). Lower Tertiary to present. Moderately deep-bodied percoids distinguished by nearly

straight line from dorsal fin to tip of jaws; jaws large, oblique; lower jaw prominent; jaws and roof of mouth adapted for expelling under force drops of water fired from surface of water at insects on overhanging tropical vegetation in mangrove swamps. Dorsal and anal fins set far back on body; dorsal fin with 4 enlarged spines at front; 5 species, with black saddle markings or spots on dorsal surface; marine and freshwaters of Indo-Pacific.

Family Ephippidae (spadefishes and batfishes). Includes Chaetodipteridae, Drepanidae and Platacidae. Eocene to present. Body deep, orbicular, greatly compressed laterally; long-based dorsal and anal fins, usually high, especially anteriorly in young; mouth quite small. About 15 species in tropics of world, few in temperate zone; size up to 40 cm (16 in.).

Family Pomacanthidae (angelfishes). Eocene to present. Body deep, laterally compressed; mouth quite small; strongly resemble Pomacentridae and Chaetodontidae (in which family they traditionally have been placed) but distinguished from these and other deep-bodied coral-reef fishes by a rather long, sharp spine projecting posteriorly from the lower part of cheek region (from the preopercle bone); profile of head from dorsal fin to tip of snout mainly convex, rarely straight, never concave as in butterfly fishes; young often differently coloured from adults; less than 100 species of conspicuous coral-reef and tropical fishes; mostly of small size, a few species up to about 45 cm (18 in.).

Family Scatophagidae (scats). Eocene to present; deep-bodied, laterally compressed; small mouth; 2 dorsal fins joined at base; bases of dorsal and anal fins scaled; about 4 species; coasts of Africa to Indo-Pacific, entering brackish and freshwater; scavengers, eating decaying plant and animal remains and fecal matter, including human; size up to 40 cm (16 in.).

Family Rhinoprenidae. One species, recently discovered, superficially similar to Scatophagidae, but really quite different. Long free filament at front of dorsal and anal fins and 1 on pectoral fin; posterior nostril of each side of head enlarged and confluent across snout. One species over muddy bottoms at river mouths in Gulf of Papua, New Guinea.

Family Enoplosidae (old wife). Eocene to present. Body laterally compressed; spinous and soft dorsal fins elevated anteriorly, as is anal fin; general appearance gives impression in side view of two separate bodies joined together at midpoint; pelvic fins large. One species in rocky areas of Australian coast; size up to 22 cm (9 in.).

Family Chaetodontidae (butterfly fishes). Oligocene to present. Body deep, disk-shaped; mouth small, jaws sometimes on end of small (or, occasionally, fairly long) beak; slope of forehead from dorsal fin to tip of snout is concave, never convex; spinous portion of dorsal and anal fins usually prominent; bases of dorsal and anal fins scaled. Body often marked with oblique dark bands and one or two eye-like spots. Over 100 species in tropics of world, all marine; typical of coral reefs.

Family Pentacerotidae (pelagic armorheads). Resemble Chaetodontidae but with higher dorsal fin and much larger dorsal, anal, and pelvic spines; dorsal fin usually strongly elevated; head rough with exposed bony plates; snout usually elongated bearing small mouth at its end; lips with "hairy" skin. Seven or 8 species, in deeper coastal waters from South Africa, Australia, and New Zealand to Japan; size up to about 50 cm (20 in.).

Superfamily Kuhlioidea

Classified together under this suborder are a number of percoid families that have in common the same pattern of main branches of a complex of taste nerves that occur on the body.

Family Kuhliidae (flagtails and aholeaholes). Typical percoids resembling North American freshwater sunfishes and basses; dorsal fin single, notch between soft and spinous portions; dorsal and anal fins folding into scaly sheaths along their bases; 6 species of tropical Indo-Pacific oceans, preferring brackish water and freshwater; size up to 45 cm (18 in.); good eating.

Families Scorpididae (sweeps and halfmoons), *Kyphosidae* (sea chubs), *Girellidae* (opaleyes), *Coracinidae* (galjoen fishes). All quite similar families recognized by combination of ovate body, small mouth, strong caudal fin that is usually weakly forked; and, especially, a spinous dorsal fin with low spines followed by a higher evenly curved or falcate soft dorsal fin; about 36 species, many partly herbivorous; tropical and warm temperate coasts in Atlantic, Pacific, Indian oceans; size usually not over 45 cm; some are good food and game fishes.

Family Oplegnathidae. Pliocene to present. Strongly resemble Scorpidae and Kyphosidae, but incisiform teeth of young become fused in adult to form a parrot-like beak to upper and lower jaws; these fishes are not related to true parrotfishes (Scaridae); several species of shorefishes mostly in tropics of Southern Hemisphere; size up to 60 cm (24 in.).

Family Theraponidae (grunters or tigerfishes). Typical percoids of small bass type; colours dull or silvery with horizontal dark stripes; dorsal fin deeply notched, spinous portion longer than soft portion; some species make grunting sounds. Less than 25 species, Indian and western Pacific oceans in freshwaters of Australia and New Guinea; small to medium size.

Family Arripidae (Australian salmon; ruffs). Not related to true salmons of Northern Hemisphere. Rather long, slender body; deeply forked tail; moderately long dorsal fin, a notch between the shorter spinous dorsal and longer soft dorsal fin. Two species; marine, young in brackish water; shallow waters off South Australia, New Zealand and adjacent islands; size up to 1 m; important food and game fishes.

Family Emmelichthyidae (bonnetmouths). Includes families Caesionidae, Erythricthyidae, Dipterygonotidae, Maenidae, Spicaridae, Centracanthidae, Merolepidae by some authors. About 25 species of 2 general body types: one with slender, elongated bodies with moderately protrusible upper jaws; the other deeper bodied and with enormously protrusible upper jaws. Some school in open waters at depths of 3–45 m (10-150 ft), others in deeper water.

Family Pomatomidae (bluefish). Resembles Australian salmon (family Arripidae), but spinous dorsal smaller and separate from soft dorsal; anal fin longer; body more robust; 1 or 2 small spines before anal fin. One species, widely distributed in tropical and subtropical seas; voracious feeder on other fishes; size up to 120 cm and 11 kg (4 ft and 25 lb); good game and food fish.

Family Nematistiidae (roosterfish or papagallo). Streamlined fish resembling jacks (Carangidae); dorsal fin remarkable for spinous portion consisting of greatly lengthened spines nearly separate from each other, the fin connected by fin membrane only at their bases, except the last ray, which is free at its base. Caudal fin deeply forked; pectoral fin long, falcate; mouth rather large. One species found only on west coast of Central America, from Gulf of California south to Panama; frequents sandy shores; size up to 35 kg (80 lb); great game fish.

Superfamily Labroidea

Family Labridae (wrasses). Paleocene to present. Various body forms, but commonly cigar-shaped, fairly slender; snout moderately long, sometimes lengthened; dorsal and anal fins long-based and low; caudal fin truncate; especially characteristic are the noticeable lips and outcurving sizable canine teeth at ends of upper and lower jaws, the large scales and the habit of swimming around the coral reefs by "rowing" with the pectoral fins. About 300 species, many beautifully coloured and marked, often with differences of colour and pattern between sexes; also differences between young and adults; size from 5 cm to nearly 2 m (2 in. to several ft), 10 kg (20 lb); marine shallow-water fishes of tropics and warm temperate zone.

Family Scaridae (parrot fishes). Eocene to present; resemble Labridae but with stouter bodies and with teeth fused to form a parrot-like beak to upper and lower jaws; large scales in regular rows; herbivorous; about 80 species, often brilliantly coloured; species separable by colour and pattern; sexes often differ in colour. Size up to 120 cm (4 ft) and 20 kg (45 lb); tropical marine fishes.

Superfamily Gadopsoidea

Family Gadopsidae (river cod). Slender, somewhat elongated, with longer based dorsal and anal fins than typical percoid. Characterized by very slender and fairly long pelvic fins located in front of pectoral fins; caudal fin rounded. One species, rivers of southern half of Australia; somewhat resembles Ophidiidae (order Paracanthopterygii) but apparently is a derivative of family Serranidae; size up to 60 cm (2 ft).

Superfamily Cirrhitoidea

Family Cirrhitidae (hawkfishes). Small colourful perchlike fishes having lower rays of pectoral fins unbranched, thickened, and separate from one another; small flag of skin projects from tip of each spine of dorsal fin; about 35 species; shallow coastal waters in warm seas.

Families Chironemidae, Haplodactylidae, Cheilodactylidae, and Latridae. Similar to Cirrhitidae; 25 to 30 marine species, mostly in shallow waters of Australia and New Zealand, some species off South America; size up to 60 cm (2 ft).

Superfamily Champsodontoidea

Family Champsodontidae. Small, elongated spiny-rayed fishes with a small spinous first dorsal fin and rather long, soft

dorsal and anal fins; pelvic fins rather large; eyes near top of head and close together; unusual is the large mouth, the jaw extending obliquely past the eyes; several species; carnivorous; deep waters of Indo-Pacific oceans.

Superfamily Trichodontoidea

Family Trichodontidae (sandfishes). Resemble codfishes, but eyes high on side of head; mouth large, oblique; lips fringed; pectoral fins with long base extending forward past pelvic fins. Two species; marine; North Pacific; to 25 cm.

Superfamily Trachinoidea

Family Opisthognathidae (jawfishes). Resemble Clinidae, but jaws large to huge, extending far past eye; dorsal fin long-based; spinous and soft portions continuous; anal fin long-based; body usually elongated, slender; eye almost at anterior tip of head; pelvic fins below pectorals. About 24 species, mostly small, in shallow tropical and temperate seas.

Family Bathymasteridae (ronquils). Resemble Opisthognathidae, but jaws not so large; no spines in dorsal or anal fins; pelvic fins slightly ahead of pectorals; about 8 species; bottom dwelling; coasts of North Pacific Ocean.

Family Mugiloididae. Includes Parapercidae and Pinguipedidae. Some resemble labrids in long dorsal and anal fins (sometimes with few spines), enlarged lips that appear to curl back, and enlarged canines at front of jaws. Body elongated, cylindrical; usually spotted and banded; eyes near top of head. Size from small up to 60–90 cm (2–3 ft); about 30 species; marine; bottom dwellers, coasts of South America, South Africa, Indo-Pacific to Japan; a few good food species.

Family Percophiidae. Includes Bembropidae and Hemerocoetidae, resemble flatheads (Platycephalidae; see SCORPAENIFORMES); body long, slender; head flattened; eyes on top of head, close together; separate spinous and soft dorsal fins; dorsal and anal fins long-based; jaws large. About 12 species; marine, from shallow to about 200 m; South America, Indo-Pacific to Japan.

Family Cheimarrichthyidae (torrent fish). Small, resembling a cottid or sculpin (family Cottidae; see SCORPAENIFORMS); eyes on top of head and close together; one species; freshwater streams of New Zealand; young in brackish water.

Family Trachinidae (weever fishes). Eocene to present. Body elongated, compressed, deep at head end, tapering to narrow, small caudal fin; a separate short-based spinous dorsal fin with black membrane and poison glands along grooves in each spine; anal and soft dorsal fin long-based; a long spine on gill cover with poison glands along grooves of spine; eye black in upper half, white in lower half; scales small, set in distinct oblique rows; 4 species, marine shallow water down to 90 m (300 ft); lie buried in sand with eyes and top of head showing; northern Europe, Mediterranean, Pacific coast of South America; venom very painful, even dangerous.

Family Trichonotidae (hairfins). Resemble Percophiidae and Mugiloididae, but body extremely elongated and dorsal fin unusually high; snout pointed; lips fringed; dive head first into sand. Several species; tropical and subtropical Indo-Pacific oceans.

Family Creediidae. Elongated little fishes resembling Percophiidae; 2 species known; coasts of Australia, Marshall and Mariana islands.

Family Limnichthyidae. Resemble Percophiidae; 5 species; New Zealand, Australia, Society, Marshall, Marianas, and Hawaiian Islands; sand burrowers.

Family Oxudercidae. A single species from coast of China at Macao; relationships still in doubt.

Superfamily Uranoscopoidea

Eocene to present. Three families; resembling toadfishes (Batrachoididae) and weever fishes (Trachinidae). Bodies rather elongated, with long dorsal and anal fins usually with few or no spinous rays; head broad, deep, flattish on top; eyes on top of head and somewhat erectile; mouth broad, oblique; lips fringed with toothlike dermal projections that interdigitate and strain water through sand when the fishes are buried to top of head and eyes; similar fringe of skin at upper end of gill slit; pectoral fins enlarged and fanlike; pelvic fins anterior to pectorals.

Family Leptoscopidae. Sand-burrowing fishes; no spines in dorsal and anal fins. Three species; marine; coasts of Australia and New Zealand; size up to 30 cm (12 in.).

Family Dactyloscopidae (sand stargazers). General features as described under superfamily; body elongated. Shape of pelvic fins is characteristic: each has 3 thickened segmented rays whose tips are free from the fin membrane, divergent and somewhat curled. Small fishes, mostly marine, up to about 10 cm (4 in.); burrow in sand with eyes and mouth protruding from surface of sandy bottom; 20 to 25 species in tropical America in Atlantic and Pacific oceans.

Family Uranoscopidae (stargazers). Head very broad and deep; posterior half of body tapering to a small truncate tail fin; eyes on top of head, projecting above surface of head; large posteriorly pointing spine on shoulder girdle above pectoral fin (in some species); dorsal fin moderately long and either with a short, nearly separate spinous portion or spines absent; anal fin moderately long and with few or no spines; large electric organs behind eyes of species of genus *Astroscopus;* about 25 species living buried on sandy bottoms usually in shallow shore areas; the only marine teleosts with electric organs; a few species at greater depths, to about 600 m (2,000 ft), Atlantic, Pacific, and Indian oceans; size up to about 9 kg (20 lb).

Superfamily Chiasmodontoidea

Family Chiasmodontidae (swallowers). Slender fishes with extremely deeply cleft mouth; large, backward-pointing teeth; dorsal fin long with spinous and soft dorsals separate; pelvic fins thoracic. Capable of swallowing and holding in their greatly distensible bellies fishes larger than themselves. About 12 species in open oceanic waters down to 500 m (1,600 ft); size up to 15 cm (6 in.); relationships in doubt.

Superfamily Notiotheniodea

Four families of codlike or sculpin-like percoids found mainly in the Antarctic, some in cold temperate Southern Hemisphere seas near Chile, Argentina, and New Zealand; superfamily includes about 75 percent of all Antarctic fishes.

Family Bovichthyidae. About 15 species in Subantarctic and south temperate seas, off Chile, Argentina, southern New Zealand, and Southern Australia; 1 species in rivers of South Australia and Tasmania.

Family Notothenidae (Antarctic cods). Miocene to present; about 50 species, most in Subantarctic waters; some species near Antarctic continent; a few in cold temperate zone, one species in rivers of southern South America. Mainly bottom dwellers of littoral zone, some deep-water species resemble true cods (Cottidae) and have a barbel on lower jaw; 2 species, large, up to 150 cm (5 ft).

Family Bathydraconidae (Antarctic dragon fishes). About 15 species; true Antarctic fishes, occurring on coasts of Antarctic continent; body greatly elongated; usually a spatulate, pikelike snout; no first dorsal fin; live on coasts of Antarctic continent to depths of 500–700 m; a few down to 2,500 m. Size up to about 50 cm, but usually much smaller.

Family Chaenichthyidae (white-blooded or icefishes). Famous white-blooded fishes of Antarctic; lack red blood cells and hemoglobin. Mostly large, up to 60 cm (2 ft) long, with scaleless body and two or three lateral lines each side; head large, snout long, spatulate, pike-like; teeth large; jaws nonprotractile. About 17 species, all in Antarctic except one in Subantarctic; mainly bottom dwellers feeding on crustaceans and small fish; most at 100 to 200 m (330–660 ft), some to 700 m (2,300 ft).

Suborder Stromateoidei

Five percoid-like families with an unusual and characteristic feature, a toothed saccular outgrowth in the gullet directly behind last gill arch. One family, the Amarsipidae, lacks the toothed saccular outgrowth in the gullet. Stromateoids may be related to fishes of superfamily Kuhlioidea.

Family Stromateidae (butterfishes). *Centrolopidae* (medusafishes), *Nomeidae* (flotsam fishes), *Ariommidae*, *Tetragonuridae* (squaretails). Eocene to present; slender to ovate, deep-bodied fishes; dorsal fin continuous or spinous portion set off from soft portion by deep notch; in the most generalized species, which resemble Kyphosidae, the soft dorsal is preceded by about 6 low, stoutish spines; other species resemble Carangidae. Eyes often with adipose (fatty) tissue around them; pelvic fins absent in some species, especially in ovate species; skeleton often weakly calcified; scaly sheath along bases of dorsal and anal fins. Young often found under jellyfishes or flotsam; adults live in deeper layers of ocean over continental shelves or are pelagic; a few species close to shore, in bays, even entering estuaries; 78 recognized species, many rarely seen; some feed largely on jellyfishes; others on crustaceans, tunicates, and small fishes; many are important commercially, especially in the Orient; family includes good-tasting species, such as butterfishes and pompanos. Tropics to temperate zone; never near oceanic islands; size 10–120 cm.

Suborder Icosteoidei

Family Icosteidae (ragfish). A single species of rare deep-sea fish of North Pacific Ocean; body highly flexible in water,

limp as a rag out of water; little is known of its anatomy. Resembles Stromateidae; presumably derived from a percoid-like ancestor; no spines in fins; pelvics and scales present in young, both absent in adult; body elongated, much compressed; up to 220 cm (7 ft).

Suborder Blennioidei

Fairly large suborder composed of 2 groups, the tropical blennies, some of which are percoid-like, and the northern blennies. The northern blennies are eel-like fishes, but the appearance of their faces and the aspects of their fins generally resemble those in the tropical blennies.

Superfamily Blennioidea (tropical blennies)

All soft, articulated fin rays in the dorsal, anal, and pelvic fins unbranched; pelvic fins, which are seldom reduced, of 1 spine and 4 or fewer rays and located in front of the pectoral fins; there are the usual 2 nostrils on each side of the head. Exact 1-to-1 ratio between the vertebrae and the dorsal rays and posterior anal rays; only 1 lateral line. About 620 species.

Family Clinidae (clinids). Eocene to present. Percoid-like fishes, some moderately elongated, rather flat-sided, usually with somewhat pointed snouts and fleshy lips; dorsal and anal fins rather high and long-based, with fin membranes conspicuously supported by thin, riblike fin rays; caudal fin fanlike, not large; pelvic fins ahead of pectorals, slender; 1 spine and usually 2 or 3 rays; pelvics in some species appear usable in walking movements; cirri (bushy tentacles of skin) often present above eyes, on anterior nostrils, and just behind head on each side. Body scaled; scales cycloid; first 3 rays of dorsal fin often higher and more or less separate from rest of dorsal. A few species "four eyed" (*i.e.*, with eye divided so as to see out of water). Most inhabit tide pools, kelp, rock crevices; some species down to 30 m (100 ft); size up to 30 cm (12 in.), most are smaller; about 180 species.

Family Tripterygiidae (threefin blennies). Pliocene to present. Much like clinids but dorsal fin divided into 3 distinct parts, the first 2 of spines only; small bottom fishes of reef and rocks. About 100 species mostly in warm seas.

Family Blenniidae (combtooth blennies). Eocene to present. Resemble clinids in fins and body shape but differ in being scaleless, in having a steep forehead and only a single row of teeth in both jaws, the teeth being close-set, long, comblike. Sometimes a pair of large to enormous curved "canines" further back in jaws; some species hop about out of water in intertidal zone; fins without pungent spines; eyes large and at top of forehead; pelvic fins with 2 rays. About 300 species, in tropical, subtropical, and warm temperate seas; of small size.

Family Chaenopsidae (pike blennies). Pliocene to present. Body very elongated; jaws long; long gill area; dorsal and anal fins long, confluent with caudal fin; no scales or lateral line; 6 species in tropical and subtropical marine shore areas of Central America and Caribbean; small fishes living in worm tubes and burrows.

Family Pholidichthyidae (engineer fish). Very elongated eel-like fish; reclusive, living under excavations; move sand and gravel in mouths; known previously only by young stage; adult recently discovered by L. Dempster. One species; marine; in tropics, Indonesia and Philippines; size up to 30 cm (12 in.); poorly known; relationships in doubt.

Family Notograptidae. Eel-shaped; dorsal and anal fins long-based and high, both confluent with caudal fin; pelvics 1-rayed, filamentous, placed before pectorals; body scaled, mouth large. Two small species; marine; western Australia.

Superfamily Stichaeoidea (northern blennies)

Eel-like fishes; single pair of nostrils; dorsal and anal fins long-based and often joined to caudal fin; pelvic fins usually present, placed a little before pectoral fins, of 1 spine and usually fewer than 4 rays; bottom fishes usually of littoral zone, some supralittoral; a few in deep water (down to 450 m, or 1,500 ft).

Family Stichaeidae (pricklebacks). Includes families Chirolophidae, Lumpenidae, Xiphiodontidae, Cebedichthydae. Eel-like; body usually scaled; dorsal fin with spines only or some soft rays at rear of fin; pectorals reduced; pelvics present or absent; some species with 3 or 4 lateral lines but most species lacking lateral line. About 54 species, most in North Pacific, some in Atlantic; often among seaweeds when tide is out; also in deeper water (to 200 m); most feed on invertebrates, a few on seaweed; most species small, a few up to 60 cm.

Family Pholidae (gunnels). Elongated; strongly compressed laterally; scaled; pelvics reduced to 1 spine and 1 ray, very small or absent; spines only in dorsal fin. About 8 species, most in North Pacific; also in North Atlantic; size small; intertidal zone or shallow water; feed on small bottom invertebrates.

Family Anarhichadidae (wolffishes). Big head and long tapering body, laterally compressed; massive dentition on jaws, roof of mouth, and throat for crushing mollusks, sea urchins, crustaceans; length to 2.3 m (7.5 ft). Nine species, in northern oceans; littoral zone of shore to 300 m (1,000 ft); good eating.

Family Ptilichthyidae (quillfish). Extremely elongated, body ending in a free, fleshy point; pelvic fins absent; dorsal and anal fins like vanes of a feather. One species, rare; North Pacific.

Family Congrogadidae. Includes Haliophidae. Resembles Stichaeidae but mouth and lips large; fins almost or entirely spineless; 3 lateral lines. About 8 species; rock-dwelling, in shallow coastal waters of Indo-Pacific from Africa to Japan.

Family Zaproridae (prowfish). A single species like a shorter, deeper bodied prickleback; pelvic fins absent; size up to 2.8 m (9 ft); deeper coastal waters to 350 m, California to Alaska.

Family Scytalinidae (graveldiver). Eel-like, with dorsal and anal fins soft-rayed and not beginning until middle of long straight body; body appears to flare out somewhat at these fins; pelvic fins lacking. One species; marine, California to Alaska; small, to 15 cm (6 in.); burrows quickly in sand or gravel in intertidal zone.

Suborder Acanthuroidei:

Modified perciod-like fishes characterized by peculiarities of bones suspending the jaws, which thereby are extended far forward as small nibbling mouths at end of more or less lengthened snout.

Family Acanthuridae (surgeonfishes). Includes Zanclidae. Percoid-like fishes characterized by one lancet-like, darkly coloured spine (like a surgeon's scalpel) or tubercles alongside the caudal peduncle (in front of base caudal fin); body deep, compressed; dorsal and anal fins long-based; scales small; eyes high on side of head; mouth small, low, sometimes extended into a beak; teeth close-set and lobate. About 75 species; many herbivorous; inhabit shallow tropical shores, a few in deeper water.

Family Siganidae (rabbitfishes). Resembling surgeonfishes but uniquely characterized by pelvic fins each having two stout spines, along the outer and inner edges, respectively; both spines grooved for carrying venom; no lancet-like spine or buckler on caudal peduncle; mouth small, rabbitlike, used for nibbling algae. About 18 marine species of small to moderate size; tropical Indo-Pacific, around coral reefs and rocky areas.

Suborder Scombroidei

Streamlined, mackerel-like or marlin-like fishes the interrelationships of which are in doubt; upper jaw not protrusible; maxillary bones of upper jaw more or less firmly attached to non-protractile premaxillaries that lie ahead of them.

Family Gemphylidae (snake mackerels). Eocene to present. Elongated, laterally compressed; mouth large, with large, cutting teeth; spinous part of dorsal fin longer than soft-rayed part, the latter often broken up into finlets posteriorly; pelvic fins usually not rudimentary. Sixteen species; tropical and temperate seas; down to 600 m (2,000 ft); length to 1 m (40 in.); some commercial value; good eating.

Family Trichiuridae (cutlass fishes). Paleocene to present. Elongated, bandlike body; large mouth with large fangs anteriorly. Scaleless; dorsal fin long, spinous anteriorly; anal fin long-based, spines short; caudal fin deeply forked on most species, absent in some (body ending in tapered point); pelvic fins reduced to scalelike spine and one small ray, or absent. Fifteen species in warm oceans; rather deep water offshore, to about 1,000 m (3,300 ft); piscivorous.

Family Scombridae (tunas and mackerels). Moderate to large, streamlined, swift-swimming, schooling fishes; body often thickly rounded, tapering to a narrow caudal peduncle bearing in some species 2 or 3 keels on its side; caudal fin widely forked or lunate (scimitar-shaped); distinguished from all fishes by series of separate finlets following spinous first dorsal and spinous anal fins; well-developed vascular system under skin of tunas is associated with sustained high-speed swimming and an elevated body temperature up to 9° C (5° F) above the surrounding water; make extensive migrations to spawning and feeding grounds; fins fit into grooves or depressions on body; about 40 species, open waters of tropics and warm seas of world; feed on fish, crustaceans, squids, and other abundant animal life; length 30 cm to 3.3 m, weight to 800 kg. Many species subjected to ever-increasing fishing exploitation and will be threatened with extinction.

Family Xiphiidae (swordfishes). Bones of upper jaw prolonged into a swordlike structure, flatter and sharp-edged, as compared with the round, shorter bill of marlins and billfish-

es; dorsal fin high and short-based, not extending beyond middle of body; caudal fin with high lobes; second dorsal fin small; pelvic fins absent; no teeth in jaws except in very young; no scales. Probably only 1 species, *Xiphias gladius*, worldwide in tropics and temperate seas; from surface to 400 m; size up to 3.6 m and 450 kg. Sword is lashed about inside schools of fishes; injured fishes are eaten. A major big-game fish; excellent eating and commercially important.

Family Istiophoridae (billfishes, marlins, sailfishes, and spearfishes). Bill round and shorter compared with sword of swordfish; dorsal fin long, extending almost the length of back of body and reaching striking height in the sailfish, *Istiophorus gladius*; pelvic fins present as thin filaments; body scaled; 2 small keels on each side of caudal peduncle; fine teeth on jaws. About 7 species; worldwide in warm seas; all are large species, the black marlin (*Makaira indica*), the largest, probably attains 900 kg (2,000 lb); record for rod and reel is 710 kg (1,560 lb); greatest game fishes in the ocean, and excellent food fishes, of considerable economic importance.

Family Luvaridae (louvars). One species; rare; resembles a dolphin (family Coryphaenidae) in its very high forehead and eye placed low almost on level with mouth; mouth small, toothless; body deep, laterally compressed; a fleshy keel on each side of caudal peduncle; pelagic in tropics and subtropics; length to 1.5 m (5 ft).

Suborder Anabantoidei

Small percoid-like fishes characterized by an accessory air-breathing chamber of labyrinthic structure on each side of head, above regular gill chamber; an unusual patch of teeth on roof of mouth, teeth which attach to floor of cranium; usually build foam nests. Freshwater fishes of Southeast Asia, India, and Africa; about 70 species in 5 families, 4 of which are closely related; the fifth, Badidae, is more generalized.

Family Badidae. Resembling Anabantidae externally but lacking accessory air-breathing chamber. Have special patch of teeth on roof of mouth and characteristic courtship behaviour of anabantids; 1 species, freshwaters of India and Southeast Asia; small, 5–8 cm (2–3 in.) long.

Family Anabantidae. Pleistocene to present; about 20 species, freshwaters of tropical Africa, Southeast Asia and Philippines; includes the "climbing perch"; size small.

Family Belontiidae (fighting fishes, some gouramies, and others). About 50 species of small freshwater fishes from tropical Africa, India, Burma, Sri Lanka, Southeast Asia, Malaya; includes Siamese fighting fish (*Betta splendens*), various species of gouramis, which have 1 ray of each pelvic fin extended into an elongated filament; paradise fishes and other popular aquarium fishes. Most of these species were formerly placed in family Anabantidae.

Family Helostamatidae (kissing gourami). Freshwater, Southeast Asia; popular aquarium fish; one species.

Family Osphronemidae (gourami). Pelvic fins each with 1 ray drawn out into a long filament; 1 species; freshwater in Sumatra, Java, Borneo, grows to about 9 kg (20 lb).

Suborder Luciocephaloidei

Family Luciocephalidae (pikehead). Pike-like in body form and long head; an accessory suprabranchial air-breathing chamber (but not labyrinthic); bony plate in throat region; peculiar jaws. One species, freshwater; Malay Archipelago and Peninsula.

Suborder Mastacembeloidei

Family Mastacembelidae (spiny eels). Eel-like; head elongated with fleshy, mobile proboscis projecting beyond lower jaw; soft dorsal fin preceded by series of erectile spines. About 50 species, freshwaters from tropical Africa to Southeast Asia up to China; up to 90 cm (3 ft), but most much smaller; some attractively coloured aquarium fishes.

Family Chaudhuriidae. One species in freshwaters of Burma; similar to spiny eels but small and without fleshy proboscis.

Suborder Kurtoidei

Family Kurtidae (nurseryfishes). Peculiar, small, percoid-like; males carry eggs, stuck under an anteriorly pointing hornlike process on top of back of head. Two species; brackish water and lower parts of streams; Indo-Malaysia and New Guinea.

Suborder Echeneioidei

Family Echeneidae (remoras). Oligocene to present; differ from percoids mainly in having a sucking disk on top of head, modified from the spinous first dorsal fin. About 10 species; warm marine seas; ride attached to large fish or other marine animals or ships; carnivorous; some are cleaner fish for other fishes; length 30 to 90 cm (1 to 3 ft).

Suborder Ammodytoidei

Families Ammodytidae and Hypoptychidae (sand lances). Eocene to present. Long slender percoid-like fishes with long, pointed head and projecting lower jaw; dorsal fin long-based, with soft rays only; pelvic fins thoracic or (usually) absent. About 18 species; most seas of the world, especially colder waters; sand burrowers; large schools near shores form an important food source for many other fishes; 20–40 cm (8–16 in.).

Suborder Gobioidei

Almost all with pelvic fins located beneath pectorals and joined together to form a vacuum cup or suction disk; some with pelvic fins close together but not in form of a suction cup; a few lack pelvics; all lack one particular bone of cranial roof. More than 800 species, most of small size, living as bottom dwellers, worldwide, in saltwater, freshwater, or brackish water, especially in tropics.

Family Eleotridae (sleepers). Pelvic fins close together or in contact anteriorly, but not united into a sucking cup; short-based spinous first dorsal fin and longer based soft-rayed second dorsal; all fins usually with rather long fin rays; no lateral line. Numerous species; along coasts and rivers in tropics and subtropics; most 10–30 cm (4–12 in.) long, one to 90 cm.

Family Gobiidae (gobies). Eocene to present. Pelvic fins united to form a suction cup; no lateral line; simple-looking but often colourful fishes with 2 dorsal fins; all fins rather conspicuously large compared with small, usually slender body; definite rows of naked lateral-line organs on head. More than 700 species; size from 1.2 cm (0.5 in., world's smallest vertebrates) to something under 30 cm (12 in.), with most less than 10 cm. Shallow coastal waters of tropics and temperate zones; usually resting on bottom or hidden; carnivorous.

Family Periophthalmidae (mud skippers). Elongated gobies with blunt head; eyes erectile, prominent, on top of head; pectoral fins with muscular base used to walk over mud and climb mangrove roots; about 6 species in estuaries and mud flats of Indian and Pacific oceans, to Japan.

Family Rhyacichthyidae (loach goby). Pelvic fins widely separated; head flattish, pointed; mouth ventral; 1 species living in torrential mountain streams of Indonesian Archipelago, up to Philippines; size up to 33 cm (13 in.).

Family Gobioidae. Elongated, eel-like gobies; dorsal fin very long, spinous part distinct from, but connected to, soft part; dorsal and anal fins confluent with caudal; eyes tiny to indistinct; mouth often obliquely upward; about 20 species, marine and brackish water of tropical America, Indian and Pacific oceans; about 8–30 cm (3–12 in.).

Family Trypauchenidae. Eel-like; scaled; dorsal and anal fins very long and confluent with caudal; pelvic fins united at base but not forming a cup or disk; eyes tiny. A blind cavity above gill cover opening to exterior by a pit. Burrowers in mud and gravel; about 12 species, along coasts and estuaries from Africa to Japan and Oceania.

Family Kraemeriidae. Rare little elongated fishes; pelvic fins separate; chin of lower jaw large, pointed, forming terminal end of head; eyes small, on top of head; 2 species, Indo-Pacific oceans.

Family Microdesmidae. Rare, small, eel-like; chin large, forming pointed end of snout; about 6 species; both coasts of tropical America, West Africa, tropical Pacific.

Suborder Callionymoidei

Family Callionymidae (dragonets). Eocene to present; resembling flatheads (Platycephalidae; see SCORPAENIFORMES); body flattened, broad, muscular; snout very short; mouth small; gill openings restricted, on top of head; eyes on top of head, close together; pelvics in front of pectorals; large spine pointing posteriorly at angle of preopercle on gill cover; bottom dwellers from tidepools down to 600 m (2,000 ft); burrow into sand; push along bottom using pelvic fins; about 40 species in tropical and temperate zones.

Family Draconettidae. Look like callionymids but are separated on differences in head skeleton; no preopercular spine; a few species; North Atlantic and North Pacific in deep water.

Critical appraisal. Classification of perciform fishes will be receiving much more study in the future. Expected changes include removal of some groups from the order, addition of some from other orders, and considerable realignment of many groups within the Perciformes. Evidence suggests removal of Gobioidei to a preperciform position, possibly as a distinct order. The snakeheads (order Channiformes) should likely be returned to the order Perciformes, associated with the anabantoid and luciocephaloid fishes. Particularly difficult problems

are the relationships and classifications of trachinoid, uranoscopoid, notothenioid, and stichaeoid fishes. Studies are presently being made on possible interrelationships between clusters within the 60 or more families of percoid fishes (suborder Percoidei). The placement and relationships of the atherinoid (sometimes called percesocine) fishes are still in dispute. The present classification follows that of Greenwood *et al.* in the removal of Atherinidae from perciform fishes (see ATHERINIFORMES) and in the removal of the ophidioid and zoarcid fishes (see PARACANTHOPTERYGII). Support for retaining the Atherinidae in order Perciformes and for other differences in the order Perciformes has been given by an American ichthyologist, W.A. Gosline, in a series of papers published between 1962 and 1971. Questions of relationships between carangid, rachycentrid, echeneid, and scombroid fishes are still unsettled. Many large and exciting problems thus remain. (W.C.F.)

BIBLIOGRAPHY. Recent books containing many excellent illustrations of perciform fishes and short accounts of the biology of each group are: JAMES E. BOEHLKE and CHARLES C.G. CHAPLIN, *Fishes of the Bahamas and Adjacent Tropical Waters* (1968), an excellent work in which each species is well illustrated, many beautifully in colour, with keys for identifying species and excellent summaries of the biology of each family; DAVID STARR JORDAN and BARTON WARREN EVERMANN, *The Fishes of North and Middle America*, 4 vol. (1896–1900, reprinted 1963); TOM C. MARSHALL, *Fishes of the Great Barrier Reef and Coastal Waters of Queensland* (1964), contains many large colour and black-and-white pictures of perciform fishes; IAN S.R. MUNRO, *The Fishes of New Guinea* (1967), 1,095 fishes illustrated by photographs, with 76 species in colour; JAMES L.B. SMITH, *The Sea Fishes of Southern Africa*, 4th ed. (1961), profusely illustrated with many colour plates; JAMES L.B. and MARGARET M. SMITH, *The Fishes of Seychelles* (1963), 880 species illustrated, many in colour; ALWYNE C. WHEELER, *The Fishes of the British Isles and North-West Europe* (1969); JOHN E. RANDALL, *Caribbean Reef Fishes* (1968), contains many fine photographs of perciforms and others.

Other books with information about perciform fishes are EARL S. HERALD, *Living Fishes of the World* (1961); NORMAN B. MARSHALL, *The Life of Fishes* (1966); and U. OKADA, *Fishes of Japan* (1955).

References pertaining to classification and relationships include: W.C. FREIHOFER, "Patterns of the Ramus Lateralis Accessorius and Their Systematic Significance in Teleostean Fishes," *Stanford Ichthyol. Bull.*, 8:79–189 (1963); and "Trunk Lateral Line Nerves, Hyoid Arch Gill Rakers, and Olfactory Bulb Location in Atheriniform, Mugilid, and Percoid Fishes," *Occ. Pap. Calif. Acad. Sci.*, no. 95 (1972); W.A. GOSLINE, "Systematic Position and Relationships of the Percesocine Fishes," *Pacif. Sci.*, 16:207–217 (1962); "The Suborders of Perciform Fishes," *Proc. U.S. Natn. Mus.*, 124: 1–78 (1968); and *Functional Morphology and the Classification of Teleostean Fishes* (1971); P.H. GREENWOOD *et al.*, "Phyletic Studies of Teleostean Fishes, with a Provisional Classification of Living Forms," *Bull. Am. Mus. Nat. Hist.*, 131:339–455 (1966); D.E. MCALLISTER, "The Evolution of Branchiostegals, and Associated Opercular, Gular, and Hyoid Bones, and the Classification of Teleostome Fishes, Living and Fossil," *Bull. Natn. Mus. Can.*, no. 221 (1968); C.T. REGAN, a series of papers on perciform classification in *Annals and Magazine of Natural History*, series 7, vol. 11 and series 9, vol. 11 (1903–23); and "Fishes," *Encyclopaedia Britannica*, 14th ed., vol. 9, pp. 305–328 (1929), which still forms the major basis of present perciform classification; D.E. ROSEN, "The Relationships and Taxonomic Position of the Halfbeaks, Killifishes, Silversides, and Their Relatives," *Bull. Am Mus. Nat. Hist.*, 127:217–267 (1964), and with COLIN PATTERSON, "The Structure and Relationships of the Paracanthopterygian Fishes," *Bull. Am. Mus. Nat. Hist.*, 141:357–474 (1969).

(W.C.F./M.W.L.)

Percussion Instruments

Two distinct groups of musical instruments, namely, idiophones and membranophones, comprise the large classification known as percussion instruments, which after the string and wind sections forms the third section of the Western orchestra. Idiophones are instruments whose own substance vibrates to produce sound (as opposed to the strings of a guitar or the air column of a flute); examples include bells, clappers, and rattles. Membranophones emit sound by the vibration of a stretched membrane; the prime examples are drums. Although many idiophones and some membranophones are tunable and hence may be melody instruments, both groups serve typically to delineate or emphasize rhythm. The term percussion instruments goes back to the German composer and organologist Michael Praetorius, who in 1619 wrote of *percussa, klopfende Instrument* (German *klopfen*, "to beat") as any struck instrument, including struck chordophones (stringed instruments). The same combination, including pre-bow chordophones, constituted the *divisio rhythmica* in the 7th-century *Etymologiae* of Isidore, archbishop of Seville.

This article covers the principal types of idiophones and membranophones found in the major areas of the world and attempts to list, within the considerations imposed by space, some of their diverse functions. (For more detailed treatment of percussion instruments in the Western orchestra, see ORCHESTRATION AND INSTRUMENTATION.) It is divided into the following sections:

Idiophones
  In Europe
    Antiquity
    The Middle Ages
    Modern times
  In Asia
  In Africa and Afro-America
    Muslim Africa
    Black Africa
    Afro-America
  In America
  In the Pacific
Membranophones
  In Europe
    Antiquity
    The Middle Ages
    Modern times
  In Asia
  In Africa and Afro-America
  In America
  In the Pacific

Idiophones

The following classification of idiophones provides some idea of the diversity of this large group of disparate instruments. Concussion instruments, consisting of two similar components struck together, include clappers, concussion stones, castanets, and cymbals. Percussion idiophones, which are instruments struck by a nonsonorous striker, form a large subgroup including triangles and simple percussion sticks; percussion beams, such as the *semanterion;* percussion disks and plaques, single and in sets; xylophones, lithophones (sonorous stones), and metallophones (sets of tuned metal bars); percussion tubes, such as stamping tubes, slit drums, and tubular chimes; and percussion vessels varying from struck gourds and pots to gongs, kettle gongs, steel drums, bells, and musical cups.

Types of idiophones

Shaken idiophones, or rattles, include vessels filled with rattling material, such as pellet bells, gourd, basketry, and hollow-ring rattles; strung rattles, such as dancers' leg rattles or anklets; stick rattles, including the sistrum, originally a forked stick with crossbars on which rattling shells, etc., were strung; pendant rattles with suspended rattling objects; and sliding rattles.

Other categories include scraped idiophones, comprising scrapers and cog rattles; split idiophones made of split hollow cane, including the Southeast Asian "tuning fork" idiophones and the chopstick; plucked idiophones, or linguaphones (so named because of their plucked tongues), such as the Jew's harp, sansa, and music box; friction idiophones, including friction sticks, simple or combined, and musical glasses; and blown idiophones, such as the 19th-century *Äolsklavier* and *piano chanteur.*

IN EUROPE

Europe received most of its idiophones either directly or indirectly from the high cultures of the ancient East or from Egypt, a country regarded by the Greeks and Romans as forming part of Asia rather than of Africa.

Antiquity. European antiquity knew many idiophones. Dancers' clappers, held pairwise in the hands of maenads

(women participants in Dionysian rites) and other female dancers, often stressed the rhythm of accompanying *auloi* (the ancient Greek reed pipes). The time-beating foot clappers of chorus leaders, attached to the right foot like a sandal, were known in Greece as *kroupalon* and adopted by Rome as the *scabellum.* Other idiophones included bells, cymbals, the unidentified *ēcheion,* and an instrument simply called "the bronze" (*chalkos*), probably a metal percussion disk. When the cult of Isis spread to Greece and Rome, her sistrum followed, always in the hands of priest or—rarely—priestess.

Use of bronze

With the exception of clappers, all these instruments were of bronze; as such they were credited with the apotropaic powers (the special protective powers against evil) accorded this metal in the East. Not only among primitive Asian islanders but in Greece itself was it customary to "sound the *chalkos* at eclipses of the moon," for example, "because it has power to purify and to drive off pollutions" (a scholiast, writing on Theocritus). According to the Roman poet Ovid, the annual visit of ghosts of the dead to their former homes was terminated by requests to depart emphasized by the clanging of a bronze plate. Thin, bronze percussion disks have been excavated at Volci, Etruria, and at Pompeii, affixed to metal handles; one from Pompeii is even garnished with pellet bells. Bells of bronze do not emit the clear, ringing tone associated with modern ones of bell metal; those of antiquity were small and made a clanging rather than ringing sound. They warded off evil spirits, averted the evil eye, served as sentinels' and watchmen's signal instruments, or were attached to the handle of a Greek warrior's shield in order to terrify the enemy by their clamour. Anklets adorned with small bells were frequently worn by jesters, dancers, and courtesans, particularly in Hellenistic times. In Rome, *tintinnabulae* ("bells") served as signal instruments or were suspended from the necks of herd animals—again to ward off evil.

Cymbals were sounded in religious rites, also at secular dances; deeply cupped, they may be seen on Greek and Roman art objects played together with a frame drum. Forked cymbals known as *crotala* travelled from Egypt to Greece and Rome, and finger cymbals were introduced from the East, chiefly for dancers, a pair being attached to thumb and middle finger of each hand.

Among the oldest instruments of mankind, rattles originally combined the functions of prophylactic amulets and children's toys; both functions coexisted as late as Roman times when *crepundia* ("rattles") served both as charms against evil and as children's toys. A number of existing Roman pellet bells (often misnamed jingles) are notable for their exquisite workmanship.

The Middle Ages. Greek and Roman idiophones were passed on to postclassical Europe, their distribution undoubtedly aided by wandering *joculatores,* or minstrels. Whereas prior to the adoption of Christianity most were ritual instruments, their function in medieval times—with the notable exception of the bell—was strictly secular.

Clappers, castanets, cymbals, and gongs. Clappers called *tabulae,* made of bone, are mentioned by Amalarius of Metz (9th century AD) as being in the hands of church cantors during liturgical services; possibly they were similar to the flat clappers played by jongleurs in the late Middle Ages and by lepers who sounded them as a warning of their approach. A three-piece clapper was used for this purpose in Spain, the *tablillas de San Lázaro* (*lazarus* being a Late Latin synonym of leper).

Lepers' warning clappers

Dancers' castanets, hollow clappers in bivalve form, were played in Spain throughout the Middle Ages; they are illustrated from the 11th century on. Already in Roman times, dancers of Cádiz are known to have played metal castanets, but those of sonorous hardwoods have been preferred since. Forked cymbals are portrayed on ivory and in miniatures of the Carolingian period. Small, cup-shaped cymbals called *acetabulae,* made of brass or silver, are mentioned by Cassiodorus (died *c.* 580) and Isidore of Seville. Interestingly, the modern practice of striking a single cymbal with a stick was anticipated by Isidore and by Suidas (10th century). Nonetheless, larger cymbals continued to be played pairwise and—at least in the later Middle Ages—were found among the "loud" instruments that accompanied dancing and were played at seigneurial festivities. Gongs appear intermittently from the 15th century on, first in the soup-plate form associated in Asia with gong chimes.

Bells. Bells were still thought to possess apotropaic powers; with their adoption, the Christian Church took over the belief that ghosts and demons could be put to flight by the sound of metal, a power henceforth augmented by the protection afforded by association with the divine cult and, more specifically, by baptism. The rite of baptizing bells is first recorded in a capitulary (civil ordinance) of Charlemagne of 789. Exorcizing celebrants sounded handbells or wore bells on their garments. St. John Chrysostom (died 407) felt compelled to protest against the custom of attaching bells to the clothing or bracelets of children in order to preserve them against demons, yet small bells or pellet bells continued to adorn the vestments of priests, a practice inherited from the ancient Near East: 51 bells ornamented the cope of Lanfranc, 11th-century archbishop of Canterbury. The Roman Catholic Church still officially recognizes the power of bells to divert the spirits of the dead; the tolling of passing bells in England, already recorded by Bede (died 735), was intended to ward off evil spirits from dying persons. Church bells announced the time of day, summoned the faithful to worship, sounded alarms, fought off lightning, broadcast the news of peace or the birth of a princeling, and praised God. In secular life handbells were played in mixed instrumental ensembles, usually one bell in each of the player's hands.

Symbolic importance of bells

Tuned bells are strung together to form chimes. Early chimes are generally represented as clapperless, hemispheric bells—inverted resting bells, in fact—struck on the outside by a hammer; less frequently, they are shown as clapper bells sounded by pulling on long cords attached to the clapper. From the 13th century, chimes were connected to clockwork and struck mechanically. As bells grew larger and the compass of the chime was extended, it became known as a carillon. The distinction today is merely one of compass, any set exceeding 1½ or two octaves in compass being considered a carillon.

Other instruments. In secular life, pellet bells were suspended from ladies' belts or worn by jesters. Before the introduction of bells in the Greek Orthodox Church in the 11th century, the *semanterion,* a percussion beam sounded by a striker, summoned the congregation, and in the Roman Catholic Church it has been used as a substitute for bells during Easter Week. Triangles first appear in the 14th century; originally a number of loose, rattling rings were threaded on the lower portion, as though the ancient sistrum had projected itself through the centuries. Scraped idiophones, known in Europe since Paleolithic times, are encountered as scraped pots intermittently from the 12th century on; such noisemakers were played—to judge by a 14th-century miniature—with other percussive instruments for merrymaking. Few instruments are as suitable for personal music making as the Jew's harp; coming from the East, it passed through Rome, but there is no evidence of its presence in the West before the mid-14th century, although it is unlikely to have disappeared completely during the intervening centuries.

Noise-making and musical improvisation

Most of the above idiophones are nontonal; with the exception of bells and percussion beams, they form part of the *musica irregularis* decried by writers such as the German organologist Sebastian Virdung in 1511 and as such were restricted to popular entertainment or signalling. Written music of this period does not help to determine whether, or how, idiophones were played with voices or with other instruments.

Modern times. *Xylophones.* Additional idiophones came into use from the Renaissance on. The xylophone was introduced from Indonesia; curiously enough, its first appearance was as the name of an organ stop in 1506, when it was termed a new stop. Virdung and the blind organist Arnold Schlick were the first to write of it, both in 1511. It remained little exploited until the Flemish carillonneurs combined it with a keyboard and transformed it into a practice instrument in the first half of the

17th century. The older form remained a folk instrument, chiefly in and east of Germany. An Austrian violinist, Ignaz Schweigl (died 1803), wrote concert pieces for it during his last years, and Saint-Saëns revived it in his *Danse macabre* of 1874. Lately, Afro-American marimbas have been incorporated into Western rhythm bands. Lithophones, stone imitations of xylophones, enjoyed a certain vogue in the 19th century, of which short-lived curiosities, the English stonemason Richardson's "rock harmonicon," with its tuned basalt bars, is best known.

Gongs. Gongs were always considered alien instruments, as indeed they are: although the word gong was known in the 16th century, its use is not further recorded until the late 18th century. Since then, gongs of indefinite pitch have been included in orchestral scores for arresting effect. For generations both flat and bossed (knobbed) gongs also performed the lowly task of summoning the family to meals.

Cymbals and castanets. Cymbals were apparently forgotten during the Renaissance; they reappear in the German composer Nicolaus Adam Strungk's opera *Esther* (1680) to provide local colour and are reintroduced through "Turkish music" (imitating the Turkish Janissary bands) to opera from 1770 on; soon afterward they acquired a permanent position in the orchestra. As folk instruments they have been played for centuries in Provence in France, probably as a continuation of an older ritual tradition.

In 16th-century Spain the saraband was usually danced to castanets or to a guitar; castanets have remained dancers' instruments ever since, either clicked together rhythmically or in sustained rolls. Modern rhythm bands now include one or two single castanets or a pair attached to a long handle for ease in clicking (the traditional manner of playing in the hand is difficult to master).

Bells. Bells grew larger until the largest ever produced, the Tsar Kolokol (Emperor Bell; 1734) of Moscow, weighing 433,200 pounds, proved too cumbersome and heavy for hanging. The hemispherical form was abandoned early as chimes became larger, culminating in tower-borne carillons brought into existence by progress in casting methods and mechanization. Chime bells were connected to town clocks and then hung in separate bell towers. In carillons of the Low Countries and northern France the bells' external hammers were automatically released by iron pegs set in a large wooden barrel (later, a metal cylinder) revolved by weight and pulley, permitting whole tunes to be played. Chorales and other melodies preludized the announcements of time on the European continent, while in Britain "chimes," or short sequences, fulfilled the office of timekeeper. Furthermore, instead of automatically playing melodies, British tower bells are rung in "changes"—a series of mathematical permutations—on bells hung dead. The role of small bells has become negligible, although handbell ringing is still a hobby in England.

Metallophones. Metallophones reached northern Europe from Indonesia in the second half of the 17th century and, like the xylophones, were promptly adopted by carillonneurs. In both the Low Countries and the regions to which such instruments spread from there, steel was the metal employed for bars. Provided with a keyboard, one served as "carillon" in Handel's oratorio *Saul*; another, struck with a beater, served as *strumento d'acciaio* (literally, "steel instrument") in Mozart's *Die Zauberflöte* (*The Magic Flute*). The modern glockenspiel (from the German *Glockenspiel*, "bell chime") was originally a bell chime, as its name indicates. Its transformation into a metallophone was in response to the need for a portable instrument; for marching bands the bars were hung in a lyre-shaped frame and called bell lyre. Symphony and opera orchestras—and today rhythm bands as well—make use of a larger, horizontal glockenspiel to which resonators have been added since 1935. Graduated steel bars struck from a keyboard by piano hammers form the celesta, patented in 1886 by Auguste and Alphonse Mustel of Paris; this has been scored in both symphony and opera and is a favourite of light orchestras. Its descendant, the vibraharp (called vibraphone in England), has characteristic tubular metal resonators that sustain the tone while imparting vibrato. These last two instruments are successfully exploited for their special tone colour in 20th-century music. Late 18th-century experiments with graduated sets of tuning forks struck by hammers from a keyboard produced the 19th-century "tuning-fork pianos," the most noteworthy of which were different models of Victor Mustel of Paris (from 1865 on).

The celesta

Tubular bells. Tubular bells are European adaptations of Southeast Asian bamboo chimes. Bells have always been both expensive and cumbersome, and substitutes started appearing in the 19th century, first in the form of steel bars, then hollow bronze cylinders, and finally brass tubes struck with wooden mallets. For use in churches, tubular bells can be played directly from the manual of an organ and the tone picked up, amplified, and fed into a system of loudspeakers in the tower.

Rattles and scrapers. Pellet bells and other rattles today play a very minor role, the former chiefly as horse or mule bells. In the present century the gourd-shaped maraca rattle has been added to the complement of rhythm band idiophones. Scrapers have all but disappeared from Europe: Virdung illustrated the scraped pot, now defunct, but other forms survive, mostly as carnival noisemakers. More sophisticated were the giant cog rattles with diameters up to 6½ feet installed in or outside church towers in the Iberian Peninsula, substitutes since the 15th century for bells during Easter Week. Scraped gourds are now part of the professional percussionists's equipment.

Plucked instruments. Jew's harps were part of the regular stock-in-trade of instrument dealers in the 16th and 17th centuries, and in the mid-18th century the playing of multiple Jew's harps is mentioned. Several of these little instruments combined in a single frame were played by virtuosos in the late 18th and 19th centuries and enjoyed enormous popularity. Miniaturization of musical clocks in the latter part of the 18th century resulted in the creation of the music box, a linguaphone provided with a metal-comb mechanism made from about 1770 on, chiefly in Switzerland. In its heyday—1810 to 1910—it was an immensely popular household instrument with a repertory of opera arias, folk songs, popular tunes of the day, and waltzes (after the midcentury). In the late 19th century it was transformed into a free-reed wind instrument (aerophone) by the substitution of free reeds for the metal comb, but neither comb nor free reed could resist the impact of the phonograph, which rendered both forms obsolete.

Friction instruments. During the 18th century several friction idiophones were introduced, among them the nail violin of Johann Wilde (*c.* 1740) with its tuned nails bowed by a violin bow. More characteristic of the period were the friction-bar instruments arising as a result of the German acoustician Ernst Chladni's late 18th-century experiments, particularly those concerned with the transmission of vibrations by friction. Chladni's own instrument, the euphone of 1790, and the aiuton of Charles Claggett of about the same time were first in a series of models, some with piano keyboard and horizontal friction cylinder or cone acting on upright bars and others with bars stroked by the player's fingers or bowed by a continuous bow.

Experimental instruments

Musical glasses are considerably older: the tuned metal cups or bowls of the East were transformed in Europe into tuned glasses and are first seen in the *Musica theoretica* (1492) of the Italian musical theorist, Franchino Gafori. One hears of them intermittently thereafter until they come to the fore in the mid-18th century as concert instruments. The rims of glasses of graduated sizes containing enough water to tune them were rubbed by the player's moistened fingers. By the 1760s they had attracted the attention of the American scientist and philosopher Benjamin Franklin, who proceeded to convert them into a more efficient and, above all, a polyphonic (many-voiced) instrument, which he called armonica—now known as the glass harmonica. Its popularity was immediate. Mozart's *Adagio und Rondo* K. 617 was written for it, as was his *Adagio für Harmonika* K. 356, both per-

formed in 1791. Efforts to combine it with a keyboard were successful from 1784 on but enjoyed only a passing vogue. Among the last to write for it was the French composer Hector Berlioz in his *Fantasia on The Tempest* of 1830; a decade later it was replaced by the growing family of free reeds.

Changing use of various types. The foregoing reveals the trend in Western musical culture to make diminishing use of nontonal idiophones. Those that succeeded in gaining access to the orchestra are used sparingly, often only as special effects (gongs) or indications of local colour (castanets). This trend is reversed in popular music of the 20th century, in which idiophones, both traditional and alien, play a basic part. On the other hand, tonal idiophones have assumed increasing importance. Those inherited from the Middle Ages, such as xylophones and Jew's harps, were expanded to the limit of their potential; others were imported and modified to meet local requirements. The new idiophones that came into being—for example, the music box and the friction-bar idiophones—proved ephemeral. Influenced by the great popularity of the piano, keyboards were adapted to numerous 19th-century idiophones in an attempt to render hitherto monophonic (single-voiced) instruments polyphonic.

IN ASIA

Clappers and cymbals. Still preserved today are actual clappers from Sumer of about 3000 BC and ancient Egyptian clappers, some of which took the form of a pair of hands. Their use seems never to have died out in Egypt. Copts still ritually strike clappers of a type that has probably been in existence since the 3rd or even 4th millennium BC. Modern Vietnam also retains clappers—often artifacts of great beauty—in its religious observances. On the Indian subcontinent circular wooden clappers, one held in each hand, are played by beggars and fakirs in some regions and are used as rhythm instruments in others. In the Far East they have a role in the theatre, as signal instruments, at funerals, or, as in the Ryukyu Islands, accompanying traditional dances. Castanets probably originated in ancient Phoenicia; they appear in Egypt in the 24th dynasty (*c.* 730–709 BC) and are typical dancers' instruments.

Ritual, social, and musical uses of clappers

Cymbals are indigenous to Asia; ancient Assyria had a unique form, funnel-shaped with long necks serving as handles. Known in ancient Israel from *c.* 1100 BC on, cymbals were the only permanent idiophones of the Temple orchestra. Egypt did not have true metal cymbals until the 24th dynasty. Today they remain in ritual use in northern India, Japan, Tibet, and Vietnam. They appear in the 5th century AD in India, where they are now also played at secular festivities. In China they play a prominent part in the theatre. Turkey, in contrast, has traditionally connected cymbals with military usage. In some part of Southeast Asia their ancient metal-connected function of dispelling evil spirits still prevails.

Xylophones, lithophones, and metallophones. Xylophones reach their highest development in Southeast Asia. They vary in form from the most primitive log type to the highly developed orchestral instruments found in Indonesia. Trough xylophones were depicted in Java in the 14th century but are not restricted to Southeast Asia; those of Japan, for instance, are rhythm rather than melody instruments, and in Burma they are associated with royalty. Xylophones of the Indonesian gamelan, or percussion orchestra (and some mainland ensembles), have various complementary compasses.

Stone chimes (lithophones) of two types occur: oblong bars like xylophone keys resting horizontally, found in Vietnam only, and vertically suspended plaques. In China their generic name is *ch'ing*; there, single sculpted musical stones and also 16-stone chimes are suspended from ornate frames. Stones forming a chime are carved in a typical L-form and struck with a mallet on their larger portion. The L-shape is very ancient; a chime of this form dating from the late Shang era (*c.* 1401–*c.* 1123 BC) was excavated at An-yang, in northeast China. Both in China and Korea, where the oldest chime goes back to the 14th century, the lithophone is a Confucian ritual instrument. Single stones and chimes are still in use in rural Annam (Vietnam).

Use of chimes in Confucian ritual

Metallophones occur in the same two forms as lithophones, those in the horizontal position being the more common. Indonesia and Indochina have metallophones constructed like xylophones, of which they are indeed metal counterparts. But in China the *fang-hsiang* with its 16 bars is a metal imitation of the lithophone. Among important components of the gamelan is the *saron*, a trough metallophone depicted as early as *c.* 800 AD on the Borobuḍur *stūpa* (Buddhist monument), Java, and the frame metallophone *gender*, now usually supplied with tubular resonators, which has been known from the 12th century on. Introduced to China by a Turkic people in the 7th century, the horizontal type of metallophone reached Korea in the 12th century and is still occasionally played there. In Japan it is a Buddhist ritual instrument.

Slit-drums. The slit-drum is an instrument made by hollowing a log or wood block through a slit. Those of east Asia and Indonesia are of great antiquity and of a high degree of development. The Chinese *mu yü* (traditionally fish-shaped) is a Buddhist and Taoist ritual slit-drum. Its Korean and Japanese counterparts are likewise ritual time markers, while in Vietnam the slit-drum is both a temple and a watchmen's instrument. On Java slit-drums can be traced to the Hindu–Javanese period (1st–9th century AD). Small models are generally suspended vertically, whereas larger ones rest horizontally; some underscore dance rhythms; others are signal instruments.

Tubular chimes and gongs. Southeast Asia is the home of tubular chimes; resonant, tuned bamboo tubes are united to form a chime in rural Annam and Java. In western Java up to 16 tubes are strung in ladder formation and suspended from a house or a tree and played with padded beaters.

Western Asia is believed to be the home of the gong, which reached China in the 6th century, Java by the 8th. Originally, it seems to have afforded protection against evil spirits, and even today the Iban people of Borneo beat gongs during a storm. Chinese gongs are both bossed (knobbed) and flat, those of Indonesia are bossed, while those of India are flat. Chinese gongs serve a whole gamut of purposes: as signal instruments of the army, as rhythm makers to accompany a song and, in the case of the bossless *lo*, as a Buddhist ritual instrument. In Japan the employment of the gong varies from temple to theatre to folk festival.

Gong chimes in Indonesian orchestras

In Indonesia and eastern Asia tuned gongs are united to form gong chimes. The Chinese upright *yung lo* is a Buddhist and Confucian cult chime and was formerly also played at court. Better known are the horizontal gong chimes of Indonesia (called *bonang* in Java), outstanding components of Southeast Asian orchestras and known from the 10th or 11th century on. Individual gongs comprising a *bonang* are often referred to as "bonang kettles," from their sharply bossed and deeply rimmed form. Frames of Thai gong chimes arch upward at both ends to form an upright semicircle. Ancient kettle gongs, products of Bronze Age culture, are found only in China, Indochina, and Indonesia. They are hung so that the striking surface is vertical and are struck in the centre of this surface (the head). Kettle gongs were cult instruments connected with rainmaking; those of the Dong Son culture area of North Vietnam may be dated from the 5th to the 3rd centuries BC.

Bells. Bells of bronze dating from about 1000 BC have been excavated in India, from the 22nd dynasty (945–730) in Egypt, and from *c.* 700 in Assyria. An intimate connection existed in ancient India between bell sound and mystic experience, and today a handbell is still rung in temple ceremonies in India and other areas of Buddhist influence. The Chinese differentiate between clapper bells (*ling*) and clapperless ones (*chung*); their temple bells, like those of Japan, are always of the *chung* type. Temple bells usually assume the form of chimes in China; one from the 6th century has 13 bells, but more modern chimes consist of 16 bells hung in two rows of a

frame. In addition to metal bells of the high-culture domains, rural areas still use primitive wooden bells with single or multiple clappers.

Rattles, scrapers, and split idiophones. The sistrum in its earliest form, that of a U, is seen in Sumer in the mid-3rd millennium BC, and a little later, about 2100/2000, a rectangular form appears in Horoz tepe, Anatolia (modern Turkey). Egyptian sistrums are characterized by being closed at the top: the *sesheshet* was shaped like a temple, the *iba* like a closed horseshoe. Sacred to the Egyptian goddess Hathor, the *iba* was played only by women, and after Hathor's metamorphosis into Isis it remained sacred to Isis.

A sliding rattle called *angklung*, found only in Indonesia, consists of several tuned bamboo tubes with cut-back tongues, inserted into a frame; they slide back and forth when the frame is shaken. When the *angklung* is constructed as a chime, each player shakes his frame when it is his turn to sound it.

Scraped idiophones are Confucian ritual instruments in east Asia. The "tiger" of China is a vessel scraper of wood with a series of notches cut along its back; known as early as the Chou dynasty (*c.* 1122–221 BC), it is scraped with a split bamboo stick. Split idiophones (made of hollow cane split between nodes) are slapped against the player's wrist or thigh in south China and Southeast Asia, producing a fixed-pitch buzz.

Plucked and friction instruments. Jew's harps have idioglott tongues (cut from the same material as the frame) in Indonesia, while both idioglott and heteroglott (the tongue made separately and attached to the frame) forms occur in China and on the Indian subcontinent. In southwest Yunnan (in southern China), three separate idioglott, tuned Jew's harps may be united by young people for the purpose of exchanging messages; such instruments are of bamboo and have the older, clothespin form. More recent are those of metal having the outline of an onion. In the Philippines and some other areas Jew's harps are important adjuncts to courtship.

Musical cups

Musical cups, the forerunners of musical glasses, are depicted on the Borobuḍur *stūpa* of *c.* 800; the Indian *rastrarang* can be played either with small sticks by percussion or by rubbing wetted fingers along the rims—the cups do not contain water. But the *jaltarang*, also of India, makes use of water for fine tuning and for the playing of *gamaka*s (ornaments) by carefully bringing the sticks into contact with the surface of the water. Similar musical cups are played in Japan in Buddhist temples and in the music of the Kabuki theatre.

IN AFRICA AND AFRO-AMERICA

Muslim Africa. Idiophones of Muslim Africa are mainly those of the Near East or derivations thereof. Outsize hollow clappers shaped like a dumbbell sliced lengthwise are clicked by Moroccan singers, who hold a pair in each hand. A percussion bowl, an inverted bowl set afloat in a larger, water-filled bowl, is beaten with a single stick by the Saharan Tuareg as their substitute for the West African water gourd. Finger cymbals are worn pairwise on finger and thumb of each hand by dancers. Basically, however, the percussion of this area is executed by drums. Idiophones do not participate in art music.

Black Africa. In contrast, Black Africa has an almost bewildering variety of idiophones. Clappers, generally of wood, are played from coast to coast. Simple percussion sticks are known in East and West Africa.

Xylophones. Xylophones vary from the primitive pit xylophone, with few wooden bars placed over a pit or trench and isolated from the ground by bundles of grass or by wooden members, to the leg xylophone, with a couple of bars placed on the outstretched legs of a woman player, to the large instruments with independent framework and tuned keys graduated in size. African xylophones are usually provided with a gourd resonator suspended from each key, often containing a mirliton device that adds a characteristic buzzing quality to the tone. Ensemble playing of several xylophones, reported by 17th-century travellers, is still practiced. In some areas xylophones form small orchestras with several players to one larger instrument. Unmistakable similarities of form, playing technique, tuning, and even of the music performed confirm the African xylophone's Southeast Asian origin.

Xylophone ensembles

Slit-drums. Slit-drums occur in the Western "Bend" and Congo Basin. They may be cylindrical, boat-, wedge-, or crescent-shaped, and zoomorphic with dorsal slit. Possibly related to the drummed tube zither, a cylindrical slit-drum with from two to five slits is encountered in West Africa; the Kissi people of Guinea not only strike the slats formed by the multiple slits but the ends of the slit-drum as well.

Bells. Bells occur in a large variety: metal and wooden, single or double, clapperless and with single or multiple clappers, played as rhythm or signal instruments. The apotropaic qualities of metal bells are recognized in Africa also, where such bells may be associated with chieftainship. Double bells, usually of metal, share a common handle: if this is straight, a dumbbell form results, but if bent, the bells lie cheek by jowl. Those of a pair differ in length or in diameter and, consequently, always in pitch. Metal double bells are devoid of clappers and are often made of sheet iron soldered down the side, whereas double bells of wood often have multiple clappers.

Water gourds, rattles, and scrapers. Water gourds—half gourds floated open side down in a pan of water and struck rhythmically with small sticks—are played in West Africa; in Dahomey their chief use is at funeral rites.

Construction of rattles

Rattles are the instruments *par excellence* of dancers, although they can also be worn on the leg simply to provide walking rhythm. Strung and gourd rattles are common, with the latter often serving in religious cults or magic rites in the Congo Basin area. In West Africa a variety of rattle sounded by external percussion was developed: the surface of a gourd is covered by a network of dried fruit seeds threaded on a cord. Sistrum-like forked sticks strung with threaded pairs of calabash fragments have an important part in fertility and initiation rites among the Mali of West Africa. The ancient horseshoe and rectangular forms of sistrum are still played in Ethiopia in the Monophysite Christian Church.

Scrapers in both solid and vessel form are widespread; a notched calabash, the *quilando*, was described by Girolamo Merolla in 1692.

Plucked instruments. The African linguaphone is known to the West by the native name sansa (zanza) or the Bantu name *mbira* and in English additionally by the misnomers "kaffir piano" and "thumb piano"; it consists of a varying number of cane or metal tongues fitted to a wooden board or resonator so as to permit one end of the tongues to swing freely; their different lengths determine the pitch. Sansas show affinity of tuning with xylophones that cannot be considered accidental, and thus they may be considered as small, portable versions of the latter.

Afro-America. West African idiophones introduced into the Americas with the slave trade are still flourishing. Clappers originating among the Yoruba (a Nigerian people) are played in Cuba; the claves, or pair of cylindrical percussion sticks of Haiti and Cuba, are now standard equipment in Western rhythm bands. The xylophone may already have entered America in pre-Columbian times. Known chiefly as marimba, one of its African names, it has been accepted in Western musical culture. Bells frequently figure as voodoo cult instruments in Caribbean and Afro-Brazilian communities. In Cuba water gourds are played at funeral rites for Afro-Cubans.

Steel drums

Steel drums originated in Trinidad during World War II and have become vehicles for popular music in the U.S. since. Actually a single "drum" forms a chime, as one head of an ordinary oil drum is divided into sections of different sizes by punched grooves, each section then being tuned to its own pitch. Today only a greater or lesser rim, depending upon the range required, is left but originally the oil drum was left intact.

Rattles play an important part in African cults of the New World, both gourd rattles with internal and external percussion and a distinct variety consisting of two metal cones joined at their widest part. (The maraca gourd rattle is probably indigenous to South America.)

Scrapers are highly popular. The notched gourd with natural handle, called guiro, is another Afro-American instrument taken over by the West. Notched turtle carapaces are scraped in the Caribbean. The jawbone of a horse, mule, or donkey, with its teeth left in, is played in both North and South America and in the Caribbean; its use among coastal Negroes of Peru goes back to the 18th century. In the United States it is still encountered in Louisiana and the Carolinas.

Sansas in the Caribbean and in South America bear such names as *marimbula* (Cuba) and *marimbao* (Brazil), clearly indicating their origin.

IN AMERICA

A wide assortment of idiophones is available to American Indians, but many of these are restricted to nonmusical uses. Concussion sticks, for instance, serve as game calls in North America, while concussion stones are invariably ritual: they are clashed to make thunder. Small, conical bells of metal and multiclapper bells of wood were known in ancient Peru.

Slit-drums have been played in Central and South America since pre-Columbian times, but their occurrence in South America is now rare. Characteristic of the well-known *teponaztli* is the form of its slits, cut to form an H with tongues of different thicknesses, thus allowing it to emit two sounds. Formerly, Zapotec warriors of Ixtepeji, Mexico, went into battle carrying an idol and singing to the accompaniment of the *teponaztli*, while Indians of 16th-century Hispaniola danced to their slit-drums.

Strung rattles are worn as leggings to emphasize a dancer's movements, but when the strung material consists of a dead enemy's teeth, as among the Brazilian Mundurucú, the rattle becomes a source of magic strength to the wearer; elsewhere, strung deer or caribou hooves attract game during the hunt. Vessel rattles of gourd and pottery imitations of gourds have been in use since Aztec times. Cult instruments in prehistoric America, they are still used by medicine men during their incantations. Hollow, seed-filled staffs were ceremonial rattles of Aztec and Maya, and descendants of these instruments are still played in North America.

Scrapers are seen on the 8th- or 9th-century frescoes of the temple at Bonampak in Chivas State, Mexico, played in a procession. Today among the Pima Indians of Arizona, scrapers play an important part in rainmaking ceremonies. Elsewhere in North America they serve as time markers. Scraped sticks of animal or human bone or of antler, with a series of notches practiced along one side, were sounded at sacrificial or memorial ceremonies of the Aztec Empire. In modern times they have been played in North America by medicine men. Split idiophones were employed in western coastal areas of North America chiefly for ceremonial purposes; some tribes slapped them against their chests, others against their palms. Finally, a curious friction-instrument vessel of Central and South America, the rubbed tortoise carapace, is a ritual instrument and taboo to the noninitiated; the instrument is rubbed with the fingers, not scraped.

IN THE PACIFIC

The different cultures of this wide area where singing and dancing form so prominent a part of the musical life include a number of primitive idiophones, some of which are played with considerable sophistication. Concussion sticks are clashed by an Aboriginal Australian singer to lend emphasis. The Maori breathe words of a song onto a carved stick held between their teeth while tapping it with a second stick. In Hawaii concussion stones were held pairwise by dancers who clicked them together like castanets. In New Britain and Papua primitive log xylophones are played, consisting of two banana stems or other logs placed on the ground with a few keys placed across them.

Aboriginal concussion sticks

A percussion gourd of New Guinea has an hourglass form, open at both ends; when plunged in and out of water, it is said to emit the sounds "uh-ah-uh-ah". Split-bamboo percussion tubes, doubling as rattles, are indigenous to the same area, while the percussion board *o-le-polotu* of Samoan and Tongan chiefs accompanies solo songs. Slit-drums can be huge. Made from a tree trunk, living slit-drums in the New Hebrides are carved with the faces of ancestors, and in New Guinea roofed drum houses are built over large, horizontal slit-drums to protect them from the weather. The Maori, one of the relatively few peoples who have no membrane drums, use their slit-drums as signal instruments. Related to these is a percussion tube of Fiji, partly hollowed and tapering at the ends and struck with two sticks on the side opposite the slit.

Slit-drums

Idioglott Jew's harps of bamboo are the major tonal idiophones of the area; here the tongue is vibrated either by the player's finger or by jerking it with a cord. A friction vessel of New Ireland consists of a rounded and hollowed block of wood, its upper portion carved into three tongues of different lengths and hence emitting as many pitches. Already rare nowadays and taboo to women and children, it is played at death commemoration rites by a male performer who rubs well-oiled hands over the tongues.

Membranophones

Musical instruments in which the sound-producing medium is a vibrating membrane fall into four main groups: kettledrums and bowl-shaped drums; tubular drums—whether cylindrical, barrel, conical, double conical, hourglass, goblet, or shallow—and rattle drums, the membranes of which are set in motion by enclosed pellets or by knotted ends of a thong or cord; friction drums, with membranes caused to vibrate by friction; and mirlitons, musical auxiliaries with membranes set in motion by directing the vibrating air column of an instrument or voice against them.

Kettledrums and tubular drums occur in both tunable and nontunable forms; friction drums and mirlitons are not tunable. The membranes of the first two groups are either glued, nailed, lapped, or laced to the body, or shell; if they are glued or nailed, the pitch can be modified by exposure to heat. Lapped and laced heads are readily tunable by tightening the lacings or screws, and wooden wedges may be inserted between the shell and lacings to further increase the membrane's tension and thus raise the pitch. The membranes of such instruments and of friction drums are set in vibration by percussion, while those of mirlitons vibrate by impact of sound waves. In all groups the shell plays a subordinate acoustical part, acting as resonator only. The greater the diameter of a head, the deeper is its sound; and the greater its tension, the higher is the pitch. In Western culture the only drums tuned to a definite pitch are kettledrums (the orchestral timpani).

Kettledrums and tubular drums may be struck by the hands, with beaters, or both combined or by the knotted ends of a thong or cord. Beaters can be cylindrical, club-shaped, straight, curved, or angled, with or without knobs or padding, or may take the form of a switch or wire brush. Friction drums are sounded by rubbing the membrane with a piece of hide or by the more usual method of working an inserted friction stick or cord up and down or by rubbing the membrane with a player's wet fingers. Acoustically, they are subject to the same laws as other membranophones, but the speed of friction is an influencing factor. They occur in Africa, America, Europe, Asia (India and Japan), and Hawaii. Mirlitons are not true musical instruments but always voice modifiers, be it a human or an instrumental voice (as when affixed, *e.g.*, to African xylophone resonators); as such they have no pitch of their own.

Methods of striking drums

Primitive ground drums, consisting in their simplest form of an animal skin stretched over the opening of a pit, are still found in many parts of the world. A skin may also be held taut by several players, each beating it with a stick. These and similar ground drums are played by women in Africa and Australia, in North America usually by men. By their very nature ground drums are nonportable; when need arose for a smaller "drum," a skin was stretched over the opening of a gourd, clay pot, or other object. Among the Swazi of South Africa such skins are not attached but held taut. Whether the skin is attached

to or held taut over a pot, etc., such instruments are known as pot drums. They are found in Asia, Africa, and America—in Africa and America often in connection with exorcism.

IN EUROPE

Antiquity. Only a few kinds of drum, none indigenous, were known to antiquity. The frame drum came from the East at an early date. The barrel drum was possibly known in Hellenistic times, for it appears in the Greco-Indian culture of Kushan. A shallow drum is depicted on a Greco-Scythian metal *gorytus*, or bow and arrow case, of the 4th century BC, but there again no evidence exists of its having been known to Greece. Frame drums, traditionally women's instruments in the Near East, remained in the hands of female players and dancers in Greece, and the Eastern playing technique was maintained: held upright on edge in the palm of one hand, the drum was tapped with the fingers of the other.

The frame drum

Probably introduced into Greece in the 6th century BC with the cult of the great mother-goddess Cybele, the frame drum is depicted as being played by maenads and was also a cult instrument in rites of the Orphic religion. From Greece the frame drum passed into Rome, where it was also associated with Cybele. Romans also had double-headed drums and spread them as far north as the Isle of Wight, in Britain, where one is represented on a mosaic pavement as a dancing girl's instrument. In the late days of the Roman Empire, frame drums became instruments of street musicians and *joculatores*; the latter are sometimes portrayed throwing them up in the air and catching them again and are probably responsible for spreading them beyond the Italian Peninsula.

The Middle Ages. Frame drums were described by Isidore of Seville in the 7th century, but they seem to have played a negligible role in medieval Europe until reintroduced by the crusaders, after which they appear both in round and square form in the hands of both men and women. Jingles and pellet bells, frequently attached from the 14th century on, enliven the sound. French medieval poems place the instrument in the hands of female dancers, and in the 14th century Boccaccio writes of it as accompanying the voice.

Double-headed drums served as time beaters from the 6th century on. From the 7th century dates the first evidence of their being played with drumsticks, a technique adopted from Asia. Hourglass drums appear in southern Europe by the 10th century, only to die out in the late Middle Ages. The small cylinder drum known as the tabor was provided with a snare (material stretched across the head to produce a rattling sound) and struck with a stick right on the snare. Played together with the three-holed pipe, it formed a lively one-man ensemble, the companion of country dances throughout western Europe for centuries; it is still in use in some areas.

The naker

With the Crusades the small kettledrum called naker, always paired, entered Europe, reaching Russia in the 11th century and the West at the turn of the 12th and 13th century. A pair of nakers would be attached by a belt to the player's waist or suspended from the shoulder, each struck alternately as the heads sounded at different pitches. Not until the mid-15th century do the wide-diameter horse-borne kettledrums reach the West. Henceforth, they were to be paired with trumpets and associated with royalty and the cavalry, whereas the marching of infantry was regulated by plebeian fife and *Landsknecht*'s (lansquenet's) drum, a cylinder drum of considerable proportions slung to the player's left. As mercenary soldiers from the late 15th century on, Swiss drummers and fifers popularized such ensembles throughout western Europe. The shallow drum, presumably of Turkish derivation, is seen in the late 14th century. Its two heads were at first nailed in Oriental fashion, but by the mid-15th century the instrument was suitably Occidentalized by the use of laced heads.

The less common friction drum first appears in the late 12th-century sculpture of the archbishop's palace at Santiago de Compostela, in Spain, complete with friction rod. Although it undoubtedly continued in existence, it is not known to be either pictured or mentioned again until the 17th century.

Modern times. By the Renaissance, Europe had a variety of drums performing specialized functions: frame drums and small tabors accompanied dance and song; larger tabors served as time beaters in small mixed ensembles; great cylinder drums with fifes were placed at the disposal of foot troops; large kettledrums and trumpets were restricted to cavalry and ceremonial music. Frame drums remained popular in the south, particularly in Spain, as dancers' instruments; from Spain they travelled to Hispaniola by the mid-16th century.

Before the late 17th century, drums in the West served principally at court or other ceremonial functions and increasingly as military instruments, although smaller models filled social needs of everyday life. The 17th-century French composer Jean-Baptiste Lully appears to have been the first to admit drums and kettledrums to the opera and even, according to Titon du Tillet, 18th-century French writer and courtier, into church music. In the year of Lully's death (1687), Daniel Speer, a contemporary German musician, commented that echo effects could be obtained on kettledrums by playing softly close to the rim and striking hard at the centre; with such new techniques, modified drumstick heads, and the possibility of notating their music (hitherto prohibited by the rules of secrecy imposed upon guild members), kettledrums, henceforth called timpani, triumphantly entered orchestra, opera and church.

Admission of kettledrums to the orchestra

"Turkish" drums, descendants of the shallow drum, were admitted to the orchestra and military band in the 18th century. The composer Christoph Willibald Gluck was an early writer for this instrument in his opéra comique *Les Pélerins de la Mecque* (*The Pilgrims of Mecca*, 1764); also known as *La Rencontre imprévue*), and Mozart scored for it in his *Die Entführung aus dem Serail* (*The Abduction from the Seraglio*, 1782). Still called "Turkish" drum in 1823 by the French musicologist Guillaume-André Villoteau, a member of Napoleon's scientific mission to Egypt, it began to be known as the bass drum around this time. Another type sometimes called bass drum is the long drum, a military instrument of Turkish origin. Great, battle-scarred specimens of the 18th and 19th centuries in museums attest its continued use in Europe in its original form. The old *Landsknecht*'s side drum became too unwieldy and was cut down to the proportions of the modern side or snare drum. It may also have been the ancestor of the tenor drum, the original function of which was to play rolls and which was gradually eased out of military bands in most countries in the 19th century in favour of the far lighter side drum. The long drum continues in use as a folk instrument in the Balkans.

The northern frame drum, or tambourine, was given the status of a salon instrument by 18th-century French society, and, combined with harp or keyboard instrument, it could be heard at fashionable soirees. Although it is rarely written for, Weber uses one for local colour in his opera *Preciosa* (1820), Berlioz in his overture *Carnaval romain* (*Roman Carnival*, 1844).

Friction drums have maintained an existence to this day in various parts of Europe, where they are played at Christmas, during the carnival season, or to greet the New Year in less industrialized areas.

Mirlitons, with their ability of amplifying and colouring tone, have been known in Europe since the 16th century, possibly earlier, but did not gain popularity until the early 19th century. From 1883 on, a French toy maker named Bigot brought out a series of bigotphones shaped like orchestral instruments, and as late as 1910 gatherings of "bigotphonists" took place in Paris. Popular in mid-20th century was the tubular kazoo.

Changes in drumming technique

Drumming in the 20th century gained an important place outside its traditional preserves of orchestra and military band: rhythm bands and experimental music opened hitherto unexplored fields. The latter was preoccupied with rhythmic stress and also exploited drum tone for its own sake as a source of independent sound and tone colour. Playing techniques consequently became

freer, players striking either the membrane or the shell, using bare hands or a large range of beaters varying from stiff to flexible and hard to soft; wire "brushes" were frequently substituted for sticks in rhythm bands. The drum was treated as a solo instrument amid a group of its peers. On the popular level, rhythm may be stressed or punctuated, often by the clarity and opposing pitches of a pair of Afro-American drums (usually referred to as Latin-American drums).

IN ASIA

Unlike the drums of Western musical tradition, those found in ancient (and parts of modern) Asia are primarily cult and ceremonial instruments. Most, if not all, Western drums are their descendants, and west Central Asia may well be their common homeland. Babylonia already had a variety of forms: cylinder, hourglass, goblet, and bowl-shaped, all of terra-cotta and all beaten with bare hands. Assyrians also carried in procession a large, conical drum played in the same manner.

Babylonian drums

Temple drums were of considerable proportions: huge frame drums existed from the 3rd millennium on in Mesopotamia, and the waist-high *lilissu* had a goblet form—a bowl on a stand—and may have been the ancestor to the kettledrum. All these were played by men, but the smaller frame drums appearing in Sumeria in about 2000 BC are depicted in the hands of women. An early Sumerian text informs us that the king's granddaughter was appointed player of the *balag di* in the Temple of the Moon at Ur around 2400 BC. Ever since, frame drums have been predominantly women's instruments. The Bible says that in ancient Israel "Miriam, the prophetess, the sister of Aaron, took a timbrel in her hand; and all the women went out after her with timbrels and dancing." They are still played throughout the Near East, in some areas in art music ensembles, in others only in popular and folk music.

Wherever the Muslim presence has made itself felt, large and small kettledrums are encountered. In Turkey classical monophonic singing is accompanied solely by a small kettledrum, played either in the centre of the head or close to the rim, depending upon whether clear or muffled beats are desired.

Association of drums with thunder

Thunder drums. In many regions of the ancient and modern East drums were believed to create or imitate thunder. Sets of thunder drums are depicted on East Turkistan art objects from the 9th century to the 11th century. In China, Lei Kung, duke of Thunder, is traditionally surrounded by numerous drums on which he creates thunder, and in Japan thundering drums were even automated by attaching a number of them to the outer circumference of a wheel that, when revolved, caused them to rattle—an early application of the rattle-drum principle. As in Africa and America, cult drums of Asia have been associated with human sacrifice; thus the Chinese *Tso Chuan* ("The Commentary of Tso," one of the classics of Confucian literature) mentions the consecration of drums in the 7th and 6th centuries BC by smearing them with the blood of a sacrificial human, usually a war captive.

Rattle drums, either containing in their cavities objects that rattle when the drum is twirled by a handle or with attached cords having knotted ends that strike the heads, can be traced back to the Shang dynasty (*c.* 1766–*c.* 1122 BC) and are now in use from Tibet and India to Japan. Still a cult drum in some areas—it is an attribute of Śiva (Shiva), the third member of the Hindu trinity—in others the rattle drum has degenerated to a mere toy. On the giant *stūpa* of Borobuḍur, Java, an archetypal image of the known cosmos, are sculpted the musical instruments in use at the time of its erection (*c.* 800). These include cylinder, conical, hourglass, and pot drums—all Indian instruments. Yet, today's Indonesian orchestra, the gamelan, virtually ignores drums in favour of metallophones. Drumming is exclusively a man's business on some islands of the Malay archipelago, where drums are kept in the men's house, out of sight of women.

Drums played an important part in early court orchestras of China and Korea, judging by their surviving elements in *gagaku*, court orchestral music of Japan. Here, the leader beats time on a drum, either cylindrical or hourglass-shaped, having projecting heads, a characteristic feature of Japanese hooped drums. Larger barrel-shaped drums with nailed heads are suspended from elaborate frames. On the Indian subcontinent a variety of membranophones participate in art music ensembles or accompany dances, while small kettledrums are played pairwise in processions; large, single kettledrums are temple instruments or may be mounted on elephants' backs in outdoor ceremonies. Many smaller drums have laced and wedged heads, the skin being further tensed by an application of tuning paste. In Burma, Thailand, and south Nias Island a number of tuned barrel drums of varying sizes are suspended in a low circular pen to form drum chimes; such a chime may consist of from 16 to 24 drums, all played by a single musician with his fingers and palms, occasionally with a drumstick. These chimes form part of the theatre orchestra and are even played in processions.

Drum chimes

Mirlitons. In East and Southeast Asia a mirliton device is applied to many woodwinds in order to obtain a characteristic modification of tone quality. To this end an additional aperture is cut in the tube and covered with a thin membrane. A different application of this principle is encountered in India, where a pair of short, wooden mirliton "trumpets" are placed against the throat of a humming performer. During the 1950s a similar technique was employed by the United States recording industry for the obtaining of anthropomorphic effects.

IN AFRICA AND AFRO-AMERICA

Muslim Africa. Drumming tradition in the Muslim north centres on smaller, portable drums, such as frame, goblet, and small kettledrums. Frame drums are made in a number of forms: circular with single membrane, square, even diamond-shaped (ancient Egypt had circular, square, and rectangular ones). Islāmic cult admits no musical instruments, and the human voice is the preponderant instrument here, hence the frame drum's important role as vocal accompaniment. It also lends rhythm for dancing and is the basic percussion instrument of Muslim art-music ensembles. In Morocco groups of women may sing to the accompaniment of a frame drum supplied with jingles, a spike fiddle (one in which the handle traverses the body and emerges at the lower end to form a spike), small cymbals, and a pair of kettledrums, each playing her own instrument.

Black Africa. Black Africa makes perhaps wider use of its membranophones than does any other culture. Numerous variant forms of drums and lacings make classification difficult; indeed, a study of types remains to be established. Some drums are played as musical instruments, others transmit messages ("talking" drums), some are restricted to cult and funerals and others—partly desacralized—to royalty, while yet others participate in everyday life. Ethiopia admits drums to the church, while West African cult drums may not be seen by the uninitiated. Drums are beaten with bare hands or with rectangled or knobbed sticks. Footed drums (*i.e.*, with a base prolonged to form "feet") attain a height of ten feet in the Loango area of west Central Africa and must be tilted to bring the head within the performer's reach. The playing head of hourglass drums may be struck with one hand and with a stick alternately. Kettledrums are royal, ritual, or ceremonial, and unlike their North African counterparts they are of wood, often with sculpted shells. Their traditional pairing with trumpets reaches as far as south of the Congo.

"Talking" drums

That portion of West Africa known as "the Bend" is the area of "talking" drums, by means of which messages are conveyed for up to 20 miles, to be relayed by another drummer. Languages of this area are characterized by pronounced high and low pitch tones (tone languages), a quality exploited when two drums—a lower pitched, or male drum, and a higher pitched, or female one—transmit low and high tones respectively. Accent, number, and pitch of the syllables are transmittable. Among the Yoruba a "talking" drum set consists of four hourglass

drums and a kettledrum; the leather lacings of the former are gripped by its player, enabling him to change the pitch as he exerts more or less pressure on them; the chief drum of the set is capable of an octave range and, in addition to tones, produces also the typical Yoruba glides by manipulation of the lacings.

East African drum chimes, of Asian heritage, are tuned to specific pitches; these instruments attained royal status in Uganda, where the largest chime consists of 15 drums requiring six musicians to play them.

Friction drums of the lower Congo area were once exclusively cult instruments but are now becoming desacralized. Whereas in Central Africa they are played only by men, women of the South African Pedi play them at female initiation rites.

African mirlitons can be most imaginative: standard material is a spider's egg membrane, and this may be added to apertures pierced in the bottom of xylophone resonators or applied to one end of an independent cane tube that is inserted into a nostril, as among the Fan of Gabon.

Afro-America. West Africans reconstructed their native drums in the New World, preponderantly as cult instruments. Small sets of two or, more often, three drums of graded sizes form an integral part of Afro-American rituals and, in the Caribbean, also of voodoo dances. Only the skins of sacrificial animals may be used for membranes among Afro-Bahians in Brazil, who baptize their new drums, prefereably with "holy" water obtained from a Catholic church. Drums in Haiti are sacred objects and may even represent the deity itself: as such they receive libations. One type of bongo drum—there exist at least four of these—has been adopted by Western rhythm bands, as has the conga.

IN AMERICA

Pre-Columbian drums of Central America appear to have been played without sticks, regardless of their size, and to have been devoid of lacings, whether they were small pottery drums, such as those excavated in Costa Rica, or the large footed drums of Mexico. Slender pottery drums of the Guatemala highlands, open top and bottom, can be dated to the late classical period (*c.* 700–1000). Skeletons of wooden cylinder drums, very shallow, have been found in Peru. In shape but not in structure they are identical to the shallow drums played by North American Indians today.

The North American shallow drums are "medicine" (magic) and dance drums, having a heavy hide head struck by a hard beater and emitting a loud, staccato sound. Larger models are suspended and struck simultaneously by up to ten drummers seated in a surrounding circle. The "dream-dance" drum, a tub-shaped shallow drum, is elaborately decorated and has a bell suspended in its interior; it is credited with great medicine power.

Cylinder drums, made of a hollowed log of cottonwood, traditionally war and dance drums of tribes of the Southwestern United States, are now sold in tourist shops. Pottery drums, either potbellied, bowl-shaped, or footed, were formerly common among the eastern and southern tribes of the United States; the potbellied type remains in use among the Pueblo Indians of the Southwest.

American Indian frame drums are usually circular; square shapes are rare, but rectangular frame drums are found in California, Mexico, and Bolivia and octagonal ones among the Cherokee of the Southeastern United States. All were war drums, regardless of whether they had one or two membranes. By adding a rattling device, a frame drum is converted into a medicine drum. The Eskimo frame drum, a shaman's instrument, is distributed over Greenland, northern Siberia, North America, and among the Lapps of northern Scandinavia and differs from other frame drums in that it has a fixed handle and is struck on the hoop, not on the membrane. It may be circular, oval, or fan-shaped.

Eskimo shaman's drum

Friction drums are rare; in Venezuela one is formed from a small barrel over which a piece of leather is stretched, with central perforation to admit a friction stick.

IN THE PACIFIC

Two forms of drum are distributed over this huge area. Elongated hourglass drums, usually with wooden handles of a piece with the drum shell, often beautifully carved and ranging up to six feet in height, are commonly in the hands of dancers and singers and are said to symbolize the transition from earth to heaven. Among the Wapenamundu of New Guinea they lack handles and are war drums. Such hourglass drums are not found in Polynesia. The other form is a conical, footed drum made of a tree trunk, reaching great dimension. Those of Polynesia are sometimes eight feet high and have elaborately carved pedestals. They are usually made of breadfruit tree wood or hollowed coconut palm, with a single, sharkskin head. Throughout the area drums are beaten with bare hands.

Whirled friction drums are met with in Hawaii, where a large nut or a calabash is furnished with a friction cord and swung through the air.

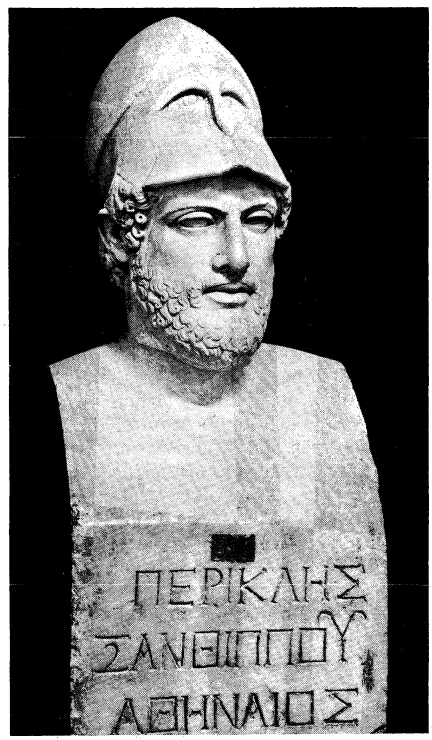
BIBLIOGRAPHY. JAMES BLADES, *Percussion Instruments and Their History* (1971), world-wide survey with details on playing techniques, by a professional percussionist; CURT SACHS, *The History of Musical Instruments* (1940), the classic of its field including important synoptic coverage of both idiophones and membranophones; SIBYL MARCUSE, *Musical Instruments: A Comprehensive Dictionary* (1964), contains short articles on individual instruments.

(S.Ma.)

Pericles

The greatest statesman of ancient Athens, Pericles brought Athenian democracy to its zenith, built the most magnificent of ancient Greek buildings, and came near to establishing Athens as the leader of all Greece.

Anderson—Alinari



Pericles, marble herm. In the Vatican Museum.

Knowledge of his life derives largely from two sources. The historian Thucydides admired him profoundly and refused to criticize him. His account suffers from the fact that, 40 years younger, he had no firsthand knowledge of Pericles' early career; it suffers also from his approach, which concentrates exclusively on Pericles' intellectual capacity and his war leadership, omitting biographical details, which Thucydides thought irrelevant to his theme. The gaps are partly filled by the Greek writer Plutarch, who, 500 years later, began writing the life of Pericles to illustrate a man of unchallengeable virtue and greatness at grips with the fickleness of the mob and

finished rather puzzled by the picture he found in his sources of Pericles' responsibility for a needless war. These sources are not all ascertainable, but they certainly preserve an invaluable amount of fact and contemporary gossip, which is sometimes nearly as useful.

Family background and education. Pericles was born (*c.* 495 BC) into the first generation able to use the new weapon of the popular vote against the old power of family politics. His father, Xanthippus, a typical member of this generation, almost certainly of an old family, began his political career by a dynastic marriage into the controversial family of the Alcmeonids. He soon left their political camp, probably on the question of relations with Persia, and took the then new path of legal prosecution as a political weapon. Perhaps outbid in his search for popular support, he was ostracized in 484 BC, though he returned in 480 to command the Athenian force at Mycale in 479, probably dying soon after. From him Pericles may have inherited a leaning toward the people, along with landed property at Cholargus, just north of Athens, which put him high, though not quite at the highest level, on the Athenian pyramid of wealth. His Alcmeonid mother, Agariste, provided him with relationships of sharply diminishing political value and her family curse, a religious defilement that was occasionally used against him by his enemies. A few days before Pericles' birth, according to the Greek historian Herodotus, Agariste dreamed she bore a lion. The symbolism, although ambiguous, is most likely to be unfavourable. That Pericles' skull was of unusual shape seems well attested, but one can hardly speculate about the possible psychological consequences.

Education by the Sophists

The only name associated with his early education is that of the musical theorist Damon, whose influence, it is said, was not just confined to music. The arrival of the Sophist philosophers in Athens occurred during his middle life, and he seems to have taken full advantage of the society of Zeno and particularly Anaxagoras, from whom he is said to have learned impassivity in the face of trouble and insult and skepticism about alleged divine phenomena.

The first known date in his life is 472 BC, when he paid for the production of the playwright Aeschylus' Persian trilogy. Nothing further is known until 463, when he unsuccessfully prosecuted Cimon, the leading general and statesman of the day, on a charge of having neglected a chance to conquer Macedonia; this implies that Pericles advocated an aggressive policy of expansion for Athens. Only rumour associates him directly with the political convulsion of the next two years, which drove Cimon into exile, swung Athens away from its alignment with Sparta, and decisively strengthened the democratic elements in the Athenian constitution; but he probably did support the democratic leader Ephialtes in this period, and his introduction of pay for juries, unfortunately undatable, is a logical consequence of Ephialtes' reforms.

Rise to democratic leadership. That Pericles immediately succeeded the assassinated Ephialtes as head of the democratic party in 461 is an ancient oversimplification; there were other men of considerable weight in Athens in the next 15 years. The outbreak of war among the Greek states in 459 put a premium on military talent, and Pericles' only recorded campaign in the next few years was a naval expedition in the Corinthian Gulf in 454, in which Athens defeated Achaea but failed to win more important objectives. Politically he is credited with some kind of rapprochement with Cimon, who is said to have been recalled and allowed to resume the war with Persia, much preferred to fighting other Greeks, but the date of Cimon's recall is uncertain, and the rumours are hard to disentangle.

Legislation on Athenian citizenship

In 451 or 450 Pericles carried a law confining Athenian citizenship to those of Athenian parentage on both sides. No source provides any background to this proposal; it is not even clear whether it was retrospective. A correct assessment is vital for understanding Pericles, but explanations vary considerably; some argue that Pericles was merely forging a low-level political weapon for use against Cimon, who had a foreign mother. The upper classes certainly had no prejudice against foreign marriages; the lower classes may well have had more, and, on the whole, it is possible to view Pericles here as championing exclusivist tendencies against immigrants who might break down the fabric of Athenian society. One hundred years later, an orator argued for firm distinctions of status on the ground that the law provided even the poorest Athenian girl with a dowry in the form of her citizenship. The law also may have passed because of a general wish to restrict access to the benefits of office and public distributions, but there was never any disposition on the part of Athenians to restrict economic opportunities for foreigners—who served in the fleet, worked on public buildings, and had freedom of trade and investment, with the crucial, but normal, exception of land and houses. To speak of this legislation as a move forward creating a "master race" is thus partly misleading, but the demagogic nature of the law seems clear.

Cimon died after 451, during his last campaign against Persia. The policy of war with Persia was abandoned and a formal peace probably made. The Persian War, begun as an ill-considered gesture in 499, could be considered ultimately successful. The city of Athens, however, was physically still much as it had been left by the Persian sack of 480, and its gods were inadequately housed.

Attempts to restore Athens' leadership among the Greek cities

Hostilities among the Greek states had also come to an end in the Five Years' Truce of 451. Pericles now embarked on a policy designed to secure Athens' cultural and political leadership in Greece. It had already dominated the alliance that had continued the Persian War after Sparta's withdrawal in 478, a leadership strengthened by the transfer of the alliance's considerable treasury from Delos to Athens in 454. If peace with Persia did not end the alliance, it may have ended the annual tribute paid to that treasury. Whether to regain this tribute, or simply to assert Athenian leadership, Pericles summoned a conference of all Greek states to consider the questions of rebuilding the Greek temples destroyed by the Persians, the payment of sacrifices due to the gods for salvation, and the freedom of the seas. Sparta would not cooperate, but Pericles continued on the narrower basis of the Athenian alliance. Tribute was to continue, and Athens would draw heavily on the reserves of the alliance for a magnificent building program centred on the Acropolis. In 447 work started on the temple later known as the Parthenon and on the gold and ivory statue of Athena (by Phidias), which it was to house; the Acropolis project was to include, among other things, a temple to Victory and the Propylaea (started 437), the entrance gateway, far grander and more expensive than any previous Greek secular building.

There was domestic criticism, however. Thucydides, son of Melesias (not the historian) and a relative of Cimon, who had inherited some of his political support, denounced both the extravagance of the project and the immorality of using allied funds to finance it. Pericles argued that the allies were paying for their defense, and, if that was assured, Athens did not have to account for how the money was actually spent. The argument ended in ostracism in 443; Thucydides went into exile for ten years, leaving Pericles unchallenged. It cannot be determined whether the glamour of the project had completely caught Athenian imagination or whether Pericles was now simply thought to be indispensable. Plutarch attributed to Pericles a desire to stimulate economic activity and employment in Athens, but these motives may be anachronistic and in actuality may not have influenced the voters very much.

Revolts within the Athenian Empire

There was also some initial allied resentment at the continuation of tribute, and some scattered revolts. Pericles met the situation in part by extending a network of Athenian settlements throughout what may now be called the empire, thus strengthening Athenian control and providing new land for the growing Athenian population. In establishing one of these, Pericles engaged in his most admired campaign, the expulsion of barbarians from the Thracian Chersonese (Gallipoli). A more serious crisis came in 447 or 446, however, when the cities of Boeotia, under Athenian control since 458, beat a small Athenian army and successfully revolted. Euboea, crucial

to Athenian control of the sea and food supplies, and Megara soon followed suit. The strategic importance of Megara was immediately demonstrated by the appearance, for the first time in 12 years, of a Spartan army north of the Isthmus in Attica. Pericles thought and acted swiftly. The details were never fully known, but, possibly by bribery and certainly by negotiation, it was arranged that Athens would give up its mainland possessions and confine itself to a largely maritime empire. The Spartan army retired, Euboea was quickly reduced, and the arrangement was ratified by the Thirty Years' Peace (winter 446–445). For Athens, the essential loss was that of Megara, which meant that a Spartan army could appear in Attica at any time. That Pericles doubted the stability of the settlement and saw the need to develop an alternative basic strategy for Athens is shown by his immediate construction of a third Long Wall to improve the defenses of Athens and the port of Piraeus. Henceforth, in effect, Athens could be turned into an island at will.

Political and military achievements. There was a break in tensions for the moment. After Thucydides' ostracism, Pericles had little domestic opposition. His position rested on his continual re-election to the generalship and on his prestige, based, according to the historian Thucydides, on his manifest intelligence and incorruptibility. From his youthful demagogy, he had moved to a more middle ground in politics, and there are traces in his later life of his being outflanked by more radical spokesmen. Athens was, Thucydides says, in name a democracy but, in fact, governed by its first man. Though Athenian democracy never gave more than severely limited powers to the executive, the assembly gave Pericles what he wanted. Thucydides, obsessed with the power of intellect, takes little note of the need of a statesman to work hard, and it is Plutarch who provided the glimpses of a man who took no interest in his own estates, who was never seen on any road but that to the public offices, and who was only recalled to have gone to one social occasion, which he left early.

Pericles' personal life

This picture is softened somewhat by what is known of his personal life. The identity of his wife, however, though certainly of wealth and high birth, is unknown. He married her in his late 20s but, as they were incompatible, divorced her some ten years later. Close to 50, he took Aspasia of Miletus into his house. By his own law, marriage was impossible, and, after the death of his two legitimate sons, their son Pericles had to be legitimated. Although she is clouded by scandal and legend, it is easy to believe she possessed great charm and intelligence. There is no reason to doubt that she was free and of good birth in her own city with its great intellectual traditions. It is clear that her own behaviour and Pericles' attitude toward her were surprising phenomena in Athens, where upper-class women were kept secluded. That Pericles was known to kiss her on leaving for and returning from work gave rise to speculation about her influence on him and, thus, on Athenian politics.

As the building program continued, Pericles demonstrated Athenian superiority in other ways. In 443 a Panhellenic colony was founded under Athenian auspices at Thurii, in south Italy, but did not form a continuing centre of Athenian influence in the west, as may have been hoped. At an unknown date, Pericles took a fleet into the Black Sea to demonstrate Athenian power and secure the corn route from south Russia. As the buildings on the Acropolis rose, celebrations of the festival of the Panathenaea grew more and more elaborate, and much was done to enhance the splendour of the Mysteries of Eleusis, symbolic, among other things, of the Athenian claim to have brought corn and civilization to mankind.

Pericles' last major campaign was the one interruption in these years. In 440, Samos, one of Athens' principal allies with a substantial fleet of its own, revolted, and, despite a victory by Pericles against superior numbers, the revolt nearly succeeded. The campaign to recover Samos, although long and costly, was completely successful, and became a model against which later Athenian generals measured their achievements.

There had been a serious possibility that Sparta and its allies might intervene on this occasion, but they did not, and the Thirty Years' Peace was upheld until the end of the 430s. Tension grew as the decade progressed, particularly with regard to Corinth, Sparta's ally, whose interests conflicted more obviously with those of Athens. By 433 the situation was serious enough for Athens' finances to be put on a war basis, and, thereafter, the drift to war continued. Pericles' policy was one of firmness, coupled with careful manipulation of the diplomatic position to keep Athens technically in the right. The firmness was a puzzle to contemporaries, particularly his determination to enforce decrees excluding Megarian trade from the Athenian Empire. Was he, it was asked, influenced by some private grievance of Aspasia? Was he trying to divert attention from personal attacks on himself and friends by making war? Thucydides tells just enough to make his own interpretation plausible, that Megara was a small matter in itself but crucial as a symbol of Athenian determination to maintain its position. Consideration of Megara's strategic importance, which Thucydides consistently undervalues, may suggest further the possibility that the Megarian decrees were not the immediate cause of the war but the first blow in a war Pericles thought inevitable and that began in spring 431.

The drift toward war

Pericles' main strategic ideas are clear. He was an admiral rather than a general, and Athens' naval resources were immeasurably superior to its land power. He would evacuate the Athenian countryside, bring the population into the Long Walls, decline battle with the Spartan army, and rely on the fleet to assure Athenian food supplies and secure the empire on whose resources the expensive naval policy depended. Expenditure on building had been counterbalanced by annual savings from the tribute, and enough capital had been reserved, he thought, for a long war, though expenditure turned out heavier than he could have calculated. This is essentially Thucydides' analysis, though he failed to explain what end to the war, other than a stalemate, Pericles wanted or expected. There are some indications that Periclean strategy included more aggressive elements, such as the recovery of Megara, which would have considerably improved Athens' position.

Weakness of Pericles' strategy

This strategy, however, had marked political weaknesses. The Athenian population had deep roots in the countryside, and great firmness was required to bring them to abandon their land to Spartan ravages without a fight. The middle class army suffered in morale, and the living conditions of the lower classes, though they were allowed activity in the fleet, deteriorated in the overcrowded city. The overcrowding had an unforeseeable consequence in a plague, which in the second summer of the war took a quarter of the population. No obvious success counterbalanced the discomforts of war, and Pericles was deposed from office and fined. He was soon re-elected, but he took no new initiatives before his death in autumn 429.

After the first campaigning season of the war, he had delivered the funeral speech over the fallen, which Thucydides reports at length. They had fallen, he said, in preserving a way of life that he described in detail. Athenian life often fell short of this Periclean ideal, but he conceived it with clarity and made it generally recognized. He conceived his Athens as "an education to Greece." If the last speech attributed to him by Thucydides is any guide, he cannot be accused of ignoring that the realities of power that made the Periclean age possible might also bring it down.

BIBLIOGRAPHY

Ancient sources: THUCYDIDES, *Histories*, bk. 1–2; PLUTARCH, *Life of Pericles* (both available in several translations).

Modern works: A.R. BURN, *Pericles and Athens* (1949), brief but good historical treatment with a critical bibliography; V. EHRENBERG, *Sophocles and Pericles* (1954), detailed discussion of aspects of the intellectual background; C.M. BOWRA, *Periclean Athens* (1971), a full picture of the period and its culture; J.K. DAVIES, *Athenian Propertied Families*, pp. 455–460 (1971), only on family and property, but with scholarly bibliography; R. MEIGGS, *The Athenian Empire* (1972), a detailed, scholarly survey.

(D.M.L.)

Periodicity, Biological

Continuous change characterizes both living organisms and their physical environment. Many of these changes occur irregularly and are termed aperiodic. Examples include the irregular variations in temperature, light, humidity, and other physical factors associated with the passage of weather systems; aperiodic, also, are the biological fluctuations in response to them. In contrast, are periodic fluctuations, relatively regular changes repeated at constant intervals of time. Clearly periodic, for example, are the natural fluctuations in daylight and in tides. Similarly periodic are the daily and tidal variations in plants and animals as they respond and adjust to these periodic changes. Phenomena that recur with such regularity are termed rhythmic. Fluctuations of organisms in direct responses to long term environmental fluctuations are termed phenological; such are migratory movements, flowering, breeding periods, etc.

GENERAL FEATURES

The range of biological rhythms

Rhythmic variations in living systems are commonplace and span a very wide gamut of intervals. Among those possessing short periods and hence a high frequency of recurrence are wing beats of insects (20 to 2,000 cycles per second), brain waves (1 to 60 c/sec), walking and chewing rhythms (1 to 8 c/sec), heartbeats (20 to 1,000 c/min), and respiratory rhythms (4 to 250 c/min).

Other biological rhythms have only a few cycles per day. Examples are the variations in activity of very young infants, in depth of sleep of adults, in metabolism of chick embryos, in spiralling of growing plant tips, and in a number of bodily processes of the human during certain illness. Longer periods include the well-known several-day recurrences of fever in malaria and the few-day to several-month reproductive rhythms of mammals. Still lower frequency rhythms include the several- to many-year cycles of reproduction in the periodic cicadas, of abundance in animal populations, of prosperity in business, and of outbreaks of wars.

Fluctuations in living creatures occur at every level from the cell to activities of the organism as a whole and are evident in altered chemistry, physiology, and behaviour, including changing responsiveness to all factors of the environment.

Periodic biological phenomena fall into two more or less distinct different categories: those correlated with periodic changes of the planet (geophysical correlates) and those having no such correlates. Biological rhythms correlated with ocean tides, days, months, and years clearly relate the activities of organisms to the fluctuations of their physical environment as a consequence of the movements of the earth. These rhythms possess properties that clearly set them apart from rhythms without external correlates, such as heartbeat and respiratory rate. The frequencies of these latter rhythms can be altered in response to changes in the immediate environment. Their frequencies also are altered readily by drugs and other chemicals and, as with metabolic rate itself, altered greatly as body-temperature changes. On the contrary, rhythmic variations with geophysical correlates have relatively fixed periods and are resistant to alteration in their frequency either by temperature or chemicals.

This extraordinary stability of the periods of the biological rhythms with geophysical correlates is not dependent upon direct responses to light and temperature rhythms. It was demonstrated more than two centuries ago that the capacity of plants to display their daily rhythm in sleep movements, the daytime raising and nighttime lowering of their leaves, persisted even when the plants were shielded from light changes and maintained in a relatively constant temperature. More recently, daily rhythms in a wide variety of living things, from single-celled forms to mammals and flowering plants, have been shown similarly to continue under constant conditions of light and temperature. Tidal rhythms in many shore organisms persist even after the organism is removed from the tidal environment. Comparably, semimonthly, monthly, and annual rhythmic variations continue even when organisms possessing them are deprived of every obvious periodic cue. Such organisms behave as if they possess the ability independently to time all Earth's major natural periods.

Adaptive significance of periodicity

A timing capacity, geared to the rhythmic variations of the natural environment, is indeed a most useful attribute. A time sense enables organisms to anticipate and exploit the most favourable portions of the day, tide, month, etc. It can be used by members of a species to synchronize their urge to breed. A time sense may serve, in conjunction with the sun, moon, and stars, as a means of geographic orientation for the migratory and homing activities of many organisms, just as a chronometer and a sextant enable a navigator to find his position or to plot a course.

BIOLOGICAL RHYTHMS AND NATURAL GEOPHYSICAL CYCLES

Living things are extraordinarily well adapted to their rhythmic environments and have become periodic in diverse aspects of their physiology and behaviour. The biological rhythms that have resulted are in part direct responses to the environment and in part indirect ones via the complex clock-timed organization.

Solar-day rhythms. Twenty-four-hour rhythms of organisms in nature are the rule; they are variously called solar day, circadian, diel, daily, diurnal, and nycthemeral. The evolution of life on earth has proceeded in the direction of maximal utilization of every habitable niche, with species specialized to deal with virtually every kind of environmental site. This specialization is related not only to space but also to time. Some organisms have become specialized for the darker, cooler, more humid nighttime; others for the lighter, warmer, drier daytime; and still others for the transitional twilight periods. The environment is in use full-time, but for maximal efficiency organisms are organized to work largely in "shifts." The night shift includes animals such as mice, cockroaches, scorpions, moths, owls, bats, skunks, opossums, and raccoon; the day shift includes songbirds, hawks, lizards, butterflies, and honeybees. Green plants alternate between the daytime use of sunlight in photosynthesis and the nighttime emphasis on growth and assimilation. Many plants undergo a daily sleep rhythm, their leaves drooping at night and rising by day. Some plants open their blossoms only at night, others only by day, with times commonly synchronized to the activity schedules of animals that pollinate them.

In each organism's daily pattern, a multiplicity of supporting chemical and physiological variations occur as the organism's whole being is thrown periodically into activity. In man, as an example, the daily alternation of sleep and wakefulness is accompanied by many changes, including activities of the nervous and endocrine systems and the liver and kidneys. Daily variations occur in body temperature, in heart and respiratory rates, and in pressure and composition of the blood. There are daily variations, also, in chemical syntheses and in cell divisions. Such rhythmically recurring phenomena are each set to the most efficient time of day for the organism and collectively contribute to the characteristic 24-hour phase map, or activity record, of that individual (Figure 1).

Daily variations in responsiveness

One of the very striking manners in which the presence of fundamental 24-hour variations is reflected in man and other animals is in changing reponsiveness or susceptibility to physical and chemical insults. The following are a few examples. Doses of X-rays that at one time of day can kill every individual exposed will at another time kill only a few individuals or none. Doses of powerful poisons, fully lethal at one time of day, leave animals unharmed at another. Noise that can induce fatal convulsions in mice at one time of day fail to do so at another. The ability of histamines to induce inflammatory response, and the effectiveness of alcohol, insulin, and even aspirin vary systematically with time of day. In insects the degree of resistance to insecticides or to exposures to high temperatures also depends upon the time of day.

Lunar tidal rhythms. Another major environmental period is the lunar tidal rhythm, the regular ebb and flow of the ocean tides, a cycle that subjects seashore plants and animals to a rhythmic change with typically two high and two low tides occurring each lunar day of about 24

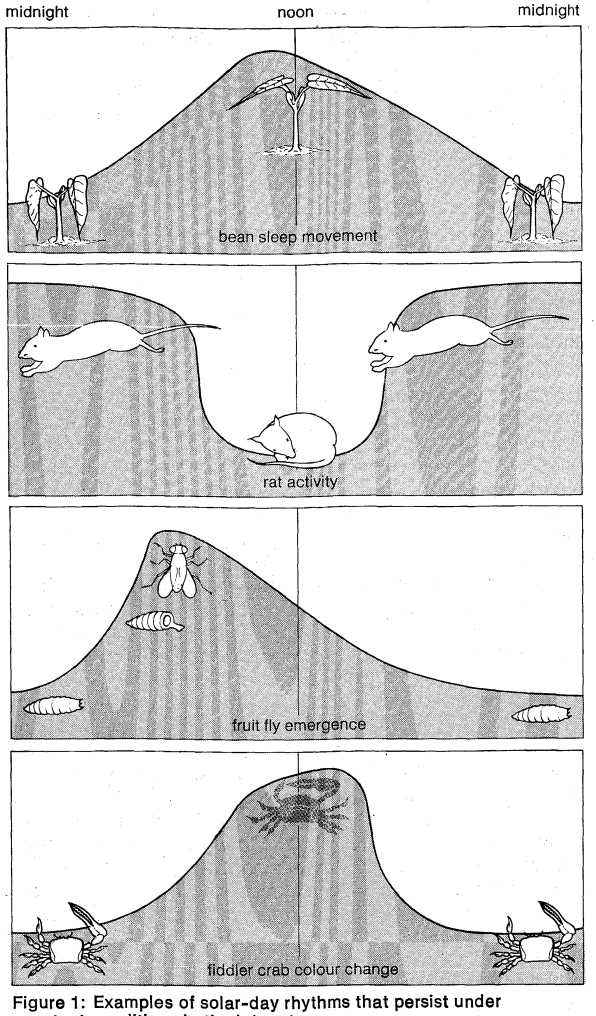


Figure 1: Examples of solar-day rhythms that persist under constant conditions in the laboratory.
Adapted from F.A. Brown, "Living Clocks," *Science*, vol. 130, pp. 1535–44 (December 4, 1959), in R.L. Smith, *Ecology and Field Biology* (1966); Harper & Row

hours' 50 minutes' duration. Some intertidal organisms, such as barnacles, green crabs, snails, clams, and oysters, are most active when submerged by the rising tide. Others such as shorebirds and fiddler crabs, are especially adapted to feed on beaches exposed at ebb tide. Some single-celled algae called diatoms move down into the sand and mud when the beaches are submerged and rise onto the surface to permit photosynthesis when the tide recedes. The diatoms and many of the shorebirds are active only when the beaches are exposed during daylight. Such organisms, therefore, display simultaneously 24-hour and tidal rhythms, which steadily undergo changing phase relations with respect to one another, coming into the same relationships exactly once each half-month. Indeed, all the organisms of the intertidal region must deal with both the daily light and temperature changes and the tidal ones. Fiddler crabs, for example, show concurrently a daily rhythm of skin colour change adaptively set to the times of day and night and a tidal rhythm of locomotor activity adjusted to the local tidal schedule.

Although the ocean tides are affected chiefly by the moon, the gravitational attraction of the sun also has an influence. The sun and moon "cooperate" at new and full moons to produce the greatest tidal ranges of each month, the so-called spring tides. At the moon's quarters the moon and sun maximally oppose one another, resulting in the smallest tidal ranges. As a consequence, inhabitants of the seashore that occupy the highest and the lowest levels of the intertidal zone experience periodic tidal submergences or exposures only for brief intervals once each half-month. One species of marine snail that inhabits the upper level of the beaches has been reported to show a semimonthly activity rhythm.

Any living creature that possesses both tidal and daily variations will exhibit day-by-day patterns of variation that systematically change from one day to the next as the peaks and troughs in the tidal cycles, like the tides themselves, occur at progressively later times of day. The specific pattern of organismic variation through any given day will repeat itself exactly a half month later.

Animal navigation

An interesting expression of these rhythmicities is evident in the reported phenomenon of celestial navigation by animals. Such animals as birds, fishes, turtles, spiders, insects, and crustaceans have been variously described to navigate or to "home" by using the relative positions of the sun or the moon or, in the case of birds, even the constellations as geographic navigational guides. Such sun-compass, moon-compass, or star-compass navigation requires that the animals be able systematically to alter their orientational angle at rates to compensate exactly for the earth's rotation relative to each of these celestial references.

Monthly rhythms. The synodic month, averaging 29.5 days, and the semimonth can be sensed by organisms, as noted earlier, through "beats" between their lunar tidal and circadian rhythms. Another kind of environmental monthly variation involves nighttime moonlight. Synodic monthly variations occur in such unlikely phenomena as rat and hamster running activity, light responsiveness of guppies, human colour vision, potato metabolic rate, and light responses of freshwater flatworms, even when shielded from all moonlight.

Among the biological activities linked to the synodic month are reproductive cycles of many different kinds of marine animals and plants. Best known and most studied of these are the monthly reproductive rhythms of palolo worms of the southwest Pacific, the fireworms of Bermuda, the grunions of the California coast, and the brown marine algae *Dictyota*. These rhythms are sufficiently precise in their relation to moon phase and season of the year as to be predictable. The changing relations of the solar-day and lunar-day physical rhythms appear to time these breeding cycles, since the phase of moon at which, for example, *Dictyota* breeds is related to the times of occurrence within the lunar day of the local high tides, which differ from place to place. The grunion's breeding is timed to within a few minutes of high tide on the local beach on the nights of its breeding, and the breeding times of palolo worms and fireworms are very accurately timed to within a few minutes in the day.

Precision timing of sea lily reproduction

The most spectacular precision in timing of a month- and year-related reproductive rhythm has been described for a deep-sea sea lily near Japan. This echinoderm liberates its sex cells once each year in October at about 3 PM on the day of one of the moon's quarters. In succeeding years the time of sex cell release changes, among the moon's two quarters, first–third–first, to progressively slightly earlier dates in October. The triplets are repeated until about the first of the month whereupon the following year it jumps abruptly to near the end of the month to start the advancing triplet progression again. The result is an 18-year cycle, which is essentially the period of regression of the moon's orbital plane. There have been reports of monthly menstrual cycles in some Primates with initiation of menstruation linked to the new moon. Reproductive swarming of mayflies has also been related to moon phase.

Little is known about the timing of such cycles as the human menstrual cycle, the approximately semimonthly estrus cycles of sheep, the three-week cycles of cattle and pigs, and the six-month cycles of dogs. Three kinds of hypotheses have been offered for their timing: (1) they are timed by fully independent periodic oscillations within the organism; (2) they are dependent for their timing on periodic interference between 24-hour and lunar-day clocks, since most of them appear to be simple multiples of a quarter month; and (3) they depend for timing on monthly variations in moonlight (directed chiefly at the primate menstrual cycle). Some recent studies have given strong support for a long-held belief that the menstrual period is in some manner regulated by the moon ("menses" means lunar month). In a study of a number of women with variable onset of menstrual periods, artificial

illumination of the bedroom through the 14th to 17th nights following the onset of menstruation resulted in a regularization of the period, with the period length coming very close to 29.5 days, the natural synodic month. That this period is a biologically significant one for the human species is further suggested by the fact that the average duration of pregnancy (from ovulation to birth) in the human is rather precisely nine 29.53-day synodic months.

Associated with the human menstrual cycle are related variations in almost every measurable bodily factor: temperature, blood chemistry, weight, heart rate, blood-cell counts, and pain thresholds are among the described variations. These factors pass through maxima and minima at different times but in such characteristic time relationships to one another that phase maps for the menstrual rhythm, comparable to those for the circadian, can be determined.

Annual rhythm. Another major rhythmic period for living things is the year. In most parts of the world substantial seasonal changes in light and temperature are correlated with rhythmic biological phenomena predominantly related to activity, reproduction, and growth. For example, in higher plants in the temperate zones, growth, flowering, and seed production typically occur over the course of the warmer months. In the oceans and fresh waters there are annual rhythms in plankton abundance. Annual reproductive rhythms in animals are typically set to the time of year that provides the optimum opportunity for birth and rearing of young. Annual rhythms include also hibernation to bridge the coldest period of the year and estivation to avoid hot-summer droughts.

The readily observable systematic annual variations in the activities of the plants and animals are accompanied by innumerable less obvious changes within their bodies, variations that pass through their peaks and depressions with such regularity that annual phase maps can be made.

Seasonal variations in human biology

Seasonal rhythms occur in man and are evident in such phenomena as birthrate and weight and in growth rate of children. There are annual variations also in innumerable other factors, among them body temperature, heart rate, basal metabolic rate, blood-cell counts, and blood chemistry. Not surprising, also, are reports of annual variations for many noninfectious diseases of the skin, nervous, circulatory, respiratory, and gastrointestinal systems, and even for mortality. Many infectious diseases of man are well known to exhibit peaks at characteristic times of year, while diseases without such evident annual fluctuations could perhaps be expected to display an annual variation in degrees of complications and mortality as a consequence of a probable annual variation in susceptibility related to the rhythmically changing biochemical and physiological states.

Too little investigated for man but difficult to treat because of the long period involved is the degree of phase shifting possible for the annual rhythms by manipulation of such physical factors as photoperiod and temperature.

Epochal rhythms. There are innumerable natural geophysical periods both shorter and longer than a year resulting from periodic activities in the sun and from relative movements of the earth, sun, moon, planets, and stars. Among the well-known shorter periods are the sun-generated ten-cycles per second electromagnetic fields, suggestively similar to the ten-cycles per second alpha rhythm of the human brain. Variations in climate, rainfall, temperature, and light have been claimed to be correlated with many longer periods. Among the periods, and perhaps the best known, are the 11- and 22-year sunspot cycles, the 18.6-year period of lunar "wobble" (nutation), and the 26,000-year precession of the earth's rotational axis. Beyond the solar system are the fluctuating galactic fields. Probably the longest physical period that has been suggested is the approximately 250,000,000-year period of rotation of the galaxy, which some believe to be correlated with terrestrial glacial periods. All these periods are accompanied by subtle alterations in the gravitational and electromagnetic fields of the terrestrial environment and so influence all living things.

Numerous several- to many-year biological periodicities involving a wide variety of phenomena have been described. Perhaps most familiar are the three–four-year cycles of lemming migrations and the eight–ten-year cycles in abundance of the Canadian lynx, varying hare, and ruffed grouse. Still other biological periodicities described from tree rings, shell deposition in mollusks, fossil depositions, business, and wars have been claimed. Very commonly reported are cycles of about 9, 11, 18, and 22 years. Although it seems plausible that such geophysically correlated biological periodicities are real, their long-period durations, together with their cycle irregularities, tend to make difficult the critical establishment of their exact lengths or causal associations with environment periodicities.

THE BIOLOGICAL CLOCK

Natural time. Many organisms, man included, when removed from their natural surroundings and placed in a laboratory under constant conditions, can often repeat a rhythmic pattern that they had experienced before such isolation. It is concluded from such behaviour that the organism can time the length of the natural day and even determine the points and measure durations of relatively detailed periods within a single day. This timing capacity is essentially uninfluenced by temperature changes and all ordinary drugs and chemicals. Every living system studied behaves as if it contains a highly dependable clock. The specialized properties of this clock suggest that it involves some unique biological mechanism indissociable from life itself; however, the clock can be investigated only indirectly, through observing the more conventional biochemical, physiological, and behavioural processes timed by it.

Biological clocks that time the solar day, the period of the earth's rotation relative to the sun, have been the most extensively studied. Except possibly for the bacteria, the same kind of clock exists in members of every major group of living creatures. Single-celled animals and plants appear to possess just as accurate and effective clocks as do the most highly differentiated multicellular organisms.

A number of general properties of the clock-timed solar-day rhythms have been discovered. One such property is that under constant conditions in the laboratory the rhythms usually change slightly, becoming instead a little longer or shorter than 24 hours. This commonly observed deviation from the 24-hour period under these conditions led to the concept of circadian (about a day) rhythms. (The explanation for such variance is discussed below.) In day-active (diurnal) vertebrates circadian rhythms, in constant light and temperature, are typically shorter than 24 hours, while in night-active (nocturnal) ones they are typically longer. The periods of circadian rhythms under these conditions rarely exceed the range of 20 to 28 hours and usually are within an hour or so of 24 hours.

The basic circadian rhythm

When an organism is returned to its normal 24-hour light–dark environment, it soon resumes an accurate solar-day rhythm. The transitions from light to darkness and darkness to light appear to be the signals for readjusting the organism's daily physiological cycles. A nocturnal mammal, for example, adjusts its time of activity to the time of darkness, a diurnal songbird to the time of daylight. Such resetting or phase-adjusting factors for the rhythms are termed daily clues, phase setters, or synchronizers.

Involved in adjusting an organism's daily activity pattern to the appropriate time of day is a rhythmic variation in the influence of the light as a resetting factor. This variation has been termed a response curve. A light change near the time of the beginning of an organism's activity pattern advances slightly the phases of the cycle to an earlier time of day; a comparable light change near the end, on the other hand, extends it. At other times during the cycle comparable light changes have little or no phase-shifting influence. Although response curves appear to possess the same general form regardless of the kind of organism, each individual possesses its own detailed form of response curve, which adjusts its activity pattern to the light–dark cycles in its own characteristic manner.

Whereas light appears to be the dominant phase setter in adjusting clock-timed rhythms to natural 24-hour cycles,

temperature is also effective. Many organisms are able to adjust their activity patterns to 24-hour rhythms in temperature in the absence of light variations. Still other phase setters have been reported, including sound, feeding time, and social interactions.

Modifiable clock-timed rhythms. The circadian rhythms that continue under experimental conditions of constant light and temperature are termed free running; they reflect the underlying natural rhythms of an organism. The free-running period can usually be altered, but only very slightly, by changing the intensity of the illumination or the temperature. A 10° C temperature increase seldom changes a circadian period more than 10 percent, and in most organisms the period is shortened. For fiddler crabs in constant darkness the free-running period, of exactly 24 hours, remains unchanged over a 20° C range. In contrast to these relatively resistant rhythms, the rates for ordinary biochemical or physiological processes typically double or triple for each 10° C rise in temperature. Free-running rhythms sometimes cease spontaneously, only to resume at a later time as if the rhythms had been continuing undetected during the intervening time. In other words, a clock-timed phenomenon within the organism may become uncoupled from the underlying clock. Circadian rhythms sometimes fade out over the course of a few cycles under constant conditions; in other cases they appear to continue indefinitely, sometimes even with increasing amplitude and sharpness.

Free running rhythms

Although chemical inhibitors or stimulants of general metabolic processes can influence rhythmic processes, they appear not to affect the underlying clock in any way. Some powerful metabolic inhibitors such as cyanide will eliminate all overt rhythmic variation in an organism, but upon removal of the inhibitor, the rhythm returns as if it had been uninterrupted. On the other hand, exposure to near-freezing temperatures inhibits rhythmic changes in such organisms as fiddler crabs and cockroaches; but when the temperature is elevated again, the restored rhythms have been phase shifted by the approximate duration of the time of chilling.

There appears to be a genetic component to circadian organization. Organisms reared from eggs in constant light and temperature may display circadian rhythms spontaneously or may be induced to do so by a single abrupt light or temperature change, a signal that itself conveys no information as to period length. Furthermore, bean plants selected and interbred for longer and for shorter circadian periods under a given set of constant environmental conditions will beget offspring with, respectively, longer and shorter circadian periods under these same experimental conditions. The length of the free-running period is clearly genetically determined in part.

The influence of heavy water, which is D_2O, on circadian rhythms is of particular interest. Treatment of organisms as different as algae and mice results in a lengthening of the free-running periods. Under a 24-hour, light–dark regime, heavy water appears to alter the phase relations of the activity patterns of organisms to the light cycles.

Clock-timed circadian rhythms have been observed during rapid geographic translocations of organisms eastward or westward by airplane to other time belts. Investigations of this type, performed with fiddler crabs and honeybees, confirmed that the organisms carry circadian rhythms still adjusted to the light cycles of their former location (Figure 2). To reset them to the new local time requires several days of exposure to the local phase setters. This phenomenon is quite familiar to persons who have travelled long distances rapidly by jet airplane. The resultant rhythmic dislocation and the need for gradual adjustment over two to ten days at the end of such a trip is often referred to as the "jet syndrome."

Time dislocation: The "jet syndrome"

It is well-known that among the innumerable periodic changes that underlie and support the overt rhythms of behaviour of organisms; the peak values occur in a characteristic sequence over the day. A description of these comprises what is called a phase map. Such a sequence and spacing reflects the order and temporal relationships of cause–effect in the normal interactions of the various

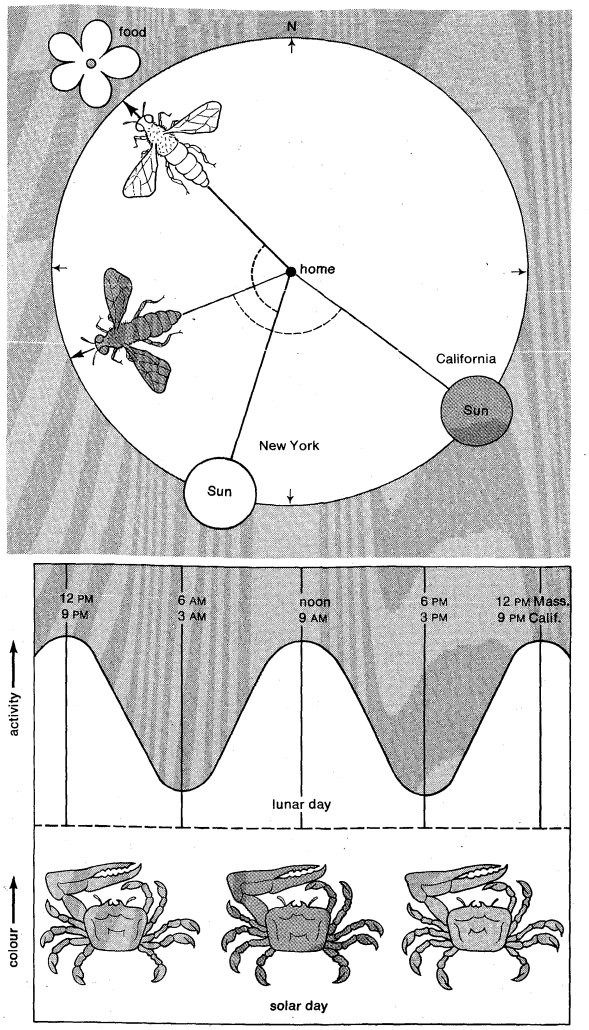


Figure 2: *The persistence of natural rhythms.*
(Above) Bees trained in New York to fly northwest to a feeding station at 1 PM fly southwest at 10 AM when transported to California, following the sun's position as learned in New York. (Below) Fiddler crabs, moved from Massachusetts to California and kept in unvarying illumination, show activity, rhythms, and colour changes identical to their East Coast counterparts, regardless of the local time.
From F.A. Brown, Jr., *Biological Clocks*, BSCS-Educational Programs Improvement Corp.

bodily processes. Phase maps may undergo transitory disruptions when an organism is compelled to make a rapid phase adjustment as, for example, after a rapid move to a new geographic longitude. Under such circumstances the various individual 24-hour components comprising the circadian phase map do not reset their phases to the new environmental times at the same rate. They may become somewhat displaced in their relations to one another. This dissociation of rhythmic components is probably the major cause of the fatigue and the lowered efficiency that characterizes the jet syndrome in human beings.

Another remarkable property of free-running rhythms is their capacity sometimes to maintain well beyond chance, and for extended periods, precise 24-hour periods and, in some cases, even to set their active period to their normal time of day without any obvious daily clues. The former has been described for fiddler crabs, gila monsters, mice, and kangaroo rats. The latter capacity has been observed in chicks and some lizards. These animals are apparently able to set their newly developed behavioral system to times of outside day and night by the use of subtle and as yet not fully identified signals.

The biological clock also times other natural environmental periods. Many plants and animals whose activities vary rhythmically with the tides have been found, when taken into the laboratory, to continue rather accurately their tidal rhythms. The period of these rhythms is about

12 hours 25 minutes, or 24 hours 50 minutes for the double-cycle, which is called a lunar day. The period of the tidal changes is correlated with the rotation of the earth relative to the moon. The tide, as does the moon, rises on the average of 50 minutes later each day.

The properties of the tidal, clock-timed rhythms differ in significant ways from the circadian ones. Although they possess the same high degree of temperature and drug independence of their period lengths, they will not synchronize to a 24-hour light–dark environment. Instead, tidal rhythms continue to move over the day at the tidal rate. This failure to follow the light cycles is obviously adaptive, since in nature the phase relationship of the tides to time of day systematically changes.

The clock-timed tidal rhythm can be adaptively adjusted to the tidal schedule of a specific beach. Fiddler crabs transported to a foreign beach with different tidal times gradually, over a few days, reset their tidal rhythms accordingly. The phase-resetting clues for the tidal rhythms remain unknown, though a number of factors such as mechanical wave action and periodic drying have been suggested.

Drifting of tidal rhythms

Clock-timed tidal rhythms commonly exhibit a tendency, under constant conditions of the laboratory, to drift away from their tidal setting until they approximate the actual time of lunar day. Oysters and fiddler crabs taken from their home beaches and kept in the laboratory gradually, during a week or so, reset their tidal rhythms to display their maximum activity at the times of the upper and lower transits of the moon, as if some subtile synchronizers associated with the lunar tides of the atmosphere were now replacing the unknown local tidal ones.

There are clocked-timed synodic monthly rhythms of 29.5 days, the period from one new moon to the next. This is the period produced by "beats" between the solar days and lunar days—in other words, the interval between the times when the solar day and lunar day begin together. The moon rises close to the same time of day once each synodic month. The organismic clock-timed monthly rhythms in some instances appear to depend upon a comparable interaction between tidal or lunar-day rhythms and solar-day ones. Interactions between 12.4-hour tidal rhythms and 24-hour circadian rhythms would yield semimonthly "beats," and between 24.8 hour double tidal cycles and 24-hour ones monthly "beats." Evidence for such clock-timed rhythms is seen in the occurrence of semimonthly colour changes in crabs kept under constant conditions. These exhibit phase differences among populations in their semimonthly rhythms accountable in terms of the differing phase relationships between their local tides and time of day.

The clock that times monthly rhythms appears to be just as dependable as those timing circadian and tidal rhythms. Monthly rhythms can, comparably, be phase shifted relative to the actual moon phases. In addition to the phase setting by changes in some factors associated with local tidal movement, moonlight variations through the month have been suggested as a phase-setting factor for semimonthly and monthly rhythms, particularly the breeding rhythms of sea worms and the menstrual cycle for Primates, including women.

Clock-timed annual rhythms, persisting in unvarying regimes of illumination and temperature and every other obvious clue as to the yearly period, have been reported for widely different kinds of organisms and phenomena. An alga, on the one hand, and the laboratory rat, on the other, have displayed annual variations in enzymatic activities. Seeds show annual variations in capacity to sprout, and newly germinated seedlings exhibit annual variations in their metabolic rate. Some birds show rather precise annual, and others *circa*-annual, reproductive cycles. Food intake of woodchucks has been reported to vary systematically with time of year. As for the other periods, the clock that times annual rhythms appears to be independent of temperature and uninfluenced by all attempts to alter its period chemically. The normal patterns of annual variation of many plants and animals can, however, be altered in response to changes in relative lengths of light and darkness (photoperiodism), much in the manner analogous to the phase resetting of the clock-timed circadian rhythm by times of light onset and termination. On the other hand, annual rhythms such as that of woodchuck food intake appear to be unalterable in phase by manipulation of either temperature or light.

Persistent clock-timed rhythms. There is another class of periodicities in organisms, rhythms that persist under the most carefully controlled laboratory conditions but that possess properties different from those described in the preceding account. These have precise solar-day, lunar-day, monthly, and annual periods. These have been discovered for respiratory or activity rates in a wide variety of animals and plants. The activity of an organism, other factors equal, varies systematically with local time of day, time of lunar day, time of month, and time of year. The potato, for example, shows a clear-cut sidereal-day (23-hour 56-minute period of rotation of the earth relative to the stars) variation in its metabolic processes. These periodisms can not be reset or phase shifted. Moreover, there is no evidence to suggest that they are adaptive. The range of their variations is least for the sidereal day and most for the annual period.

Fluctuating activity. Aperiodic, hour-by-hour fluctuations in metabolic rate or activity of organisms under presumably constant conditions continue to show correlations with weather changes, indicating beyond reasonable doubt that changing physical factors somehow continue to reach and influence the organisms, even in their experimental isolation. How do these external factors pervade the controlled environments of the laboratory? The answer perhaps may be found in a very highly specialized sensitivity of organisms to weak geoelectromagnetic fields that pervade the universe. Aperiodic biological fluctuations do coincide with fluctuations in such factors as background radiation, terrestrial magnetism, and primary cosmic radiation.

Influence of magnetism and radiation

The exogenous derivation (external to the organism) of these variations is indicated by the universality of response: the rhythmic patterns of these geophysically dependent fluctuations tend to show, among organisms spanning the gamut of kinds of living creatures, either the same general form or mirror images of parts or the whole of them, suggesting that all the organisms are responding to the same specific geophysical changes and that they have freedom only to determine the sign, positive or negative, of their responses to them.

The nature of the biological clock. As more information becomes available, it seems clearer that there are two kinds of rhythms, one resulting from response to the subtle geophysical environment, the other being rhythmic organismic responses to more obvious changes.

The timer for the first of these kinds of rhythms obviously lies outside the organism, and hence these rhythms are termed geophysically dependent. The detailed character of the biological cyclic pattern, however, is just as obviously determined by the nature of the biological response. The widespread general uniformity of the pattern for all organisms suggests that the response is generated at a very fundamental level of cellular organization. It must still depend upon the organism as an organized continuum whose composition and reactivity at any given point in time are related to past and future. The organism is, therefore, not a fully passive system in the generation of its exogenous metabolic periodisms. It may be an effective stimulus-filtering system.

The clock that times the overt, adaptive circadian, tidal, monthly, and annual rhythms is far less evident. Two kinds of hypotheses have been advanced. One, the internal-timer hypothesis, is that living systems can oscillate, fully independent of all environmental periodisms with close to, but not exactly, all major geophysical frequencies. Free-running rhythms are assumed to reflect directly the periods of the somewhat inaccurate internal clocks. Exact organismic solar-day, tidal, monthly, and annual periods are assumed to depend upon continuing corrections of the poorly accurate clocks by the environmental light, temperature, and tidal rhythms. The rates at which the clocks run are believed to be speeded or slowed to a small degree by changed levels of temperature and illumi-

The internal-timer hypothesis

nation as well as by certain chemical factors. The clock periods are assumed to be inherited, the approximation to the geophysical periods having resulted from natural selection.

By describing the rhythms as oscillations resembling in properties those of physical systems that become entrained or synchronized to other slightly different imposed frequencies and treating them mathematically as such, surprisingly satisfactory descriptions of the observed dynamics of biological rhythms and phase setting of the rhythms can be obtained, useful predictions made, and new experiments suggested. Despite intensive research, progress has not yet been made in elucidating an internal oscillator system with the clock properties. The problem is compounded by the inability to differentiate between a biochemical or biophysical reaction that is in reality a clock element and one that is merely timed by the clock.

The external-timer hypothesis

Another hypothesis is that living things have evolved the capacity to use the subtle rhythmic input of geophysical periods to time their own processes to the most adaptive rhythmic patterns. In terms of this hypothesis the clocks are not inaccurate but quite precise (24-hour, 24.8-hour, 29.5-day, and 365¼-day clocks, for example). As long as the organism remains in its natural rhythmic environment, the clocks and the environmental rhythms of light, temperature, and tides have the same frequencies. When, however, the organism is placed under constant conditions, its circadian rhythm gradually drifts off phase. There will be phase advances alternating with phase delays in a cyclic manner as the constant light and temperature react at different points along the response curve. The recurring patterns will gradually shift to earlier hours if the advancing response is stronger than the delaying one or shift toward later hours if the delaying one is the stronger. This corresponds well with the observation that circadian periods deviate slightly in practice from 24 hours.

The self-setting to specific clock hour of solar day or lunar day or time of month or of year sometimes observed in constant conditions, as well as the very existence of the geophysically dependent periodisms, would find consistent explanation. The accuracy of the clocks during rapid east–west translocations would require either influence of a universal time-varying geophysical factor, phase-synchronized at all longitudes, as, for example, ionospheric charge, or a capacity to rephase its own rhythmic pattern in relation to a rapidly shifted exogenous local-time cycle effected by the metabolic inertia inherent in a continuing functional system.

In brief, some of the properties of biological periodicities definitely require for their explanation that timing information continuously flows inward from the environment, even in what have commonly been deemed constant conditions. Many other phenomena are equally compatible with either hypothesis, internal or external timing. Possibly both kinds of clocks are utilized by organisms. An environmentally dependent clock element may be responsible for the remarkable resistance to timing interference by temperature and chemical agents and for the long-term, relatively precise persistence of daily rhythms and particularly for the stability of the longer period monthly and annual ones. An environmentally independent clock element, on the other hand, may be essential for bridging shorter intervals of time within cycles but with perhaps a precision by itself inadequate to account for the observed regularity and dependability of clock-timed rhythms.

FACTORS AFFECTING BIOLOGICAL PERIODICITIES

The rhythmic patterns of variation of organisms differ from one species to another and also from one individual to another, even in response to essentially the same external conditions. The nervous and endocrine systems in interaction with the parts of the body that are coordinated and integrated by them contribute to these patterns. Under ordinary circumstances, activities comprising the patterns are arranged to meet an organism's demands anticipated over the period. Pathological or otherwise induced alterations in the organism alter the cycle forms. The periods, following those of the physical environment, usually remain unchanged. Reported, however, have been instances in blinded mice of circadian periods becoming free-running ones differing from 24 hours, resembling the state of affairs commonly seen when comparable mice are held in continuous darkness. Light, operating through the eyes, would appear in such instances to be the essential phase-setting agent.

The role played by the animals' total functional complex in determining the cycle form has been illustrated in respiratory patterns of unhatched chick embryos and in the activity of human infants. In conditions of constant temperature and light, chick embryos up to six days old show a daily activity variation with several peaks each day. By the seventh day, when they have just completed the development of a functioning sensory-neuro-motor system, the daily variation changes to one with a broad daytime higher activity and lower nighttime one. Well known, too, is the change in sleep-activity rhythms of infants. Up to about 15 weeks there are several cycles a day, but afterward the cycle simplifies to essentially a single longer period of wakefulness by day and sleep by night. It might be argued that the infant simply adopts the social pattern of the household, but in the chick the change in pattern with time seems clearly genetically determined.

The sleep-activity cycle of infants

It was mentioned earlier that evolution, involving natural selection, has suited species to fit each niche at a particular time within a cycle. In the continuing course of events a species is usually rigorously held to its rhythmic schedule by the competition of other species present at other times. It is an interesting observation, however, that some large day-active game animals in Africa, whose populations had been decimated by hunters, altered their natural patterns to become night active but resumed diurnal habits when their numbers again had increased. It is likely that comparable adaptive phenomena may be more widespread in nature than has been reported.

Man in his cultural evolution has greatly modified his patterns of rhythmic behaviour. The eight-hour occupational period accurately clock timed to particular hours of the day provides its imprint. There is evidence of diverse kinds that an 8-hour work schedule at one time of day is not exactly like that of another time. The efficiency of man varies with a characteristic 24-hour pattern, performance being best from 7 to 11 AM and worst at 3 AM (Figure 3). Other daily rhythmic factors concurrently operate.

From J. Hardy, *Physiological Problems in Space Exploration* (1964); Charles C. Thomas, Publisher

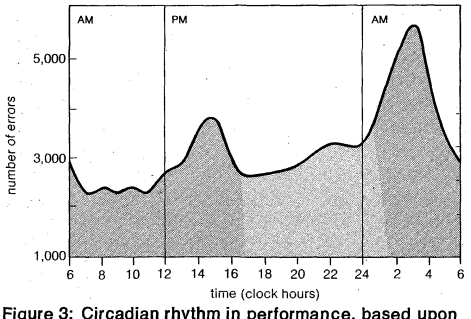


Figure 3: Circadian rhythm in performance, based upon the number of errors made in the routine checking of gas meters at different clock hours.

The week, with its day or two of rest or altered scheduling, obviously imposes a weekly rhythm upon man. It has been suggested that the week is a natural period for him, as also for other organisms as well, perhaps related to quarterly lunar phases.

Man is increasingly modifying his environment, from the conditioning of his living and working quarters through more extensive changes associated with life in submarines, sea labs, and space vehicles. At a fearsome rate, technological man is altering the earth's atmosphere physically and chemically with consequent modification of illumination, temperature, and a host of other periodic parameters. One of the vast areas of concern for man and for all living creatures is what such changes might ultimately do to the periodicities that characterize life on this

planet and whose critical regulatory roles are only beginning to be appreciated.

BIBLIOGRAPHY. "Biological Clocks," *Cold Spring Harbor Sympos. Quant. Biol.*, vol. 25 (1960); "Rhythmic Functions in the Living System," *Ann. N.Y. Acad. Sci.*, vol. 98 (1962); and J. ASCHOFF (ed.), *Circadian Clocks* (1965), comprise the published proceedings of three broadly oriented international symposia on biological rhythmicity participated in by the leading investigators of the subject. F.A. BROWN, JR., *Biological Clocks* (1962), is a brief, elementary account of the nature of, and problems related to, the phenomena of plant and animal periodicities. More advanced and detailed general accounts of bio-rhythmicity are: E. BUNNING, *The Physiological Clock*, rev. 2nd ed. (1967), particularly good for the earlier literature and classical views; J.L. CLOUDSLEY-THOMPSON, *Rhythmic Activity in Animal Physiology and Behaviour* (1961), best for its treatment of ecological significances; J.E. HARKER, *The Physiology of Diurnal Rhythms* (1964), emphasizing the periodicities associated with day–night; and B.M. SWEENEY, *Rhythmic Phenomena in Plants* (1969). In a class by itself is A. SOLLBERGER, *Biological Rhythm Research* (1965), a comprehensive, well-indexed, and documented reference work containing more than 2,500 literature citations, well-balanced in distribution over the subject. F.A. BROWN, JR., J.W. HASTINGS, and J.D. PALMER, *The Biological Clock: Two Views* (1970), is a succinct presentation of the two currently held alternative hypotheses, internal and external timing of the rhythms, together with the evidence upon which they are based. G.G. LUCE, *Biological Rhythms in Psychiatry and Medicine* (1970), is a readable, informative, extensively documented account of occurrence and significances of daily, lunar, and annual periodicities for man.

(F.A.B.)

Periodic Law

Dmitry I. Mendeleyev discovered in the mid-19th century that the chemical elements show a periodic recurrence of properties when they are arranged in a certain order, which is approximately the order of increasing atomic weight. The statement of this fact, which has been of inestimable value in the development of the science of chemistry, is called the periodic law; a tabular arrangement of the elements that brings those with similar properties into juxtaposition is called the periodic table or periodic system of the elements.

It was recognized during the second decade of the 20th century that the order of elements in the periodic system is that of their atomic numbers, the integers which are equal to the positive electrical charges of the atomic nuclei expressed in electronic units; and in subsequent years great progress was made in explaining the periodic law in terms of the electronic structure of atoms and molecules. This clarification has increased the value of the law, which is used as much today as it was at the beginning of the 20th century, when it expressed the only known relationship among the elements.

History of the periodic law. The classification of the chemical elements is a comparatively recent problem, for it was not until about the beginning of the 19th century that chemists clearly recognized that the chemical elements represent unique, chemically indivisible species, which appear in various combined forms as chemical compounds. The early years of the century witnessed a rapid development in analytical chemistry—the art of distinguishing different chemical substances—and the consequent building up of a vast body of knowledge of the chemical and physical properties of both elements and compounds. This rapid expansion of chemical knowledge soon necessitated classification, for on the classification of chemical knowledge are based not only the systematized literature of chemistry but also the laboratory arts by which chemistry is passed on as a living science from one generation of chemists to another. The elementary substances are few in number (105 now are known and physicists are hopeful that more will be discovered in high-energy experiments), with distinctive properties; whereas compound substances are extremely numerous, with properties correspondingly diversified and often overlapping. Relationships could be discerned more readily among the compounds than among the elements; it thus occurred that the classification of elements lagged many years behind that of compounds. In fact, no general agreement had been reached among chemists as to the classification of elements for nearly half a century after the systems of classification of compounds had become established in general use.

J.W. Döbereiner in 1817 showed that the combining weight, meaning atomic weight, of strontium lies midway between those of calcium and barium; and some years later he showed that other such "triads" exist (chlorine, bromine, and iodine [halogens] and lithium, sodium, and potassium [alkali metals]). J.-B.-A. Dumas, L. Gmelin, E. Lenssen, Max von Pettenkofer, and J.P. Cooke expanded Döbereiner's suggestions between 1827 and 1858 by showing that similar relationships extended further than the triads of elements, fluorine being added to the halogens, and magnesium to the alkaline-earth metals; while oxygen, sulfur, selenium, and tellurium were classed as one family, and nitrogen, phosphorus, arsenic, antimony, and bismuth as another family of elements. Triads of elements

Attempts were made subsequently to show that the atomic weights of the elements could be expressed by an arithmetic function, and in 1862 A.-E.-B. de Chancourtois proposed a classification of the elements based on the new values of atomic weights given by Stanislao Cannizzaro's system of 1858. De Chancourtois plotted the atomic weights on the surface of a cylinder with a circumference of 16 units, corresponding to the approximate atomic weight of oxygen. The resulting helical curve brought closely related elements onto corresponding points above or below one another on the cylinder, and he suggested in consequence that "the properties of the elements are the properties of numbers," a remarkable prediction in the light of modern knowledge.

Several years later, in 1865, J.A.R. Newlands proposed a system of classification of the elements in the order of increasing atomic weights, the elements being assigned ordinal numbers from unity upward and divided into seven groups having properties closely related to the first seven of the elements then known: hydrogen, lithium, beryllium, boron, carbon, nitrogen, and oxygen. This relationship was termed the law of octaves, by analogy with the seven intervals of the musical scale.

Then in 1869, as a result of an extensive correlation of the properties and the atomic weights of the elements, with special attention to valency (that is, the number of single bonds the element can form), Mendeleyev proposed the periodic law, by which "the elements arranged according to the magnitude of atomic weights show a periodic change of properties." Lothar Meyer had independently reached a similar conclusion, published after the appearance of Mendeleyev's paper in a Russian journal and an account of it in German.

Mendeleyev's periodic table of 1869 contained 17 columns, with two nearly complete periods (sequences) of elements, from potassium to bromine and rubidium to iodine (see Figures 2 and 3), preceded by two partial periods of seven elements each (lithium to fluorine and sodium to chlorine), and followed by three incomplete periods. In an 1871 paper Mendeleyev presented a revision of the 17-group table, the principal improvement being the correct repositioning of 17 elements. He, as well as Lothar Meyer, also proposed a table with eight columns obtained by splitting each of the long periods into a period of seven, an eighth group containing the three central elements (such as iron, cobalt, nickel; Mendeleyev also included copper, instead of placing it in Group I), and a second period of seven. The first and second periods of seven were later distinguished by use of the letters "a" and "b" attached to the group symbols, which were the Roman numerals. The first periodic table

With the discovery of the noble gases helium, neon, argon, krypton, radon, and xenon by Lord Rayleigh (John William Strutt) and Sir William Ramsay in 1894 and the following years, Mendeleyev and others proposed that a new "zero" group to accommodate them be added to the periodic table. The "short-period" form of the periodic table, Figure 1, with Groups 0, I, II, . . . VIII, became popular, and remained in general use until about 1930.

Based upon an earlier (1882) model of T. Bayley, J.

Group 0	I		II		III		IV		V		VI		VII		VIII
	a	b	a	b	a	b	a	b	a	b	a	b	a	b	
	H 1														
He 2	Li 3		Be 4		B 5			C 6		N 7		O 8		F 9	
Ne 10	Na 11		Mg 12		Al 13			Si 14		P 15		S 16		Cl 17	
Ar 18	K 19		Ca 20		Sc 21		Ti 22		V 23		Cr 24		Mn 25		Fe 26, Co 27, Ni 28
		Cu 29		Zn 30		Ga 31		Ge 32		As 33		Se 34		Br 35	
Kr 36	Rb 37		Sr 38		Y 39		Zr 40		Nb 41		Mo 42		Tc 43		Ru 44, Rh 45, Pd 46
		Ag 47		Cd 48		In 49		Sn 50		Sb 51		Te 52		I 53	
Xe 54	Cs 55		Ba 56		57–71*		Hf 72		Ta 73		W 74		Re 75		Os 76, Ir 77, Pt 78
		Au 79		Hg 80		Tl 81		Pb 82		Bi 83		Po 84		At 85	
Rn 86	Fr 87		Ra 88		Ac 89		Th 90		Pa 91		U 92		Np 93		Pu 94, Am 95, Cm 96

*Lanthanum and the lanthanons

Figure 1: Short-period form of periodic system of elements.

Thomsen in 1895 devised a new table. This was interpreted in terms of the electronic structure of atoms by Niels Bohr in 1922. In this table, Figure 2, there are periods of increasing length between the noble gases; the table thus contains a period of 2 elements, two periods of 8 elements, two periods of 18 elements, a period of 32 elements, and an incomplete period. The elements in each period may be connected by tie lines with one or more corresponding elements in the following period. The principal disadvantage of this periodic table is the large space required by the period of 32 elements and the difficulty of tracing a sequence of closely similar elements along the slanting tie lines. A very satisfactory compromise, shown in Figure 3, is to compress the period of 32 elements into 18 spaces by omitting the 14 lanthanons (lanthanides) and the 14 actinons (actinides) and listing them in a special double row below the other periods.

Alternate long forms of the periodic table have been proposed from time to time. One of the earliest of these, described by A. Werner in 1905, is notable because it divides each of the shorter periods into two parts, one at either end of the table over the elements in the longer periods that they most resemble. The multiple tie lines connecting the periods in the Bayley-type table are thereby dispensed with. This class of table, too, can be greatly simplified by removing the lanthanon and actinon elements to a separate area. By the mid-20th century this version of the table (Figure 4) had become the most commonly used.

Predictive value of the periodic law

The great value of the periodic law was made evident by Mendeleyev's success in 1871 in finding that the properties of 17 elements could be correlated with those of other elements by moving the 17 to new positions from those indicated by their atomic weights. This change indicated that there were small errors in the previously accepted atomic weights of several of the elements and large errors for several others, for which wrong multiples of the combining weights had been used as atomic weights (the combining weight being that weight of an element that combines with a given weight of a standard). Mendeleyev was also able to predict the existence, and many of the properties, of the then undiscovered elements ekaboron, ekaaluminum, and ekasilicon, now identified with the elements scandium, gallium, and germanium, respectively. Similarly, after the discovery of helium and argon, the periodic law permitted the prediction of the existence of neon, krypton, xenon, and radon. Moreover, Bohr pointed out that the missing element 72 would be expected, from its position in the periodic system, to be similar to zirconium in its properties rather than to the rare earths; this observation led G. de Hevesy and D. Coster in 1922 to examine zirconium ores and to discover the unknown element, which they named hafnium.

In spite of the corrections made by the redetermination of atomic weights, some of the elements in the Mendeleyev and Lothar Meyer periodic tables of 1871 were still required by their properties to be put in positions somewhat out of the order of atomic weights. Thus, with three pairs of elements—argon and potassium, cobalt and nickel, and tellurium and iodine—the first element of each pair has the greater atomic weight but the earlier position in the periodic system. The eventual solution to this difficulty was found only when the structure of the atom was better understood.

About 1910, Sir Ernest Rutherford's experiments on the scattering of alpha particles by the nuclei of heavy atoms led to the determination of the nuclear electrical charge. The ratio of the nuclear charge to that of the electron was noted to be roughly one-half the atomic weight. In 1911 A. van den Broek suggested that this quantity, the atomic number, might be identified with the ordinal number of the element in the periodic system (following the lead of Newlands, it had become customary to number the elements according to their position in the table). This suggestion was brilliantly confirmed in 1913 by H.G.J. Moseley's measurements of the wavelengths of the characteristic X-ray spectral lines of many elements, which showed that the wavelengths did indeed depend in a regular way upon the atomic numbers—identical with the ordinal numbers of the elements in the table. There is no longer any uncertainty about the position of any element in the ordered series of the periodic system. All of the elements from 1 to 105 are now known, and it is to be expected that new elements beyond 105 will be made by nuclear reactions.

That the exact atomic weight of an element is of small significance with regard to its position in the periodic system is shown by the existence of isotopes of every element, atoms with the same atomic number but with different atomic weights. The chemical properties of the different isotopes of an element are essentially the same, and all of the isotopes of an element occupy the same place in the periodic system in spite of their differences in atomic weight.

Detailed understanding of the periodic system has developed along with the quantum theory of spectra and the electronic structure of atoms, beginning with the work of Bohr in 1913. Important forward steps were the formulation of the general rules of the old quantum theory by W. Wilson and A. Sommerfeld in 1916, the discovery of the exclusion principle by W. Pauli, in 1925 (see ATOMIC STRUCTURE), the discovery of the spin of the electron by G.E. Uhlenbeck and S. Goudsmit in 1925 (see ELECTRON), and the development of quantum mechanics (see MECHANICS, QUANTUM) by W. Heisenberg and E. Schrödinger during the same year. The development of the electronic theory of valence and molecular structure (see MOLECULAR STRUCTURE), beginning with the postulate of the shared electron pair by G.N. Lewis in 1916, also played a very important part in explaining the peri-

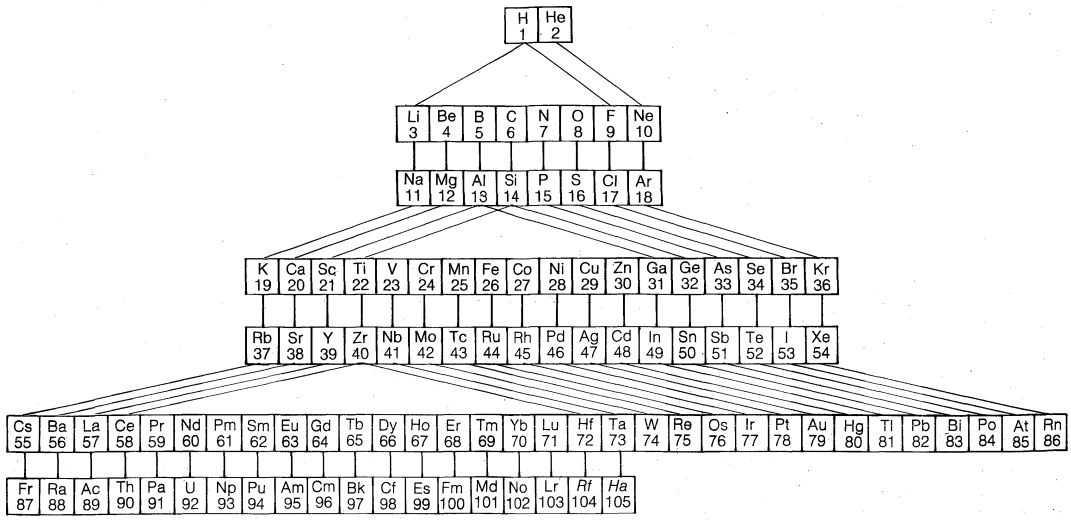


Figure 2: Periodic system of elements with periods demarcated by noble gases.

odic law (see CHEMICAL BONDING; and CHEMICAL REACTION, MECHANISMS OF).

The long form of the periodic table. The long-period form of the periodic system of the elements, presented here in two forms, Figures 3 and 4, contains all of the chemical elements that have been discovered or made; they are arranged, in the order of their atomic numbers, in seven horizontal periods, with the lanthanons (lanthanum, 57, to lutetium, 71) and the actinons (actinium, 89, to lawrencium, 103) indicated separately below (unless otherwise stated, Figure 3 will be used as reference). The periods are of varying lengths. First there is the hydrogen period, consisting of the two elements hydrogen, 1, and helium, 2. Then there are two periods of eight elements each: the first short period, from lithium, 3, to neon, 10; and the second short period, from sodium, 11, to argon, 18. There follow two periods of 18 elements each: the first long period, from potassium 19, to krypton, 36; and the second long period, from rubidium, 37, to xenon, 54. The very long period of 32 elements, from cesium, 55, to radon, 86, is condensed into 18 columns by the omission of the lanthanons (which are indicated separately below), permitting the remaining 18 elements, which are closely similar in their properties to corresponding elements of the first and second long periods, to be placed directly below these elements. The last period is incomplete. The properties of these elements indicate that this period is a second very long period of 32 elements, which would be completed at element 118.

Periods

The six noble gases, helium, neon, argon, krypton, xenon, and radon, occur at the ends of the six completed periods and constitute the 0 group of the periodic system. (It is customary to refer to horizontal series of elements in the table as periods and vertical series as groups.) The seven elements lithium to fluorine and the seven corresponding elements sodium to chlorine are placed, in Figure 3, in the seven groups, I, II, III, IV, V, VI, and VII, respectively. The 17 elements of the fourth period, from potassium, 19, to bromine, 35, are distinct in their properties and might well be considered to constitute 17 subgroups of the periodic system. Largely for historical reasons (the use of the short-period form of the periodic system,

Groups

								H 1									He 2
0	I	II	III	IV	V	VI	VII	0									
He 2	Li 3	Be 4	B 5	C 6	N 7	O 8	F 9	Ne 10									
Ne 10	Na 11	Mg 12	Al 13	Si 14	P 15	S 16	Cl 17	Ar 18									
0	Ia	IIa	IIIa	IVa	Va	VIa	VIIa	VIII			Ib	IIb	IIIb	IVb	Vb	VIb	VIIb	0
Ar 18	K 19	Ca 20	Sc 21	Ti 22	V 23	Cr 24	Mn 25	Fe 26	Co 27	Ni 28	Cu 29	Zn 30	Ga 31	Ge 32	As 33	Se 34	Br 35	Kr 36
Kr 36	Rb 37	Sr 38	Y 39	Zr 40	Nb 41	Mo 42	Tc 43	Ru 44	Rh 45	Pd 46	Ag 47	Cd 48	In 49	Sn 50	Sb 51	Te 52	I 53	Xe 54
Xe 54	Cs 55	Ba 56	La 57	*	Hf 72	Ta 73	W 74	Re 75	Os 76	Ir 77	Pt 78	Au 79	Hg 80	Tl 81	Pb 82	Bi 83	Po 84	At 85	Rn 86
Rn 86	Fr 87	Ra 88	Ac 89	* *	*Rf* 104	*Ha* 105												

*Lanthanons	Ce 58	Pr 59	Nd 60	Pm 61	Sm 62	Eu 63	Gd 64	Tb 65	Dy 66	Ho 67	Er 68	Tm 69	Yb 70	Lu 71
* *Actinons	Th 90	Pa 91	U 92	Np 93	Pu 94	Am 95	Cm 96	Bk 97	Cf 98	Es 99	Fm 100	Md 101	No 102	Lr 103

Figure 3: Long-period form of periodic system of elements.

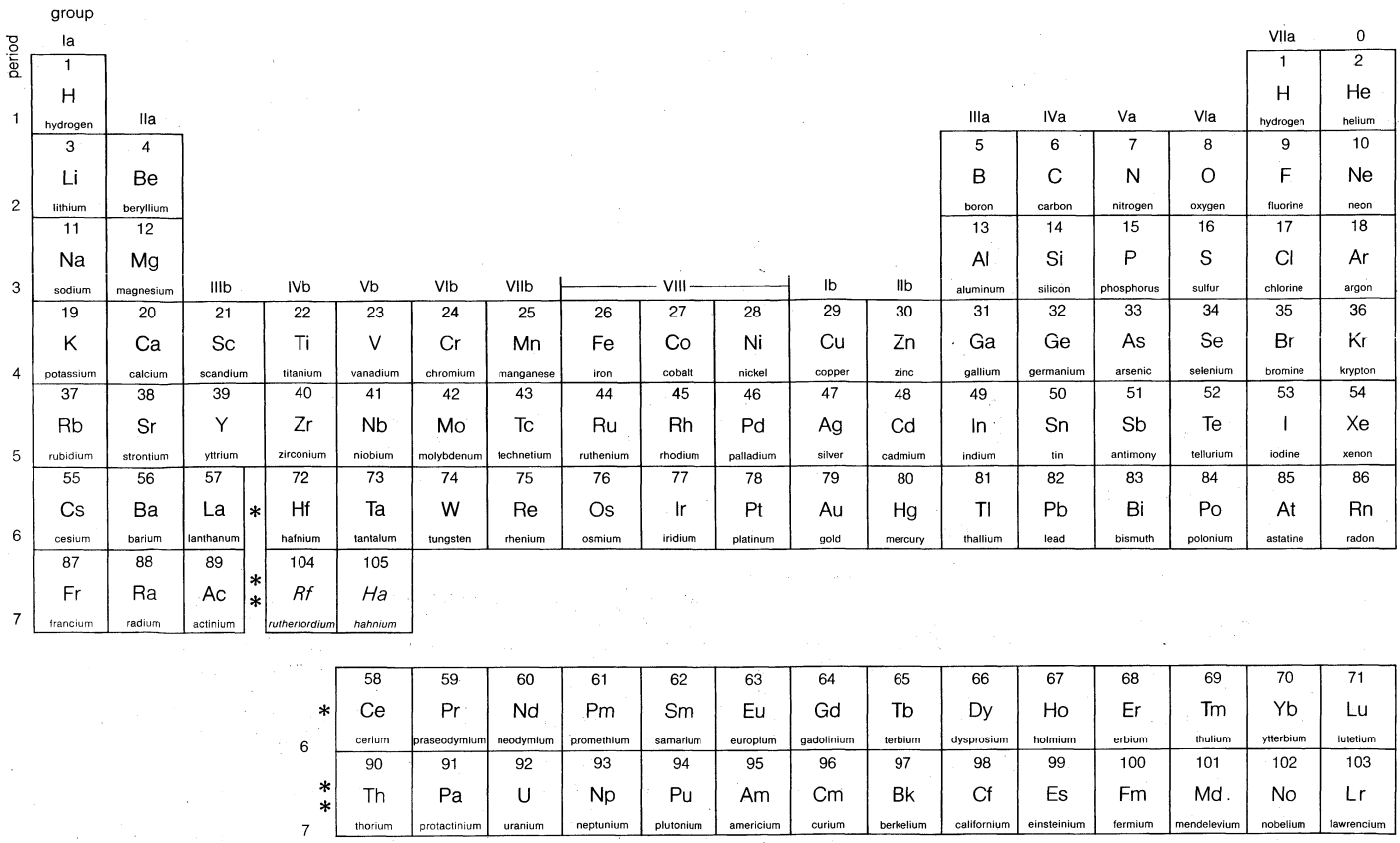


Figure 4: A modern version of the periodic table of elements.

Figure 1), however, it is customary to describe them as constituting 15 subgroups, the three consecutive elements iron, cobalt, and nickel and their congeners in the succeeding periods being placed in a single group, VIII. The elements potassium, 19, to manganese, 25, are assigned to subgroups Ia, IIa, IIIa, IVa, Va, VIa, and VIIa, respectively; and the elements copper, 29, to bromine, 35, are assigned to subgroups Ib, IIb, IIIb, IVb, Vb, VIb, and VIIb. (In some modern versions of the periodic system, the members of subgroups a and b in Groups III and above are the reverse of the arrangement shown here. There is no one standard form of the system.)

The first group, the alkali metals, thereby includes, in addition to lithium and sodium, the metals from potassium down the table to francium, but not the much less similar metals of subgroup Ib (copper, etc.). Also the second group, the alkaline-earth metals, is considered to include beryllium, magnesium, calcium, strontium, barium, and radium but not the elements of subgroup IIb. The properties of the elements do not indicate unambiguously whether the boron group might be considered better to include the subgroup IIIa or the subgroup IIIb, although the former is often so included. The other four groups are as follows: the carbon group, IV, consists of carbon, silicon, germanium, tin, and lead; the nitrogen group, V, includes nitrogen, phosphorus, arsenic, antimony, and bismuth; the oxygen group, VI, includes oxygen, sulfur, selenium, tellurium, and polonium; and the halogen group, VII, includes fluorine, chlorine, bromine, iodine, and astatine.

Although hydrogen is included in Group I in some version of the table, it is not closely similar to either the alkali metals or the halogens in its chemical properties. It is, however, assigned the oxidation number $+1$ in compounds such as hydrogen fluoride, HF, and -1 in compounds such as lithium hydride, LiH; and it may hence be considered as being similar to a Group Ia element and to a Group VII element, respectively, in compounds of these two types, taking the place first of Li and then of F in lithium fluoride, LiF. Hydrogen is, in fact, the most individualistic of the elements: no other element resembles it in the way that sodium resembles lithium, chlorine resembles fluorine, and neon resembles helium. It is a unique element, the only element that cannot conveniently be considered a member of a group.

A number of the elements of each long period are called the transition metals. These are usually taken to be scandium, 21, to zinc, 30 (the iron-group transition metals); yttrium, 39, to cadmium, 48 (the palladium-group transition metals); and hafnium, 72, to mercury, 80 (the platinum-group transition metals). By this definition, the transition metals include Groups IIIa to VIIa, VIII, and Ib and IIb (Figure 3) in the long periods.

Periodic properties

The periodicity in properties of the elements arranged in order of atomic number is strikingly shown by the consideration of the physical state of the elementary substances and such related properties as the melting point, density, and hardness. The elements of Group 0 are gases that are difficult to condense. The alkali metals, in Group I, are soft, metallic solids with low melting points. The alkaline-earth metals, in Group II, are harder and have higher melting points than the adjacent alkali metals. The hardness and melting point continue to increase through Groups III and IV and then decrease through Groups V, VI, and VII. The elements of the long periods show a gradual increase in hardness and melting point from the beginning alkali metals to near the centre of the period and then at Group VIb an irregular decrease to the halogens and noble gases.

The valence of the elements (that is, the number of bonds formed with a standard element) is closely correlated with position in the periodic table, the elements in the main groups (the "a" groups, as opposed to the transition elements or "b" groups) having maximum positive valence, or oxidation number, equal to the group number, and maximum negative valence equal to the difference between eight and the group number.

The general chemical properties described as metallic or base forming, metalloid or amphoteric, and nonmetallic, or acid forming, are correlated with the periodic table in a simple way: the most metallic elements are to the left and to the bottom of the periodic table, and the most nonme-

tallic elements are to the right and to the top (ignoring the noble gases). The metalloids are adjacent to a diagonal line from boron to polonium. A closely related property is electronegativity, the tendency of atoms to retain their electrons and to attract additional electrons. The degree of electronegativity of an element is shown by ionization potential, electron affinity, oxidation-reduction potential, the energy of formation of chemical bonds, and other properties. It is shown to depend upon the element's position in the periodic table in the same way that nonmetallic character does, fluorine being the most electronegative element and cesium (or francium) the least electronegative (most electropositive) element.

Atomic sizes

The sizes of atoms of elements vary regularly throughout the periodic system. Thus, the effective bonding radius (or one-half the distance between adjacent atoms) in the elementary substances in their crystalline or molecular forms decreases through the first short period from 1.52 angstrom (angstrom unit = 10^{-8} cm) for lithium to 0.73 angstrom for fluorine; at the beginning of the second period, the bonding radius rises abruptly to 1.86 angstrom for sodium and gradually decreases to 0.99 angstrom for chlorine. The behaviour through the long periods is more complex: the bonding radius decreases gradually from 2.31 angstrom for potassium to a minimum of 1.25 angstrom for cobalt and nickel, then rises slightly, and finally falls to 1.14 angstrom for bromine. The sizes of atoms are of importance in the determination of coordination number (that is, the number of groups attached to the central atom in a compound) and hence in the composition of compounds. The increase in atomic size from the upper right corner of the periodic table to the lower left corner is reflected in the formulas of the oxygen acids of the elements in their highest states of oxidation, as shown in the Table.

Formulas of Oxygen Acids of Elements in Their Highest Oxidation State

$H_2Be(OH)_4$	H_3BO_3	H_2CO_3	HNO_3		
	$HAl(OH)_4$	H_4SiO_4	H_3PO_4	H_2SO_4	$HClO_4$
		H_4GeO_4	H_3AsO_4	H_2SeO_4	
		$H_2Sn(OH)_6$	$HSb(OH)_6$	H_6TeO_6	HIO_4
				and	and
				H_2TeO_4	H_5IO_6

The smallest atoms group only three oxygen atoms about themselves; the next larger atoms, which coordinate a tetrahedron of four oxygen atoms, are in a diagonal belt; and the still larger atoms, which form octahedral oxygen complexes (stannic acid, antimonic acid, telluric acid, paraperiodic acid), lie below and to the left of this belt. Only the chemical and physical properties of the elements are determined by the extranuclear electronic structure; these properties show the periodicity described in the periodic law. The properties of the atomic nuclei themselves, such as the magnitude of the packing fraction and the power of entering into nuclear reactions, are, although dependent upon the atomic number, not dependent in the same periodic way.

The electronic structure of atoms as the basis of the periodic system. The noble gases, helium, neon, argon, krypton, xenon, and radon, have the striking chemical property of forming few chemical compounds. This property would depend upon their possessing especially stable electronic structures (that is, structures so firmly knit that they would not yield to accommodate ordinary chemical bonds). During the development of modern atomic physics and the theory of quantum mechanics (*q.v.*), a precise and detailed understanding was obtained of the electronic structure of the noble gases and other atoms that explains the periodic law in a thoroughly satisfactory manner.

Quantum numbers

The Pauli exclusion principle states that no more than two electrons can occupy the same orbit—or, in quantum-mechanical language, orbital—in an atom and that two electrons in the same orbital must be paired (that is, must have their spins opposed). The orbitals in an atom may be described by a principal quantum number, n, which may assume the values 1, 2, 3, . . . , and by an azimuthal quantum number, l, which may assume the values 0, 1, 2, . . . , $n - 1$. There are $2l + 1$ distinct orbitals for each set of values of n and l. The most stable orbitals, which bring the electron closest to the nucleus, are those with the smallest values of n and l. The electrons that occupy the orbital with $n = 1$ (and $l = 0$) are said to be in the K shell of electrons; the L, M, N, . . . shells correspond respectively to $n = 2, 3, 4, \ldots$. Each shell except the K shell is divided into subshells corresponding to the values 0, 1, 2, 3, . . . of the orbital quantum number l; these subshells are called the s, p, d, f, . . . subshells, and they can accommodate a maximum of 2, 6, 10, 14, . . . electrons. (There is no special significance to the letter designations of the quantum numbers or of the shells and subshells.)

The approximate order of stability of the successive subshells in an atom is indicated in the Table below. The number of electrons in the atoms of the elements increases with increasing atomic weight, and the added electrons go, of necessity, into successively less stable shells. The most stable shell, the K shell, is completed with helium, which has two electrons. The L shell is then completely filled at neon, with atomic number 10. The atoms of the heavier noble gases do not, however, have a completed outer shell but instead have s and p subshells only. The outer shell of eight electrons is called traditionally an octet. The d subshells and f subshells subsequently are also filled with electrons after the initially less stable orbitals are occupied, an inversion of stability having occurred with increasing atomic number.

The electron occupancy of the shells in the noble gas atoms is as follows:

Electrons in		*K*	*L*	*M*	*N*	*O*	*P*	*Q*	shell
He	2:	2							
Ne	10:	2	8						
Ar	18:	2	8	8					
Kr	36:	2	8	18	8				
Xe	54:	2	8	18	18	8			
Rn	86:	2	8	18	32	18	8		
	118:	2	8	18	32	32	18	8	

The numbers 2, 8, 18, and 32 correspond to filling the s; s and p; s, p, and d; and s, p, d, and f subshells, respectively.

Periods and electron shells

The first period of the periodic table is complete at helium, when the K shell is filled with two electrons. The first and second short periods represent the filling of the $2s$ and $2p$ subshells (completing the L shell at neon) and the $3s$ and $3p$ subshells (at argon), leaving the M shell incomplete. The first long period begins with the introduction of electrons into the $4s$ orbital. Then, at scandium, the five $3d$ orbitals of the inner M shell begin to be occupied. It is the successive occupancy of these five $3d$ orbitals by their complement of ten electrons that characterizes the ten elements of the iron-group transition series. At krypton the M shell is complete and there is an octet in the N shell. The second long period, of 18 elements, similarly represents the completion of an outer octet and the next inner subshell of $10d$ electrons.

The very long period of 32 elements results from the completion of the $4f$ subshell of 14 electrons, the $5d$ subshell of 10 electrons, and the $6s$, $6p$ octet. The filling of the $4f$ orbitals corresponds to the sequence of 14 lanthanons, and that of the $5d$ orbitals to the 10 platinum-group transition metals.

The next period, which is incomplete, involves the $5f$ subshell of 14 electrons, the $6d$ subshell of 10 electrons, and the $7s$, $7p$ octet. The filling of the $5f$ orbitals corresponds to the actinons, the elements beginning with thorium, 90.

The successive periods of the system hence correspond to the introduction of electrons into the following orbitals:

The very short period, H, He.	2 elements. $1s$.
The first short period, Li to Ne.	8 elements. $2s$, $2p$.
The second short period, Na to Ar.	8 elements. $3s$, $3p$.
The first long period, K to Kr.	18 elements. $3d$, $4s$, $4p$.
The second long period, Rb to Xe.	18 elements. $4d$, $5s$, $5p$.
The very long period, Cs to Rn.	32 elements. $4f$, $5d$, $6s$, $6p$.
The incomplete period, Fr to 118.	32 elements. $5f$, $6d$, $7s$, $7p$.

There are advantages to replacing the K, L, M, . . . shells by a different grouping of the subshells, in which those with nearly the same energy are grouped together, in close correlation with the periodic system. The new set of shells is the following:

The helium shell.	2 electrons. $1s$.
The neon shell.	8 electrons. $2s$, $2p$.
The argon shell.	8 electrons. $3s$, $3p$.
The krypton shell.	18 electrons. $4s$, $4p$, $3d$.
The xenon shell.	18 electrons. $5s$, $5p$, $4d$.
The radon shell.	32 electrons. $6s$, $6p$, $5d$, $4f$.
The eka-radon shell.	32 electrons. $7s$, $7p$, $6d$, $5f$.

Periodicity in properties of the elements. *The main groups.* The periodicity of properties of the elements is caused by the periodicity in electronic structure. The noble gases are chemically unreactive, or nearly so, because their electronic structures are stable—their atoms hold their quota of electrons strongly, have no affinity for more electrons, and have little tendency to share electrons with other atoms. An element close to a noble gas in the periodic system, on the other hand, is reactive chemically because of the possibility of assuming the stable electronic configuration of the noble gas, by losing one or more electrons to another atom, by gaining one or more electrons from another atom, or by sharing electrons. The alkali metals, in Group I, can assume the noble-gas configuration by losing one electron, which is loosely held in the outermost (valence) shell, to another element with greater electron affinity, thus producing the stable singly charged positive ions. Similarly the alkaline-earth metals can form doubly charged positive ions with the noble-gas electronic configuration by losing the two loosely held electrons of the valence shell; the positive ionic valences of the elements of the first groups are hence equal to the group numbers. The elements just preceding the noble gases can form negative ions with the noble-gas configuration by gaining electrons; the negative ionic valences of these elements are equal to the difference between eight and their group numbers. The covalence (or number of shared electron pairs) of an atom is determined by its electron number and the stable orbitals available to it. An atom such as fluorine, with seven electrons in its outer shell, can combine with a similar atom by sharing a pair of electrons with it; each atom thus achieves the noble-gas configuration by having three unshared pairs and one shared electron pair in its valence shell.

Properties and electronic structure

The properties of elements in the same group of the periodic system are, although similar, not identical. The trend in properties from the lighter to the heavier elements may be attributed to changes in the strength of binding of the outer electrons and especially to the increasing size of the atoms.

The transition elements. The nature of the three transition series can be seen by taking the iron elements as an example, since the other two series are similar to it. In potassium and calcium the valence electrons, outside of the argon shell, occupy the $4s$ orbital. Additional valence electrons, in scandium, titanium, etc., enter not the $4p$ subshell but the $3d$ subshell, which is then completely filled, along with $4s$, in zinc. The elements potassium, calcium, scandium, titanium, vanadium, chromium, and manganese have maximum oxidation numbers that range from one to seven in that order, corresponding to their successive positions following argon. The atoms titanium to manganese are, however, not similar to the corresponding atoms carbon to fluorine or silicon to chlorine in the short periods because they are not directly followed by a noble gas, whose stable structure they might achieve by gaining or sharing electrons. Instead, these atoms have a strong tendency to form bipositive or tripositive ions by losing two or three electrons, the remaining valence electrons occupying the $3d$ orbitals, which for these positive ions are more stable than the $4s$ orbital.

The iron series

The elements copper and zinc, by losing one or two electrons, respectively, can form the ions Cu^+ and Zn^{2+} with completed M shells of electrons. If the $4s$ and $4p$ orbitals were much less stable than the $3d$ orbitals, instead of about equally stable, this electronic configuration would be similar to the neon configuration; and nickel would be a noble gas, copper an alkali metal, and zinc an alkaline-earth metal. The actual, different properties of nickel, copper, and zinc result from their ability to use the $4s$ and $4p$ orbitals in forming chemical bonds. Nevertheless, the effect of the 18-electron-shell configurations (of Cu^+, Ag^+, Au^+) is evident in the normal valences of the elements of subgroups Ib, IIb, IIIb, IVb, which are equal to the group numbers. Some of these elements may also assume other valences, one or two greater (by losing additional electrons, leaving the $3d$ subshell incomplete) or two less (by having the $4s$ orbital occupied by an unshared electron pair).

It is convenient to consider ten of the elements in each long period or very long period to be transition elements, corresponding to the introduction of ten electrons in a d subshell. There is some arbitrariness in the selection of these elements, because the sharp differentiation of s, p, and d subshells for the outer electrons of an atom in a molecule or crystal is not possible; these subshells are precisely defined only for a system with spherical symmetry; *i.e.*, for a system containing only one nucleus.

The lanthanons. The existence of a long series of very closely related elements, the lanthanons, or rare-earth elements, has been clarified by the development of the understanding of the electronic structure of atoms. With increase in the atomic number of the elements, the seven $4f$ orbitals become increasingly stable until at cerium, 58, these orbitals begin to be occupied. In the tripositive gadolinium ion, 64, the $4f$ subshell is half filled, with one electron in each of the seven orbitals; and in the lutetium ion, 71, it is completely filled with its complement of 14 electrons.

The lanthanons thus comprise the 14 metals from cerium to lutetium, with the same electronic structure except for the variation from 1 to 14 in the number of $4f$ electrons. Often lanthanum, with the same electronic structure but with no $4f$ electrons in the terpositive ion, is also included among the lanthanons, increasing their number to 15.

The $4f$ orbitals are deep within the atoms; there is a stable noble-gas shell of eight $5s$ and $5p$ electrons outside them, as well as—in the neutral atoms—two valence electrons in the $6s$ orbital, and perhaps one in a $5d$ orbital. Because the $4f$ orbitals are in an inner shell, the properties of the atom or ion are changed only slightly by changes in the number of electrons that occupy them; hence the lanthanons are closely similar to one another in physical and chemical behaviour and are often called an inner transition series. This similarity is made strikingly evident by the great difficulty of separating these elements from the complex mineral mixtures in which they occur in nature, a circumstance that caused their discovery to be spread over a period of more than a century.

The lanthanons all have +3 as their principal oxidation state, corresponding to the loss of the three valence electrons to more electronegative atoms. Cerium, praseodymium, and terbium also occur in oxidation state +4, and samarium, europium, and ytterbium in oxidation state +2.

The actinons. The only important general question about the periodic system, to which the answer remains doubtful, concerns the nature of the last period. Because thorium is more closely similar to hafnium than to cerium in many of its properties, and uranium is more closely similar to tungsten than to neodymium, it had been thought that the uranium elements constitute the beginning of a transition period similar to the two long periods and involving the occupancy of the $6d$ subshell. G.T. Seaborg, however, pointed out that the properties of the transuranium elements suggest that these elements are members of a second inner transition series of the lanthanon type beginning at thorium and corresponding to a filling of the $5f$ orbitals, as the lanthanons correspond to the filling of the $4f$ orbitals. It is true that uranium is unlike the lanthanons in that the +6 oxidation state is the most stable of its oxidation states (+6, +5, +4, +3), but neptunium, instead of being similar to rhenium, is reported not to have a stable +7 state; instead, neptunium and plutonium are said to have the same oxidation states as uranium but with a general shift in stability toward the lower oxidation states.

Second inner transition series

An answer to this question can be given by the consideration of the stability and nature of the electronic orbitals involved. Although d orbitals contribute to the formation of strong covalent bonds, inner f orbitals are not useful in

this way, as is shown by the small tendency of the lanthanons to form covalent compounds. The fact that uranium does form covalent compounds shows that the *6d* orbitals are stable enough to be used for this purpose; and the stability of the +6 oxidation state shows that the *5f* orbitals of uranium are not sufficiently stable to attract any of the six valence electrons very strongly. Hence the uranium period differs from the immediately preceding period in that the *5f* orbitals are less stable relative to the *6d* orbitals than are the *4f* relative to the *5d* orbitals.

With increase in atomic number, however, the *5f* orbitals increase in stability more rapidly than the *6d* orbitals, and ultimately the *5f* orbitals are filled. It is likely that this process begins at neptunium, and presumably it is completed at element 103, which is similar to lutetium in structure and properties, with probably a greater tendency to the formation of covalent compounds.

The *6d*, *7s*, and *7p* subshells are thus successively occupied, with element 118, eka-radon, as the next noble gas, though probably able to form some chemical compounds. The 14 elements from thorium to lawrencium are called the actinons or actinides.

Thus this great generalization, the periodic law, which first showed its value through Mendeleyev's prediction of the existence and properties of undiscovered elements, is again being put to the same use, the undiscovered elements of present interest being those that have greater atomic number than the known elements. The undiscovered elements are being sought not in nature's minerals but in preparations made by man.

BIBLIOGRAPHY. J.W. VAN SPRONSEN, *The Periodic System of Chemical Elements* (1969), a thorough treatment of the history of the periodic system and the periodic table; LINUS C. PAULING, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed. (1960), a general discussion of atomic and molecular structure and of valence in relation to the periodic system and the elements.

(L.C.P.)

Perissodactyla

The order Perissodactyla is a group of herbivorous mammals characterized by the possession of either one or three hoofed toes on each hindfoot. The name (Greek *perissos*, "odd," and *daktylos*, "finger") was introduced to separate the odd-toed ungulates from the even-toed ones (Artiodactyla), all of which had previously been classified as members of a single group.

GENERAL FEATURES

The Perissodactyla comprise three families of living mammals: six species of horses (Equidae), four species of tapirs (Tapiridae), and five species of rhinoceroses (Rhinocerotidae). These families are remnants of a group that flourished during the Tertiary Period (from 65,000,000 to 2,500,000 years ago), a time when it was much richer in species and in variety of form than at present and played a dominant role in the fauna of the world.

The horses, asses, and zebras are long-legged, running forms with one functional digit in each foot and with high-crowned, molariform (*i.e.*, modified for grinding) cheek teeth. The tapir is a rather rounded, piglike, semi-amphibious forest and woodland animal with a small proboscis (trunklike snout) and a coat of short, bristly hairs. Tapirs have primitive features, such as four hoofed toes in the forefoot and three in the hind, and they have rather simple molar teeth. Rhinoceroses are massive, graviportal (ponderous) creatures with a thick and nearly hairless hide and three digits on each foot. They bear hornlike structures on the head.

The Perissodactyla are of particular scientific interest because their fossil history is so well-known. The evolution of horses from the tiny "dawn horse" (*Hyracotherium*, formerly *Eohippus*) to the present form is a classic sequence, knowledge of which has played an important role in evolutionary thought. The order also provides a notable example of parallel evolution. Following completely different evolutionary paths, both perissodactyls and artiodactyls (*e.g.*, cattle, antelope, swine) independently evolved features such as high-crowned grinding teeth and elongated limbs with a reduced number of digits, in adaptation to a similar running (cursorial), herbivorous mode of life.

Living perissodactyls are of medium or large size. Asses and tapirs, the smallest representatives of the order, attain a length of approximately two to 2.5 metres (6½–eight feet), stand one metre or more at the shoulder, and weigh up to 250 or 300 kilograms (550–650 pounds). The largest forms are the Indian and square-lipped rhinoceroses (*Rhinoceros unicornis* and *Ceratotherium simum*, respectively), which are four to five metres long and measure up to two metres at the shoulder. The maximum weight has not been well established, but a figure of 2,070 kilograms (4,560 pounds) has been recorded. *Baluchitherium*, relative of the rhinoceroses known from the Oligocene, (about 30,000,000 years ago), was the largest known land mammal, standing about 5.5 metres (18 feet) at the shoulder.

Indian and square-lipped rhinoceroses

Ecologically, the Perissodactyla were the dominant large herbivore group during the Tertiary, a position now held by the Artiodactyla. All feed either by grazing (*i.e.*, cropping grasses) or by browsing (taking shoots and leaves from trees and bushes). The Equidae in particular were abundant and important members of the Old World fauna until their numbers were reduced by modern man. Zebras are still numerous and ecologically important in a few parts of Africa. The importance of the domestic horse and the ass in the history of mankind is very great indeed. Both have served extensively as pack, draft, and riding animals. The horse is sometimes eaten by man, and its flesh is widely used as pet food; through centuries of domestication, it has been developed into a number of different breeds (see HORSE).

The living wild Equidae are confined to the Old World. Zebras and the true wild ass (*Equus asinus*) are African, with the zebras confined to the southern and eastern parts, while the ass originally ranged over northern and northeastern Africa.

The wild horse (*Equus caballus*), ancestor of the domestic horse, occupied the low country north of the great mountain ranges from Europe across central Asia; it may now be extinct as a wild animal. The half-asses, races of *E. hemionus*, were found in the arid zone of Asia from Persia to the Gobi Desert, as well as in Arabia, Syria, and northwestern India.

The living rhinoceroses are also Old World forms, with two species in Africa and three in Asia. There are three species of tapirs in the New World tropics, one in Middle America and two in South America. The fourth species of tapir is Asiatic.

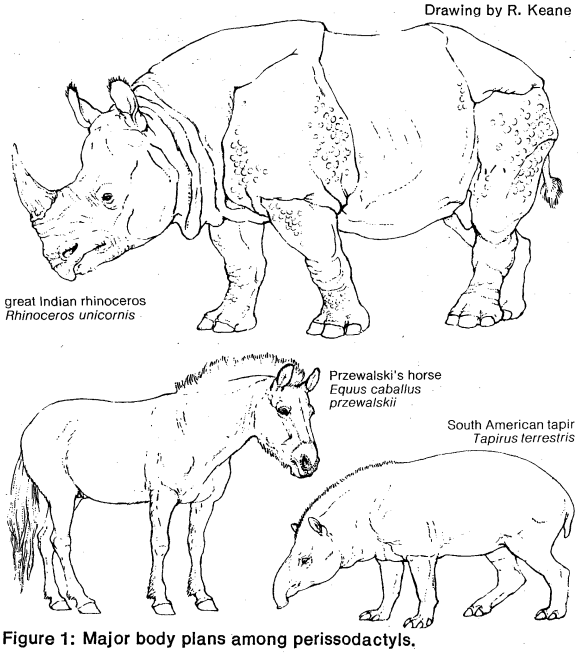


Figure 1: Major body plans among perissodactyls.

NATURAL HISTORY

Distribution, ecology, and conservation. The Equidae are highly specialized for a cursorial, herbivorous mode of life. They are absent from forests and other densely vegetated regions, but apart from this limitation, the range of the group is relatively unrestricted by the type of vegetation, climate, and topography. The species replace one another geographically for the most part, and each occupies a somewhat different habitat.

Dietary needs as range factors

Grass plays a major role in the diet; the zebras, for example, are known to feed on tall, coarse grasses avoided by most antelopes. Some species also take shrubs, herbs, and even bulbs. Water requirements vary in different species. In South Africa, Burchell's zebra has been found to drink about once every 36 hours. By contrast the mountain zebra (*Equus zebra*), Przewalski's horse (*Equus caballus przewalskii*) and the half-ass, all living in semidesert areas, are reported to survive if they can drink once in three or four days. The ass, too, can manage with less water than the horse. The mountain zebra and Przewalski's horse dig for water in dry riverbeds and depressions.

Zebras. The mountain zebra occupies parts of the arid rocky escarpment separating the interior plateau of the southern African subcontinent from the coast lowlands. The race *E.z. hartmannae*, still relatively numerous over much of its original range in South West Africa and southern Angola, enjoys legal protection and is represented in game reserves. Conflict between these zebras and farm livestock for the meagre pasture of the semidesert regions, however, has led to a reduction in zebra numbers. The race *E.z. zebra* was originally common in the mountain ranges of the Cape Province, but now survives only as a remnant of perhaps 100 animals. About one-half of these have sanctuary in a national park.

Burchell's zebra, (*E. quagga*) formerly inhabited a great area of grassland and savanna from the Cape to the southern Sudan. The southernmost race (*E. q. quagga*), which was only partly striped, became extinct in the 19th century. The populations of the other races have been much reduced in many places and the range of the species has shrunk considerably. There are large populations in reserves, however, and the species is not in any immediate danger of extermination. Grevy's zebra (*E. grevyi*), which shares a narrow zone in northern Kenya with Burchell's zebra, is confined to sparsely wooded, semidesert plains and low hills in northern Kenya, southern and eastern Ethiopia, and western Somaliland. Its status appears to be generally satisfactory.

Asses. The true ass (*Equus asinus*), ancestor of the domestic donkey, is the equid of arid North Africa whose range extends south to approximately 6° N latitude. Its natural distribution probably included all habitable parts of North Africa. At present, asses are known from semidesert country extending from the east bank of the Nile (in the Sudan) to the Red Sea, and in parts of Eritrea, Ethiopia, and Somalia. There are also isolated pockets in the Tibesti Mountains in the Sahara, and in the countries of central and western Africa. There is a great deal of uncertainty about the identity of all asses now described as "wild." Some may be merely feral (escaped or released) donkeys, and interbreeding with feral donkeys is likely to have occurred in many, if not all, existing populations.

The two surviving races of wild horses

The wild horse. The wild horse was widely distributed in Eurasia north of the mountain chains. The Romans encountered it in Spain. Two races have survived to modern times. A gray race, known as the tarpan, was the horse of southern Russia. It became extinct in the Ukraine during the mid-19th century. Przewalski's horse (*E.c. przewalskii*), a small, reddish-brown race considered a species by some authors, still survived as a truly wild animal until 1950, and small numbers may yet exist in the remote, semidesert steppe country on the boundary between Mongolia and China. Wild horses enjoy legal protection in Mongolia and China, but nomadic pastoralists have been encroaching on previously uninhabited country and competing with the horses for pasture and the scarce water supplies.

The half-asses, races of *Equus hemionus*, occupied the dry belt from Mongolia through Central Asia to Syria, with a northern limit at about 50° N latitude. The chigetia or kulan (*E. h. hemionus*), which was formerly widespread over an immense region of the Gobi, now occurs only in semidesert steppe country in central Mongolia. Hunting and competition for water by pastoral tribesmen are responsible for its decline. The kulan is slightly smaller than the kiang, (*E. h. kiang*), which is found on the cold, arid steppes of Sadak, Nepal, Sikkim, and western Tibet at altitudes of 4,270 metres (14,000 feet) and more. The kiang is now said to be rare but not endangered. The Persian onager, (*E. h. onager*) lives in a lower semidesert or desert environment, with a range that formerly included northeastern Iran, northwestern Afghanistan, and Russian Turkestan. It is now extremely rare and unlikely to survive outside northeastern Iran and the Badkhyz Reserve in Turkmeniya. A small nucleus has sanctuary in the semidesert salt plains of the recently established Kavir Protected Region in Iran. The Indian wild ass is a closely related, probably identical, form sometimes distinguished as the race *E. h. khur*. A fairly small population occupies salt flats in the Rann of Kutch, a remnant of the thousands found there at the end of World War II. The Syrian onager (*E. hemionus hemippus*) is the smallest member of the group and stands about one metre (3 feet) at the shoulder. It was once found in the desert region of Palestine, Syria, and Iraq, and was domesticated by the ancient Sumerians before the introduction of the domestic horse into Mesopotamia. This race may survive in the Djezireh Desert, Syria, or north of the Syrian–Turkish border; if so, the number must be extremely small.

"White" and "black" rhinos of Africa

Rhinoceroses. The two African species of rhinoceros are the black or prehensile-lipped rhinoceros (*Diceros bicornis*) and the white or square-lipped rhinoceros (*Ceratotherium simum*). The terms black and white are misleading, since both species are grayish to brownish, but the names are well established in common usage.

The black rhinoceros was originally widespread from the Cape to southwestern Angola and throughout eastern Africa as far as Somalia, parts of Ethiopia and The Sudan. Its range also extended westward through the northern savanna zone to Lake Chad, the northern Cameroons, Volta, and the northern Ivory Coast. The animal was extremely numerous in some parts. It now occupies a much smaller area, within which it is found in scattered pockets, many of them in parks and reserves. South Africa has an important population of some 300 of them in its Zululand reserves. The species still occurs in northern South West Africa and southwestern Angola, Rhodesia, Mozambique, Malawi, Zambia, and Zaire. Northern Tanzania and Kenya have more black rhino than any other country but the future of the animals outside the parks and reserves is far from secure. The species is also found in parts of Somalia, Ethiopia, northern Uganda, and The Sudan, and remnants remain in the Central African Republic and northern Cameroon. The decline in numbers is largely the result of expanding human settlement and of poaching to obtain the horns, which fetch high prices.

The black rhinoceros occupies a variety of habitats, frequenting open plains, sparse thorn scrub, savannas, thickets, and dry forests, as well as mountain forests and moorlands at high altitudes. It is a selective browser, and grass plays a minor role in its diet. Succulent plants, such as euphorbias, assume great importance in dry habitats, and where these plants are abundant the animals appear to be able to survive without free water. Where water is available, drinking is regular and frequent; the animals also may dig for water in dry riverbeds.

The much larger white rhinoceros is a grazing species with a broad, square muzzle. It prefers short grasses seven to ten centimetres (about three to four inches) high. The animal makes much use of shade trees for resting and is dependent on surface water. The range of the white rhinoceros is markedly discontinuous. South of the Zambesi River it was once extremely common over a fairly large area of bushveld. It has since become confined to the game reserves in Zululand, where the population has

risen to some 1,700; some of the animals have been redistributed to several other parks and reserves in southern Africa.

A northern race formerly inhabited the southern Sudan and adjacent areas of Uganda and Zaire, extending westward into the Central African Republic. It has also been much reduced, but considerable populations still survive in the Parc National de la Garamba (Zaire) and in the Bahr el Ghazal region of The Sudan, and a small remnant has found sanctuary in the Ruwenzori National Park, Uganda.

The smallest living rhinoceros

The smallest of the three Asian rhinoceroses (also the smallest living member of the family) is the Sumatran, or Asiatic, two-horned rhinoceros, *Didermocerus* (or *Dicerorhinus*) *sumatrensis*, standing one to one and a half metres (three to five feet) at the shoulder. It was originally found from eastern Pakistan and Assam throughout Burma, much of Thailand, Indochina (Cambodia, Laos, and Viet Nam), Malaya, Sumatra, and Borneo. Small isolated populations still occur in a few widely separated localities in Burma, Thailand, Malaya, Sumatra, and Sabah, and possibly in other nearby territories. The total population is thought to number between 100 and 170. Some of the survivors in Sumatra are protected in reserves.

The Javan, or lesser one-horned, rhinoceros (*Rhinoceros sondaicus*) occupied the islands of Java, Borneo, and Sumatra, the Malay Peninsula, and a region extending northwards through Burma into Assam and eastern Bengal. It is now restricted to the Udjung Kulon Reserve in western Java where there are at least 25 and perhaps as many as 50 to 60 animals.

Both the Sumatran and Javan rhinoceroses inhabit forests as well as marshy areas and regions of thick bush and bamboo, climbing actively in mountainous country. They are mainly browsers.

The great Indian, or one-horned, rhinoceros (*Rhinoceros unicornis*) is more or less equivalent in size to the square-lipped rhinoceros and is distinguishable from the smaller Javan rhinoceros by the presence of a large horn, tubercles on its skin, and a different arrangement of skinfolds. It previously occupied an extensive range across northern India and Nepal from Assam in the east to the Indus Valley in the west. It is found in a range of habitats —open grassland, savanna, forests, and hilly country— and appears to be mainly a grazer, raiding grain fields in some areas. Hunting and the pressure of expanding human populations have greatly reduced both the range and numbers of this animal. It is now found almost entirely in eight reserves or sanctuaries in India, notably the Kaziranga Sanctuary in Assam (estimated population 300) and in the Rapti Valley region of the Nepal Terai. The total population is estimated at about 600 animals, and the prospects for survival appear to be reasonably good.

Tapirs. The Malayan tapir (*Tapirus indicus*), largest member of the family Tapiridae, is found in Sumatra and the Malay Peninsula, as far north as the Burmo-Siamese border in latitude 18° N. It is found from sea level to high altitudes and occupies forests and thickets but may feed in more open areas. It is still abundant and widespread.

Distribution of New World tapirs

The three New World species occupy distinct, nonoverlapping but contiguous ranges. The mountain tapir (*Tapirus pinchaque*), the smallest and most primitive, inhabits the temperate-zone forests and bordering grasslands of the Andes in Colombia and Ecuador and in northern Peru, up to altitudes of nearly 4,600 metres (about 15,000 feet). Agricultural and pastoral expansion have resulted in some decline in the status of this species, but it is still fairly common. The Central American, or Baird's, tapir (*T. bairdii*) is the largest of the American species. It is essentially middle American, with a range extending from Mexico into coastal Ecuador, and it occupies undisturbed climax rain forest. It is shy and adjusts poorly to the disturbance caused by settlement. This disturbance, together with the destruction of habitat accompanying human occupation, has greatly reduced its range and numbers. The species is said to be much in need of active conservation. The most widespread species is the Brazilian tapir (*T. terrestris*), which is found throughout the Brazilian subregion east of the Andes and in a small area west of the Andes in northwestern Venezuela and northern Columbia. Like the other species, it is largely a forest form requiring the proximity of water. The three New World tapirs are mainly browsers and are remarkably similar in habits.

Behaviour. *Expression and communication.* The Equidae communicate by means of calls and changes in facial expression. Six different sounds are made by Burchell's zebra. A whinny, consisting of a series of two- or three-syllabic "ha" sounds, serves to maintain contact between members of a group. The repertoire includes an alarm call ("i-ha"), an alarm snort, a drawn-out snort of satisfaction, and a squeal of pain and fear. Other species utter similar sounds, the whinny of the horse and the bray of the ass being well-known examples. Characteristic facial expressions have been described for greeting ceremonies (mouth open, ears up), threat (mouth open, ears back), and submission (mouth open, nibbling movements, ears down). In all species studied except the horse, females assume a particular expression ("mating face") when permitting the male to mount.

In the rhinoceroses and tapirs snorting, squealing, bellowing, and, in some forms, whistling sounds play a major role in communication. Visual signals are not well developed in these nonsocial animals, but a few facial expressions are used.

Feeding. The Perissodactyla are mainly grazers or browsers. The quality and quantity of grasses available to grazing species may vary considerably with the season and the area. The animals may accordingly move great distances to reach attractive sources of food. Migrations of Burchell's zebras to succulent pastures during the rainy season are a feature of the Serengeti Plains and the Etosha National Park in Africa. The distribution of asses, half-asses, and horses inhabiting arid areas largely follows that of rainfall and pasture.

The food of the browsers is fairly readily available throughout the year; thus, species in this category are relatively sedentary. The browsing rhinoceroses may break down trees and shrubs, and use their forelimbs to help get at otherwise inaccessible leaves and twigs. Food is plucked with the lips. In the tapirs, the upper lip is fused with the short proboscis. The rhinoceroses (excluding the white) have a pointed upper lip with a fingerlike process that is used to pluck leaves and twigs. The white rhinoceros, with its broad square muzzle, is the most specialized grazing rhinoceros, feeding on grass.

Herd behaviour among equids

Social organization and territory. In both the mountain and Burchell's zebras the family group is the basic social unit. It generally consists of a single adult male and two or three adult females with their foals. The groups are stable, apparently because of strong mutual ties among the females rather than because of herding by the male. The stallion is dominant, and there is a hierarchy among the mares, the highest ranking (alpha) animal usually leading the group. Other males are either solitary or live in bachelor groups of two or three, sometimes up to ten. Juveniles leave the family group when they attain sexual maturity at one and one-half to two years. When large aggregations occur on favoured grazing grounds, the groups retain their identity. Among Grevy's zebras and wild asses, territorial males, mixed groups of mares and foals and of stallions are found. There is no evidence for territorial behaviour among any of the zebras except Grevy's. Individual groups occupy home ranges that overlap to some degree with those of other groups.

The social organization of other equids is not as well documented. Observers studying the wild horse and half-asses have noted that females and juveniles form a group dominated by a single stallion, which keeps them together by active herding; the unattached males are solitary or live in small herds.

The pattern of social organization among the rhinoceroses is quite different from that of the Equidae. Dominant adult males of the white rhinoceros occupy territories that, in the Natal reserves, average about 200 hectares

(500 acres). Within its area a male may tolerate subadult or aged bulls, which have subordinate status. Adult females accompanied by their calves inhabit home ranges encompassing the territories of six or seven dominant bulls. Juveniles consort with other juveniles or with calfless females, but groups of more than two usually do not stay together long.

The black rhinoceros is basically solitary. Adults of both sexes usually occupy home ranges of 1,000 hectares (2,500 acres) or more, the size depending on the characteristics of the environment; occasionally the ranges may be as small as 200 hectares, however. There is a good deal of overlap in the utilization of these home ranges.

The Asiatic rhinoceroses also are essentially solitary, but detailed information on the nature of the areas they inhabit is not available. Individual great Indian rhinoceroses are said to occupy tracts as small as 8 to 20 hectares (20 to 50 acres).

Little is known about the social organization or territorial behaviour of the tapirs. All species are reported to be found alone or in pairs.

Scent marking and territories

Male zebras and horses follow mares in estrus. The stallion, after smelling the spot where a mare has urinated or defecated, exhibits "flehmen" (a characteristic display in which the head is lifted and the upper lip raised) and then urinates or defecates on the same spot. In similar fashion, members of stallion groups often urinate or defecate consecutively; communal dung heaps formed by five to eight animals often arise in this way. The significance of such behaviour is not clear.

Among the Rhinocerotidae excretory products play an important role in marking territories and home ranges. Dominant male white rhinoceroses defecate almost entirely on heaps within their territories. They then scatter the material by kicking vigorously, presumably leaving an individual scent mark in this manner. In addition, they urinate in a ritualized fashion, spraying the urine in powerful jets in a manner peculiar to them and shown by no other sex or age group. Other members of the population also use dung heaps (either in territories or in communal areas, such as along paths) but not exclusively, and they do not scatter dung.

In the black rhinoceros, dung-scattering behaviour does not appear to be exclusive to dominant males. The function of the communal heaps may be mainly to establish the presence of the inhabitant in his home range, and to maintain contact between known animals.

Dung heaps and urine spraying are also observed among other species of rhinoceroses and among tapirs; their significance is presumably of a similar nature.

Fighting. The pattern of fighting is related to the amount of lethal equipment the various groups possess. The Equidae, unarmoured, do not employ stylized fighting techniques to reduce the danger of serious injury—as among certain other species. Fighting is largely confined to adult males competing for estrous mares. Various techniques occur in the zebras, which may serve as an example of the family. Circling, neck fighting, biting (either in a standing or sitting position), rearing combined with biting and kicking, and kicking on the run all are used, either alone or in combination. No set pattern is followed.

Fights among rhinoceroses consist of charges and striking with the horns, usually accompanied by vocal threats. Goring is not common, the stylized pattern having probably been evolved to minimize the danger of serious injury from the formidable horns.

Mutual grooming and greetings

Amicable behaviour. Mutual grooming is well known among horses. Two animals stand facing in opposite directions and groom each other by nibbling at the root of the tail and the base of the neck. Burchell's zebra behaves similarly and so, presumably, do other members of the family.

Zebras greet each other simply by nose-to-nose contact, except that adult stallions go through a ceremony involving nose-to-genital contact. Nose-to-nose greeting is also characteristic of tapirs and rhinoceroses. The latter also rub their bodies together.

Courtship and mating. Courtship is relatively simple among the social equids. The true ass is apparently exceptional. The partners are strangers when the first approaches are made and the female requires violent subjugation by the male, which bites, kicks, and chases her before she will stand for him. This may be the result of separation of the sexes outside the mating season. The wild horse and Burchell's zebra are not at all violent. The stallion often grooms the mare before attempting to mount. The estrous mare (especially, or exclusively, young mares in the case of Burchell's zebra) adopts a typical posture with legs slightly apart, tail lifted, and, except in the horse, a characteristic facial expression (the "mating face" already mentioned).

The more or less solitary rhinoceroses and tapirs go through a more elaborate courtship, presumably because the partners are strangers. After a chase, the male and female may engage in low-intensity fighting, ending with the male laying his head on the female's rump and then mounting and copulating for an extended period. Several males may mate with an estrous female.

Relations between parent and offspring. The perissodactyls bear well-developed (precocial) young, usually a single offspring. After the mother has assisted in removing the placenta and has licked her offspring in the usual mammalian fashion, the young animal soon attempts to stand. A Burchell's zebra foal has been observed to stand quite firmly 14 minutes after birth, and a black rhinoceros calf 25 minutes after birth.

Newly born equids follow any nearby object during the first few days of life. At this time, zebra mares drive away all other zebras from their foals. The behaviour ensures that the foal will form a bond with its mother during the initial period of imprinting. Foals follow their mothers closely and are groomed frequently.

Although precocial, black rhinoceros calves appear to have a lying-out period; that is, an initial period when they rest quietly in thick cover except when being suckled. Thereafter they follow their mothers closely. A young white rhinoceros tends to walk ahead of the mother and may be guided by her horn.

Most young perissodactyls remain with their mothers until the next offspring is born. A young rhinoceros may, therefore, accompany its mother until it is two and one-half years old, or older. Although grazing starts early, suckling proceeds for a considerable time, perhaps for its psychological rather than its physiological value.

Play. As in other mammals, play is a prominent form of behaviour among young perissodactyls. Zebras up to the age of one year frequently engage in running games. Foals gallop wildly about on their own, jumping and kicking up their heels, sometimes chasing other animals, such as gazelles, mongooses, or birds. In groups they play catching games, running after one another in close succession. Mock fighting sometimes takes place. Groups of adults have also been seen to chase foal groups in play, and indeed stallion groups carry out playful gallops. Stallions also engage in play greeting and in mock fights.

Running games and mock fighting

Playful romping and mock fighting with the horns are common among rhinoceros calves. Young tapirs play running games.

Rolling and wallowing. Behaviour for the care of the body is widespread among the perissodactyls. Equids frequently roll in dry, loose soil forming rolling hollows—a common feature of zebra country.

Wallowing, which may help regulate body temperature, probably is mainly a form of self-grooming; it is practiced by all species of rhinoceroses. They often spend hours lying in pools during the middle of the day in hot weather. Mud of suitable consistency induces wallowing, which may be followed by sand bathing. Prolonged rubbing on tree trunks or suitable stumps follows a wallow; old rubbing stumps and stones may take on a shine from repeated use.

Tapirs may have the most pronounced tendency to bathe and wallow, but few details of their behaviour are known. They are also said to enter water when disturbed.

Reproduction. Female equids of all the species for which information is available attain puberty at about one year, but are not normally successfully mated before the age of two to two and one-half years, and possibly as

late as three to four years in the case of Grevy's zebra. Zebras probably breed until about 20 years of age. The domestic species are seasonally polyestrous (repeatedly fertile), coming into breeding condition in spring and, unless mated, undergoing repeated estrous cycles at intervals of approximately three weeks until the end of the summer. The wild species studied also tend to mate seasonally; most young are born in spring and summer.

Reproductive potential of equids

The gestation period of equids is between 11 and 12 months. In most species a postpartum estrus occurs, usually within two weeks of the birth of the young; thus, the maximal potential reproductive rate is one young per year. This potential is not always attained. Only about 50 percent of domestic mares that are mated produce foals, and nearly half of a study group of Burchell's zebra mares bore only one foal in three years.

The gestation period of three species of rhinoceroses is about 15 to 17 months. For the Sumatran rhinoceros the period is said to be only seven months. No information is available for the Javan rhinoceros. Female white and black rhinoceroses attain sexual maturity at the age of four to five years and are capable of calving at intervals of approximately 2½ years. Rhinoceroses probably breed until between 30 and 40 years old. The white tends to have a mating peak in spring, corresponding with the flush of green grass, and a calving peak in autumn.

The Malayan and Brazilian tapirs have gestation periods of 13 months' duration. The Brazilian tapir is reported to mate before the onset of the rainy season.

Drawing by R. Keane based on (A, B, C, D) A.S. Romer, *Vertebrate Paleontology*, Copyright 1933, 1945, 1966 by the University of Chicago; all rights reserved

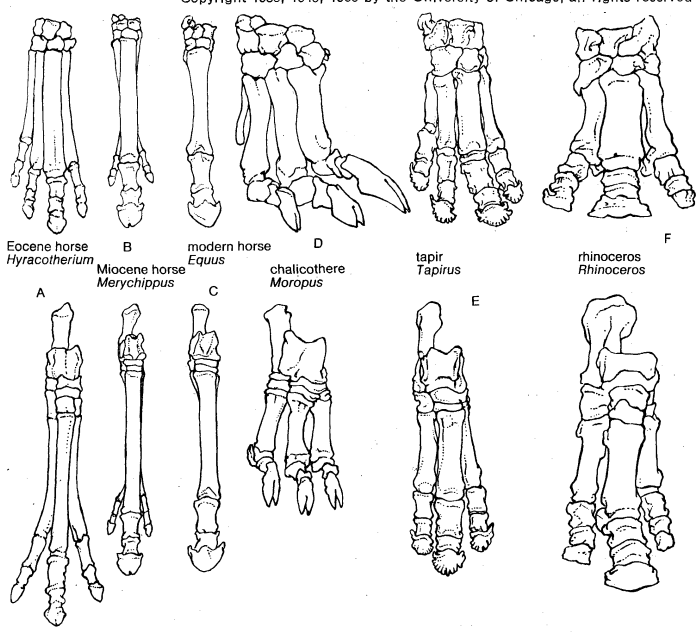


Figure 2: Forelimbs (above) and hindlimbs (below) of some recent and fossil perissodactyls.

FORM AND FUNCTION

Integument. The skin of rhinoceroses is extraordinarily thick. The great Indian and Javan rhinoceroses are covered with large, practically immovable plates, separated by joints of thinner skin to permit movement. The two species differ in the arrangement of the folds. The hair of all rhinoceroses is sparse or absent except that of young Sumatran rhinoceroses, which have a dense coat of crisp, black hair. The skin of the tapirs is also thick with a sparse covering of short hairs arranged in irregular groups. The equids have a normal hide with a well-developed hairy coat.

Structure of rhinoceros horns

The "horns" of the rhinoceroses are noteworthy structures of epidermal origin. The horn is composed of a mass of fused epidermal cells that are impregnated with a tough, fibrous protein (keratin) and that rest on a roughened bony cushion on the fused nasal bones. The male Javan rhinoceros has a short horn about 25 centimetres (10 inches) long; the horn of the female is rudimentary. The great Indian rhinoceros has a single horn up to 60 centimetres (25 inches) long. The other species of rhinoceroses have a second horn that stands on a protuberance of the bones (frontals) between the eyes. Hornlike structures were also present in the titanotheres and extinct rhinoceroses.

In all living perissodactyls the terminal digital bones are flattened and triangular, with evenly rounded free edges, and are encased by keratinous hooves derived from the integument. The single hoof of the equids—the only mammals to walk on the tips of single digits—is the most highly developed structure of this kind among mammals. The keratinous wall is analogous to the nail of mammals that have claws or nails.

Skeleton. *Backbone.* The vertebral column acts as a firm girder, with high dorsal (neural) spines on the thoracic vertebrae, above the forelimbs and ribs. Spines and ribs serve as compression struts above and below. The column balances largely on the forelegs and is pushed from behind by the hindlegs, which are the main propellants. This skeletal structure permits running and also enables great weights to be borne in such animals as the rhinoceroses. There are never fewer than 22 thoracolumbar (trunk) vertebrae.

The neck, or cervical, vertebrae are opisthocoelous; *i.e.*, with the bodies (centra) of the vertebrae hollowed behind to take the convex heads of the succeeding centra. This feature facilitates rotatory movement of the neck and is most highly developed in the horses.

Limb girdles. The shoulder blade is long and narrow with a small coracoid process (a ridge to which muscles are attached) and a low spine. There is no clavicle (collarbone). The pelvic girdle has a broad, vertically raised ilium to which are attached the large gluteal (thigh) muscles, important for locomotion, and the abdominal muscles, which carry the weight of the belly.

Limbs. There is a clear evolutionary tendency in the Equidae for the limbs to become long and slender, with a reduction in the number of digits in the swift-running forms. These changes are accompanied by an increase in rigidity and specialization for movement fore and aft. The upper (proximal) segments, the humerus of the forelimb and the femur of the hindlimb, have remained short. In contrast the lower (distal) parts, consisting of an anterior radius and posterior ulna in the forelimb and an anterior tibia and posterior fibula in the hindlimb, have become longer and slimmer. The humerus is short and broad. Its articulation with the radius and ulna only permits fore-and-aft movement. The proximal end of the stout femur has a third projection, or trochanter, in addition to the usual mammalian two, which serves as an additional point of attachment for the large locomotory muscles of the hindlimbs.

Evolution of swift-running forms

The anatomical feature of the order now considered most significant is that the axis of symmetry of the limbs passes through the third or middle toe, the most strongly developed and the one on which most of the weight is borne. This is called the mesaxonic condition and is contrasted with the paraxonic condition of the Artiodactyla, in which the axis passes between the third and fourth toes.

Originally the five toes of the limb were held in the semidigitigrade position—*i.e.*, with the weight of the body being borne on the soles of the toes and on the lower ends of the elongate metacarpal and metatarsal bones of the forefeet and hindfeet. The upper (proximal) ends of these bones were raised above the ground, a condition still to be seen in the tapirs.

As the third digit became increasingly dominant, it became longer and thicker. The upper ends of the third metacarpus and metatarsus broadened and forced the other digits to the side. The first (inner) digit was the first to disappear. The earliest known forms already bore only three digits on the hindfoot. The loss of the first toe on the front foot led to the four-toed condition common in the Eocene; the fifth digit persisted, although it was somewhat weak. The tapirs, with four toes in front and three behind, have retained this early condition.

The fifth digit of the hindfoot was the next to disappear

in the evolutionary sequence, and all known forms beyond the lower Oligocene had only three toes in each foot. The living rhinoceroses illustrate this condition. A massive limb with a broad foot is essential in such heavy animals. The feet have cushions of elastic connective tissue well suited to bear the weight of the body.

In the most highly specialized forms, the second and fourth digits also underwent reduction. These digits are retained in the living Equidae only as functionless, vestigial slivers of bone on either side of the third metacarpal and metatarsal.

Growth of the third digit's terminal joint

With reduction in the number of digits, the third has assumed a progressively more vertical position. Its terminal joint, or phalanx, has become larger and the hoof surrounding it bigger and thicker. At the same time the ulna, the smaller bone of the forelimb, decreased in size until, in the modern Equidae, its upper end fused with the radius and its lower end remained merely as part of the articulating surface of the radius with the wrist bones (carpals). In the hindlimb, the fibula became reduced in similar fashion. It articulates with the ankle bones (tarsals) only in a few extinct forms, such as the titanotheres. In the tapirs and rhinoceroses it is slender. In the equids the proximal end remains as a small splint of bone, while the distal end has fused with the tibia.

The arrangement of the small bones of the carpus ("wrist") and tarsus (ankle) is another characteristic feature of perissodactyl limbs. In most other mammalian orders, the carpals and tarsals provide the limbs with regions of flexibility. Their development in the perissodactyls has been toward increasing rigidity, following the trend in the limbs as a whole. The basic mammalian arrangement is one of three rows of bones: a proximal row of three, adjacent to the lower arm and leg bones; an intermediate row of four (the centralia); and a distal row of five (carpalia of the forefoot, tarsalia of the hind), one to each digit. The centralia, carpalia, and tarsalia are numbered one to four or five, beginning on the side of the thumb and big toe. In all orders the number of wrist and ankle bones has been reduced. Usually only one of the centralia, known as the navicular, remains, while there are the same number of carpalia and tarsalia as there are digits.

An interlocking plan is characteristic of the Perissodactyla. In the forefoot, the third distal carpal (the magnum or capitate) is enlarged and interlocks with the proximal carpals. The elongated third metacarpal thrusts up against these interlocked bones. In the equids, distal carpal I (the trapezium) is absent, and the arrangement in the hindfoot is similar. In the most advanced, modern forms metatarsus III thrusts against the enlarged and flattened tarsal III (ectocuneiform), and this in turn is in contact with the large, flat navicular (centrale II and III). The navicular abuts on the flattened astragalus (or talus), the intermediate bone of the proximal row. The articulation between the upper surface of the astragalus and the tibia is pulley-like and permits only fore-and-aft movement of the limb.

Teeth. The full complement of mammalian teeth consists of three incisors, one canine, four premolars, and three molars in each half of each jaw. The arrangement may be expressed by the formula $\frac{3 \cdot 1 \cdot 4 \cdot 3}{3 \cdot 1 \cdot 4 \cdot 3} = 44$ teeth. The figures represent the number of incisors, canines, premolars, and molars in each half of the upper (above the line) and lower (below) jaws, respectively.

Fully-toothed early hoofed mammals

The Condylarthra, a group of mammals that first appeared in the Paleocene (about 65,000,000 years ago) and were ancestral to most of the later and recent hoofed mammals, had a full complement of teeth. In many of the early perissodactyls, only the first lower premolar had been lost. The subsequent evolutionary sequence led to losses and specializations of the incisors and canines. Lengthening of the facial part of the skull resulted in the formation of a gap, the diastema, between the incisors and the premolars. The first upper premolar was reduced or lost in consequence. The canines, when present, were situated in this diastema, as they are in male horses.

Among the living perissodactyls, the tapirs have the least specialized battery of teeth. In this, as in many other features, they have remained primitive. The dental formula of the family Tapiridae is: $\frac{3 \cdot 1 \cdot 4 \cdot 3}{(2\text{–}3) \cdot 1 \cdot 3 \cdot 3} = 40\text{–}42$ teeth. The first upper premolar is noteworthy in being the only premolar with a milk, or deciduous, predecessor.

The cutting teeth are reduced in the Rhinocerotidae. Incisors and canines are absent in the two African forms, although rudimentary incisors occasionally are found in young animals. The Asiatic species have one or two upper, but generally no lower, incisors in each half of the jaw. They have no upper canines; the Javan and Sumatran rhinoceroses have one short, sharp lower canine on each side. There are three upper and lower molar teeth in all five species. The white and Sumatran rhinoceroses have three premolars, and the others have three or four premolars. The dental formula for the family is set up, therefore, as follows: $\frac{(0\text{–}2) \cdot 0 \cdot (3\text{–}4) \cdot 3}{0 \cdot (0\text{–}1) \cdot (3\text{–}4) \cdot 3} = 24\text{–}30$ teeth. The dental formula of the Equidae is $\frac{3 \cdot 1 \cdot (3\text{–}4) \cdot 3}{3 \cdot 1 \cdot 3 \cdot 3} =$ 40–42 teeth. The canines are rudimentary in females and the first upper premolar is extremely small, often falling out.

The form of the premolars and molars is of great interest and their evolutionary history has been studied in some detail. Primitive, browsing members of the order had brachydont cheek teeth (*i.e.*, with low crowns and long, narrow root canals), with separate low, rounded cusps—the bunodont condition. Increasing specialization for grazing resulted in fusion of the cusps into ridges (lophs), thus teeth of this kind are called lophodont. Lower molars typically have two transverse lophs, the protoloph and the metaloph. In the upper molars these ridges are fused with a longitudinal ridge (ectoloph), which runs along the outer edge of the tooth. Further development leads to a convoluted arrangement of the lophs, such teeth being termed selenolophodont. Two extinct perissodactyl groups did not follow this trend. In the titanotheres the inner cusps remained bunodont, and in the chalicotheres there was no protoloph.

Associated with these changes in the tooth surfaces is a tendency for the crown to become higher. High-crowned teeth are termed hypsodont. Their roots remain open until a fairly advanced age, a condition associated with continuous growth, which compensates for wear. The hollows between the lophs of hypsodont teeth are filled with a deposit of secondary cement, which strengthens the teeth and makes them more resistant to wear. A further evolutionary trend is for premolars to become as large as molars. Where the process of molarization is complete, as in horses, all grinding teeth are identical.

From *Horses: The Story of the Horse Family in the Modern World and Through Sixty Million Years of History*, by G.G. Simpson. Copyright 1951 by Oxford University Press, Inc. Reprinted by permission

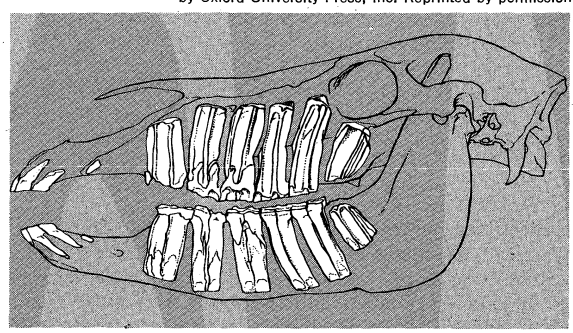


Figure 3: Skull and dentition of horse, *Equus caballus*, cut away to show roots of teeth.

The Tapiridae have primitive brachydont premolars and molars. They possess two simple transverse lophs, and are thus termed bilophodont. They lack secondary cement on the crowns.

The Sumatran rhinoceros, the most primitive of the living rhinoceroses, and the Javan rhinoceros have similar brachydont, lophodont cheek teeth. The great Indian rhinoceros, which is less of a specialized browser, has hypselodont (hypsodont and selenodont) premolars, with a layer of cement on the crowns. The black rhinoceros

has brachydont and lophodont teeth, with a thin layer of cement. The white rhinoceros is more specialized, for the cheek teeth are hypselodont and have a thick cement layer.

The grinding teeth of the Equidae are highly specialized, high crowned, with a complicated selenodont surface and thick cement deposits.

Comparative size of the stomach

Digestive system. The stomach of perissodactyls is small, simple, and undivided. In the horse its capacity is only 8.5 percent of the whole digestive system. The comparable figure for the ox is 71 percent. The intestine is very long and the cecum (blind gut) and colon are huge and sacculated (*i.e.*, with many blind pockets). Here food is macerated and fermented and the fibrous portions are dissolved. The liver has no gall bladder.

EVOLUTION AND PALEONTOLOGY

The Perissodactyla appeared early in the Eocene, about 55,000,000 to 40,000,000 years ago. Together with most other ungulate mammals, they were probably derived from the Condylarthra. The condylarths were abundant in Europe and North America, mainly during the Paleocene (65,000,000 to 55,000,000 years ago). Condylarths were unspecialized mammals, rather carnivore-like in appearance. The larger species attained the size of tapirs. The limbs were fairly short and primitive but the third toe was somewhat enlarged and the phalanges ended in hooves. There was a full complement of teeth with some specializatin of bunodont molars for herbivorous diet.

Horses. The earliest horses appeared during the lower Eocene in Europe and North America. They are generally known as *Eohippus* ("dawn horse"), but *Hyracotherium* is the correct taxonomic designation. Some species of these little forest-dwelling, browsing animals were no larger than a terrier. They had moderately long, slender limbs with only four toes in the forefoot and three in the hindfoot, all equipped with hooves. The molars were essentially bunodont (with low, rounded cusps) and the premolars simple.

Hyracotherium-like animals persisted in Europe until the end of the Eocene. Another group, the paleotheres or "native" European horses, evolved as a specialized side branch, which died out in the Oligocene. North America was the centre of horse evolution. During the Eocene, *Hyracotherium* was succeeded by forms such as *Orohippus* and *Epihippus*, which are known only from that epoch. They were notable only for minor advances, such as the development of a molar configuration in two premolars.

The Oligocene (40,000,000 to 26,000,000 years ago) saw a major change with the appearance of three-toed horses, *Mesohippus*, *Miohippus*, and others. All of the premolars were similar to the molars, low-crowned but lophodont (ridged). *Anchitherium* was an early Miocene form as large as a modern pony, which migrated from North America to Europe. These primitive three-toed horses or anchitheres survived until Pliocene times, some of their descendants attaining the size of a rhinoceros, as in the late giant representative *Megahippus*.

The main course of horse evolution entered a third stage in North America in the Miocene Period (26,000,000 to 7,000,000 years ago). A line of grazing horses developed, almost certainly to exploit the new grasslands that were spreading over the surface of the earth. The degree of lophodonty of the molariform teeth increased, changing the pattern of the crests on the surface and increasing their grinding efficiency. Of greater importance, these teeth became hypsodont (high-crowned) and thus maintained a good grinding surface as grinding of the harsh, siliceous grass caused them to wear down. Another substance, cement, came to supplement the dentine and enamel forming the teeth of earlier types, and provided additional material to resist abrasion. The evolution of these specialized teeth was a tremendous advance.

Trends in horse evolution

The limbs of the grazing horses became increasingly rigid and specialized for fore-and-aft movement, better fitting the animals for running in open country. In *Merychippus* the ulna was fused with the radius and the fibula was much reduced. In some advanced forms the central toe was much larger than the two lateral toes and carried most of the weight of the body on a hoof much like that of modern horses.

A number of evolutionary lines developed during the Pliocene, which lasted from 7,000,000 to 2,500,000 years ago. *Pliohippus* of North America is probably the line from which modern horses have come. The genus *Equus* is characteristic of the Pleistocene when it developed in North America and spread to all continents except Australia. By the end of the Pleistocene, horses had become extinct in the New World.

Titanotheres. Another group entirely, the titanotheres (Brontotheriidae), evolved independently from *Hyracotherium*-like ancestors and became abundant in North America during the Eocene and Oligocene, but disappeared by the Miocene. They were also found in Asia and Eastern Europe. The end forms, such as *Brontops* and *Brontotherium*, were huge, the largest standing 2½ metres (8 feet) at the shoulder. They had long, low skulls and a small brain. Many species bore a pair of large, hornlike processes on the front of the head.

By courtesy of the Field Museum of Natural History, Chicago



Figure 4: Reconstruction of fossil Perissodactyla. *Brontops*, an Oligocene titanothere, some individuals of which were about 2.5 metres (eight feet) at the shoulder. Detail of a painting by Charles R. Knight.

Chalicotheres. The chalicotheres (Chalicotheriidae) were moderately large animals that appeared in Eurasia and North America during the Eocene. Thereafter they evolved mainly in the Old World, disappearing from America in the mid-Miocene, but persisting in Asia and Africa until they died out in the Pleistocene. Early members of the group such as *Paleomoropus*, from the lower Eocene, resembled contemporary equids. The Miocene *Moropus* typifies the peculiar characteristics of later forms. It was horselike, but the front legs were longer than the rear. Each foot bore three toes which ended in large, fissured phalanges bearing claws that may have been used for digging up roots and bulbs.

Tapirs. The fossil and living tapirs, together with similar extinct forms such as the lophiodonts, constitute a small, fairly uniform group.

Homogalax, *Lophiodon* and other tapir-like animals appeared in the Eocene. Some were close to the ancestry of both rhinoceroses and tapirs, and other evolutionary lines left no descendants. *Protapirus* from the Oligocene of Europe and North America was the forerunner of the modern tapirs. These tapirids persisted until the Pleistocene, when climatic changes led to their extinction.

Rhinoceroses. The fossil history of the rhinoceroses is far more complex. They evolved from the early tapiroids but diverged from that group. Unlike the equids, they have tended to grow large, with short, stout limbs with

little reduction in the digits. Although the premolars have tended to molarize, the molar cusp pattern is simple; the teeth seldom become hypsodont, and cement occurs rarely.

The Hyracodontidae, or running rhinoceroses, from the Eocene and Oligocene of North America and Asia were the most primitive. *Hyracodon* had long, slender legs with three toes on each foot and typically rhinocerotid cheek teeth.

The amynodonts are known from the late Eocene and Oligocene of Eurasia and America and lived in Asia until the Miocene. They were a side branch, perhaps derived from primitive hyracodonts. *Metamynodon* and some other forms were about as large as hippopotamuses and may have lived in rivers. The premolars were simple and the incisors reduced, but canines and molars were enlarged.

True rhinoceroses are probably also descended from early hyracodonts. They became numerous in the Oligocene as large animals with molariform premolars. Many side branches evolved. *Trigonias* still had four front toes but *Caenopus*, another Oligocene representative, had the three toes common to all later rhinoceroses. Like its relatives of that period, *Caenopus* was fairly small, about the size of a tapir, and hornless.

One spectacular group contained giant hornless creatures, such as *Baluchitherium* and *Indricotherium* from the Oligocene and Miocene of Asia; the largest of known land mammals, they stood 5.5 metres (about 18 feet) high at the shoulder and had a long neck and long forelegs.

Rhinoceroses died out in North America during the Pliocene, but in Eurasia many different species survived to the Pleistocene. One of the most familiar of these later rhinoceroses is the woolly rhinoceros (*Coelodonta*), which was depicted by Stone Age artists and is known from nearly intact specimens. The surviving rhinoceroses represent remnants of once varied and abundant stock; the genera have little direct relationship to one another.

CLASSIFICATION

Importance of skeletal features

Distinguishing taxonomic features. Skeletal features are of greatest importance in classifying the Perissodactyla. They are, of course, the only criteria applicable to fossil forms. Distinguishing characteristics of the skull include the relative length of the facial region, length and form of the nasal bones, presence of a postorbital bar and of hornlike structures. The form and number of teeth, presence of a diastema, number of molariform premolars, the height of the cheek teeth, their cusp structure and the presence or absence of cement on the grinding surfaces all are particularly valuable taxonomic features. The limb structure is also a useful aid to classification, especially the length of upper and lower portions, the degree to which the ulna and the fibula are reduced and fused, respectively, with the radius and the tibia, the number and form of carpal and tarsal bones, and the number and relative size of the digits.

Among living perissodactyls, body size, form of the upper lip, number and length of horns, structure of the skin and colour pattern are the most important features for classification.

Annotated classification. The classification presented here follows that of U.S. paleontologist George Gaylord Simpson, which is generally accepted but modified in some fairly minor ways by other authors. The term Mesaxonia, introduced by the 19th century paleontologist O.C. Marsh, is essentially synonymous with Perissodactyla. Simpson used it as a convenient designation for the superorder containing the single order Perissodactyla, as he used the parallel name Paraxonia for the superorder with the one order Artiodactyla.

Groups indicated by a dagger (†) are known only as fossils.

ORDER PERISSODACTYLA

Herbivorous ungulates with either 3 digits or 1, at least in the hindfeet. Early forms digitigrade, this feature either retained or completely replaced by the unguligrade condition in later representatives. The body weight borne mainly or entirely on the third digit through which the long axis of the limb passes (mesaxonic condition). The talus (heel bone) with only 1, proximal, keeled surface, articulating with the tibia, no duplication of keel on the distal surface, as in artiodactyls. Nasals broad at the posterior end, alisphenoid canal present. Posterior premolars molariform; cheek teeth bunodont in early forms, typically lophodont or selenolophodont, rectangular; wide diastema separates them from incisors and canines, which may be reduced or absent. Testes inguinal, occasionally scrotal; mammary glands inguinal, with 2 teats; uterus bicornuate. Fifteen living species, but more than 200 fossil forms.

Suborder Hippomorpha

Superfamily Equoidea

Dentition complete, upper molars with 6 tubercles, the 2 external ones united to form an ectoloph, median and internal tubercles generally fused into a single loph. Tendency to molarization of premolars and reduction of lateral digits.

†*Family Palaeotheriidae* (paleotheres or "native horses"). Lower Eocene to lower Oligocene; Europe and Asia. Size variable, from that of fox (shoulder height about 40 centimetres or 16 inches) to that of a rhinoceros; body form tapirlike. Three digits in each foot. Dentition complete or first premolars absent; premolars molariform, separated from canines by diastema. Cheek teeth brachydont, lophodont, with or without cement. Orbits without postorbital bar. Eurasian "native horses," seven genera.

Family Equidae (horses and relatives). A progressive group with tendency to molarization of premolars and development of cement.

†*Subfamily Hyracotheriinae* (hyracotheres). Eocene: Europe and North America. Size about that of a fox; 4 toes in forefoot, 3 in hindfoot. Dentition complete, premolars simple or last 2 slightly molariform, molars brachydont, bunodont, diastema short. Skull short, without postorbital bar. Three genera. Basic stock from which other members of family evolved.

†*Subfamily Anchitheriinae* (anchitheres). Lower Oligocene to lower Pliocene. Fossils from North America, Europe, Asia. Browsing horses. Early forms collie-sized (shoulder height about 60 centimetres), end forms as large as a rhinoceros. Three toes in each foot, central toe somewhat enlarged; metapodials elongated. First premolars reduced, second to fourth molariform. Cheek teeth brachydont, lophodont; ectoloph (longitudinal outer crest) above typically W-shaped cross-lophs retained distinct cusps; hypostyle at back edge of upper molars prominent. Lower cheek teeth with 2 crescent-shaped lophs and double cusp at junction. Diastema long, no postorbital bar. Six genera.

Subfamily Equinae (horses, asses, half-asses, and zebras). Lower Miocene to Pleistocene of North America; Pleistocene of South America; lower Pliocene to Recent of Europe, Asia, and Africa. Most forms as large as ponies or modern horses. Feet with 3 toes in early forms, 1 in later forms (but metapodia 2 and 4 persist as remnants), limbs with progressively elongated lower segments and reduced or fused ulna and fibula. Calcaneum without articulating surface for fibula. First premolars reduced or absent, others molariform; cheek teeth hypsodont, selenolophodont, with cement; canines in long diastema. Skull long, with postorbital bar in upper Miocene and later forms; nasals long and narrow. Some Recent forms with striped coat.

†*Superfamily Brontotherioidea*

†*Family Brontotheriidae* (titanotheres). Lower Eocene to lower Oligocene. Fossils from North America and Asia. Early forms as large as tapirs, end forms huge, standing 2.5 metres at the shoulder. Graviportal browsers; 4 toes in forefoot, 3 in hindfoot. Dentition complete, premolars small and simple in early forms; later forms with incisors and first premolars reduced or absent, remaining premolars only partly molarized. Molars large, bunolophodont, above with W-shaped ectoloph as in early equids, lower with double V resembling equid pattern but lacking reduplication of cusp at union of V's. Skull elongated, low, without postorbital bar, with large rough, hornlike process in later forms. About 39 genera.

†Suborder Ancylopoda

†*Family Chalicotheriidae* (chalicotheres). Upper Eocene to lower Pliocene. Fossils from North America, Europe, Asia. Early forms equivalent in size (and similar) to contemporary horses, *Hyracotherium*; later representatives larger, up to size of a modern horse. Most with forelegs longer than hindlegs and 3 toes in each foot; rudimentary metacarpal V retained in forefoot, most with clawlike digits. Incisors or canines reduced or absent; first premolar absent, remainder small and simple. Molars bunolophodont, cusp pattern similar to Brontotheriidae. About 13 genera. Aberrant perissodactyls which may have used claws to dig for bulbs and roots.

Suborder Ceratomorpha
Superfamily Tapiroidea

Brachydont forms, molars with simple ectoloph and strongly developed transverse protoloph and metaloph. Forefoot generally with 4 toes, hindfoot with 3.

†*Family Isectolophidae.* Lower to upper Eocene. Fossils from North America and Asia. Limb structure similar to that of *Hyracotherium*, teeth already tapiroid. Four genera.

†*Family Helaletidae.* Lower Eocene to middle Oligocene. Fossils from North America and Asia. Skull advanced in structure; premolars much simpler than molars, molars of simple lophodont form. Ten to 12 genera.

†*Family Lophiodontidae* (lophiodonts). Lower to upper Eocene; Europe and probably Asia. Skull not tapir-like, lophs of molariform teeth oblique. About 6 genera.

Family Tapiridae (tapirs). Lower Eocene to Recent. Fossils from North and South America, Europe, Asia, Africa. Recent species in Central and South America and Southeast Asia. Limited to tropical equatorial forests of South America and Asia. Browsers. Skull greatly modified; premolars completely molarized, first sometimes absent; transverse lophs strongly developed in molariform teeth, ectoloph simple. Limb structure primitive; 4 toes in forefoot; 3 in hindfoot. Six genera but only 1 extant.

Superfamily Rhinocerotoidea

Molars brachydont, upper with 2 transverse lophs fused with a well-developed ectoloph; lower with 2 asymmetrical crescents. Premolars more or less molariform. Rough processes on nasals (for horns) frequently present; forefoot with three or four toes.

†*Family Hyrachyidae.* Lower to upper Eocene; North America and probably Asia. Small animals; forefoot 4-toed; molariform teeth with separate tubercles. Four genera.

†*Family Hyracodontidae* (hyracodonts or running rhinoceroses). Middle Eocene to upper Oligocene. Fossils from North America and probably Asia. Relatively small; feet 3-toed, legs long and slender. Molariform teeth typically rhinocerotid. About 7 genera.

†*Family Amynodontidae* (amynodonts). Upper Eocene to middle Oligocene. Fossils from North America, Europe, Asia. Hippopotamus-like; forefoot 4-toed, hindfoot with 3 toes. Skull short, heavy, premolars and incisors reduced, premolars simple, canines and molars enlarged. Six genera.

Family Rhinocerotidae (rhinoceroses). Middle Eocene to present. Fossils from North America, Europe, Asia, and Africa. Now restricted to tropical Africa and Asia.

Large; most forms 3-toed. Skull elongated, raised at posterior end. Dentition incomplete, upper canines always absent, some incisors usually enlarged as cutting teeth; premolars molariform, last upper molar simple. About 34 genera, 4 extant.

Status of the chalicotheres

Critical appraisal. Systems of classification frequently differ in the status given to the chalicotheres. Simpson takes the view that their ancestry was probably equid and almost certainly hippomorph. He holds that the significance of their claws has been over-emphasized and has tended to distract attention from their true affinities. Accordingly he places them in the superfamily Chalicotheroidea of the suborder Hippomorpha. In the classification presented above they are given their own suborder, Ancylopoda, following the views of Alfred S. Romer, another authority on the group.

The forms united in the family Lophiodontidae by Simpson, followed here, are thought by some recent workers to warrant separation into three families. Romer distinguishes the families Lophialetidae, Deperetellidae, and Lophiodontidae. The affinities of certain primitive genera such as *Hyrachus* and *Colonoceras* remain controversial; Simpson places them in the family Hyrachyidae, superfamily Rhinocerotoidea. Romer considers them to be early tapiroids and assigns them to the Helatidae (superfamily Tapiroidea). The difference of opinion is slight, for it is generally agreed that the hyrachids are close in the common stem of the tapiroids and rhinocerotoids.

BIBLIOGRAPHY. Information on the ungulates as a group is contained in J. SIDNEY, "The Past and Present Distribution of Some African Ungulates," *Trans. Zool. Soc. Lond.* vol. 30 (1965), a comprehensive account, both geographical and temporal; and J.R. ELLERMAN and T.C.S. MORRISON-SCOTT, *Checklist of Palaearctic and Indian Mammals, 1758 to 1946*, 2nd ed. (1966).

The following papers are concerned with the biology of certain perissodactyls. W. VON RICHTER, "Untersuchungen über angeborene Verhaltensweisen des Schabrackentapirs (*Tapirus indicus*) und des Flachlandtapirs (*Tapirus terrestris*)," *Zool. Beitr.* Neue Folge, 12:67–159 (1966), one of the few detailed studies of tapirs; C.A.W. GUGGISBERG, *S.O.S. Rhino* (1966), a readable general account of the living rhinoceroses, their biology and conservation; H. KLINGEL, "Soziale Organisation und Verhalten freilebender Steppenzebras," and "Soziale Organisation und Verhaltensweisen von Hartman und Bergzebras," *Z. Tierpsychol.*, 24:580–624 and 25:76–88 (1967–68), descriptions of the behaviour and social organization of zebras; G.G. SIMPSON, *Horses: The Story of the Horse Family in the Modern World and through 60 Million Years of History* (1961), an interesting account of the natural history and evolution of the Equidae.

(R.C.Bi.)

Permafrost

Permafrost, or perennially frozen ground, is naturally occurring material that has a temperature colder than 0° C (32° F) continuously for two or more years. This layer of frozen ground is designated exclusively on the basis of temperature. Part or all of its moisture may be unfrozen, depending upon the chemical composition of the water or depression of the freezing point by capillary forces. Permafrost with saline soil moisture, for example, may be colder than 0° C for several years but would contain no ice and would not be firmly cemented. Most permafrost, however, is consolidated by ice; permafrost with no water, and thus no ice, is termed dry permafrost. The upper surface of permafrost is called the permafrost table. In permafrost areas the surface layer of ground that freezes in the winter (seasonally frozen ground) and thaws in summer is called the active layer. The thickness of the active layer depends mainly on the moisture content, varying from less than a foot in thickness in wet, organic sediments to several feet in well-drained gravels.

This article treats the origin and physical characteristics of permafrost. For further information on general soil properties, see SOILS; for discussions of climatic regimes and permafrost in the past, see CLIMATE; CLIMATIC CHANGE; and PLEISTOCENE EPOCH. See also ROCKS, PHYSICAL PROPERTIES OF, for information on the strength of natural materials, and GLACIATION, LANDFORMS PRODUCED BY, for relevant descriptions of physical features.

From T. Pewe, *The Periglacial Environment; Past and Present*, (1969); McGill-Queens University Press

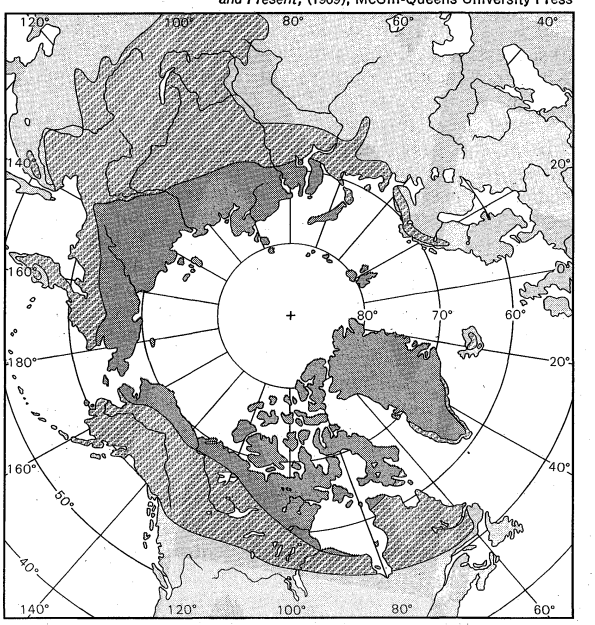


Figure 1: Distribution of permafrost in the Northern Hemisphere.

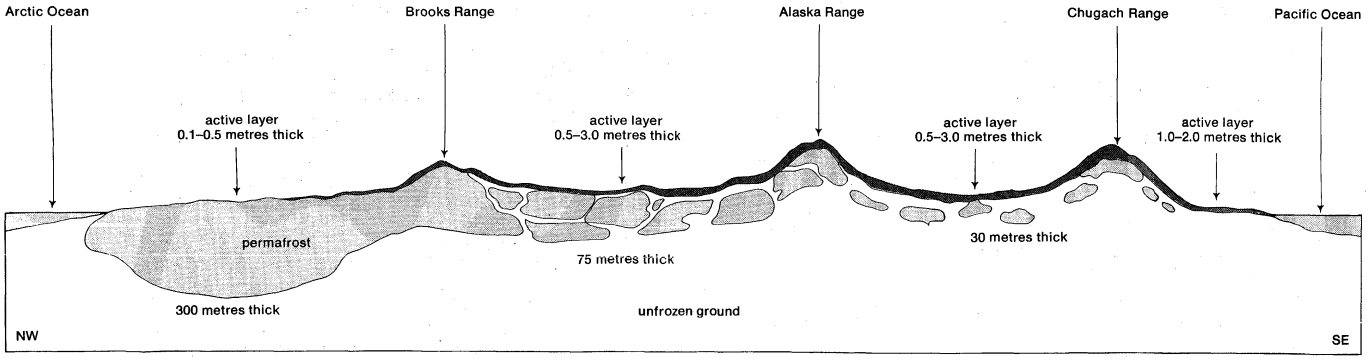


Figure 2: Approximate distribution of permafrost in Alaska.

Permafrost is widespread in the northern part of the Northern Hemisphere (Figure 1), and it is estimated to underlie 20 percent of the world's land surface. It occurs in 85 percent of Alaska, 50 percent of the U.S.S.R. and Canada, and probably all of Antarctica. In the Northern Hemisphere, permafrost is more widespread and extends to greater depths in the north than in the south. It is 5,000 feet (1,525 metres) thick in northern Siberia, 2,000 ft (610 m) thick in northern Alaska, and thins progressively toward the south (Figure 2). Permafrost can be differentiated into two broad zones; the continuous and the discontinuous, referring to the lateral continuity of permafrost. In the continuous zone of the far north, permafrost is nearly everywhere present except under the lakes and rivers that do not freeze to the bottom. The discontinuous zone includes numerous permafrost-free areas that increase progressively in size and number from north to south. Near the southern boundary, only rare patches of permafrost exist.

Permafrost forms and exists in a climate where the mean annual air temperature is 0° C or colder. Such a climate is generally characterized by long, cold winters with little snow and short, relatively dry, cool summers. Permafrost, therefore, is widespread in the Arctic, sub-Arctic, and Antarctica. There is a fairly good correlation between the southern boundary of Northern Hemisphere permafrost and the position of the 0° C mean annual air isotherm.

Although the existence of permafrost was known to the inhabitants of Siberia for centuries, scientists of the Western world did not take seriously isolated reports of thick frozen ground existing under northern forest and grasslands until 1836. Then, Alexander Theodor von Middendorff measured temperatures to depths of approximately 350 ft (107 m) of permafrost in the Shargin shaft, an unsuccessful well dug for the governor of the Russian-Alaskan Trading Company, at Yakutsk, and estimated that the permafrost was 700 ft (214 m) thick.

Carcass preservation

A most interesting aspect of permafrost is the preservation of carcasses and partial carcasses of modern and extinct Ice Age mammals within it. Frozen carcasses have been known in the north for centuries. The first direct report of a carcass of a mammoth in frozen ground of Siberia was by E. Yssbrants Ides in 1692. At least one or two almost complete, frozen mammoths have been reported from Siberia. Although no such famous frozen carcasses as those found in Siberia have been reported from North America, partial carcasses of mammoth, bison, musk ox, moose, horse, lynx, caribou, and ground squirrel have been found in Alaska.

Perhaps the most celebrated find of an Alaskan frozen carcass was the partial forequarters of a baby mammoth collected in 1948. The well-preserved hide of the head, neck, trunk, and one front leg was about one-quarter of an inch thick and almost hairless. A most interesting and important discovery of a partial carcass of an extinct superbison was made in 1951 in central Alaska (Figure 3); the age given for it by the radiocarbon laboratory of the Geological Survey of Sweden is 31,400 years.

Stories have been reported for many years of mammoth meat served at banquets, mammoths dying with buttercups in their mouths, and mammoths perishing as cataclysmic climatic changes took place that turned tropical



Figure 3: Partial carcass of an extinct bison *(Bison crassicornis)* discovered in frozen silt in Dome Creek, near Fairbanks, Alaska. An age of 31,400 (+2,040 or −1,815) years was obtained by testing a piece of hide for radioactive-carbon activity.
Troy L. Pewe

climates into frigid Arctic conditions. It is not necessary, however, to entertain fantasies about the distribution, preservation, and extinction of Ice Age mammals. Work in Alaska demonstrates that all the carcasses are between 10,000 and 70,000 years old and could not have been older because they would not have survived the preceding interglacial warm period when the ground and its enclosed carcasses thawed. In order to preserve the carcasses in nature's deep freeze, it is evident that most of the permafrost present today, in central Alaska at least, has existed since latest glacial times. Geological relationships and radiocarbon dates indicate that the carcasses are remains from the last glacial cold period and are not "a million years old" as casually announced on occasion. All evidence suggests that most carcasses represent natural deaths in a cold, harsh climate.

Permafrost affects many of man's activities in the Arctic and sub-Arctic, causing problems not experienced elsewhere and requiring that conventional engineering techniques be modified at additional costs. Agriculture, mining, water supply, sewage disposal, and construction are seriously affected by subsidence caused by thawing of perennially frozen ground as well as by associated problems brought on by soil flowage and frost action. Without a thorough understanding of thermal and mechanical problems unique to permafrost, impassable roads, unusable air strips, and abandoned buildings may result.

ORIGIN AND STABILITY OF PERMAFROST

In areas where the mean annual air temperature becomes colder than 0° C, some of the ground frozen in the winter will not be completely thawed in the summer; therefore, a layer of permafrost will form and continue to grow downward gradually each year from the seasonally frozen

ground. The permafrost layer will become thicker each winter, its thickness controlled by the thermal balance between the heat flow from the earth's interior and that flowing outward into the atmosphere. This balance depends upon the mean annual air temperature and the geothermal gradient. The average geothermal gradient is an increase of 1° C for every 100–200 ft (30.5 to 61 m) of depth. Eventually the thickening permafrost layer reaches an equilibrium depth at which the amount of geothermal heat reaching the permafrost is on the average equal to that lost to the atmosphere. Thousands of years are required to attain a state of equilibrium where permafrost is hundreds of feet thick.

Geothermal gradient

An example of the change of temperature of frozen ground with depth and the upper and lower limit of permafrost is illustrated in Figure 4. The annual fluctuation

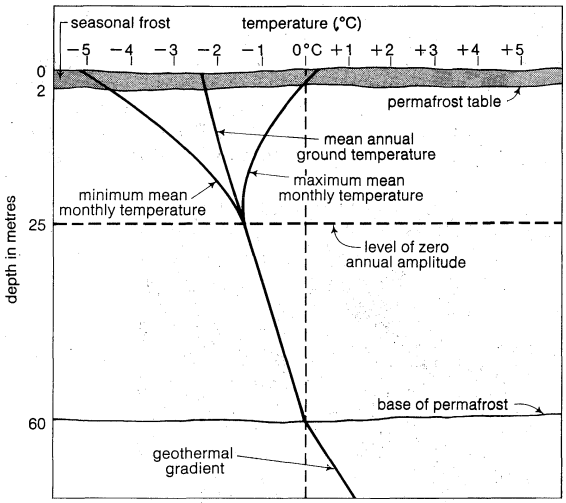


Figure 4: Hypothetical example of a temperature profile and thickness of permafrost in central Alaska.

of air temperature from winter to summer is reflected in a subdued manner in the upper few feet of the ground. This fluctuation diminishes rapidly with depth, being only a few degrees at 25 ft (7.6 m), and is barely detectable at 50 ft (15.2 m). The level of zero amplitude, at which fluctuations are hardly detectable, is 30–50 ft (9.2–15.2 m). If the permafrost is in thermal equilibrium, the temperature at the level of zero amplitude is generally regarded as the minimum temperature of the permafrost. Below this depth the temperature increases steadily under the influence of heat from the earth's interior. The temperature of permafrost at the level of zero amplitude is close to the average temperature of the ground surface, being, however, 2°–6° C warmer than the mean annual air temperature determined from the U.S. Weather Bureau-type surface installations (at a height of 5 ft [1.5 m] above ground level). The temperature of permafrost at the depth of minimum annual seasonal change varies from near 0° C at the southern limit of permafrost to −10° C in northern Alaska and −13° C in northeastern Siberia.

Climatic controls

As the climate becomes colder or warmer, but maintaining a mean annual temperature colder than 0° C, the temperature of the permafrost correspondingly rises or declines resulting in changes in the position of the base of permafrost. The position of the top of permafrost will be lowered by thawing when the climate warms to a mean annual air temperature warmer than 0° C. The rate at which the base or top of permafrost is changed depends not only on the amount of climatic fluctuation but also on the amount of ice in the ground and the composition of the ground, conditions that in part control the geothermal gradient. If the geothermal gradient is known and if the surface temperature is stable for a long period of time, it is, therefore, possible to predict from a knowledge of the mean annual air temperature the depth of permafrost in a particular area remote from water bodies.

Permafrost is the result of present climate. Many temperature profiles show, however, that permafrost is not in equilibrium with present climate at the sites of measurement. Some areas show, for example, that climatic warming since the last third of the 19th century has caused a warming of the permafrost to a depth of a few hundred feet. Here some of the existing perennially frozen ground would disappear with rising base of permafrost if present surface conditions were to persist for several thousands of years. Actually, in such areas much of the permafrost is a product of a colder, former climate.

It is thought that permafrost first occurred in conjunction with the onset of glacial conditions near the beginning of the Quaternary Period about 2,500,000 to 3,000,000 years ago. The oldest permafrost known in North America, however, is about mid-Quaternary in age, or less than 2,000,000 years old. In the sub-Arctic at least, most permafrost probably disappeared during the interglacial times and reappeared in glacial times. Most existing permafrost probably formed in the cold (glacial) period of the last 70,000 years.

LOCAL THICKNESS AND ICE CONTENT

The thickness and areal distribution of permafrost is directly affected by snow and vegetation cover, topography, bodies of water, the interior heat of the earth, and the temperature of the atmosphere as mentioned earlier.

The most conspicuous change in thickness of permafrost is related to climate. At Barrow, Alaska, the mean annual air temperature is −12° C and the thickness is 1,300 ft (396.5 m). At Fairbanks, Alaska, in the discontinuous zone of permafrost in central Alaska, the mean annual air temperature is −3° C and the thickness is about 300 ft (91.5 m). Near the southern border of permafrost, the mean annual air temperature is about 0° or −1° C and the perennially frozen ground is only a few feet thick.

If the mean annual air temperature is the same in two areas, the permafrost will be thicker where the conductivity of the ground is higher and the geothermal gradient is less. A.H. Lachenbruch of the U.S. Geological Survey reports an interesting example from northern Alaska. The mean annual air temperatures at Cape Simpson and Prudhoe Bay are similar, but permafrost thickness is 1,000 ft (305 m) at Cape Simpson and about 2,000 ft (710 m) at Prudhoe Bay because rocks at Prudhoe Bay are more siliceous and have a higher conductivity and a lower geothermal gradient than rocks at Cape Simpson.

Effect of water bodies

Bodies of water, lakes, rivers, and the sea have a profound effect on the distribution of permafrost. A deep lake that does not freeze to the bottom during the winter will be underlain by a zone of thawed material. If the minimum horizontal dimension of the deep lake is about twice as much as the thickness of permafrost nearby, there probably exists an unfrozen vertical zone extending all the way to the bottom of permafrost. Such thawed areas extending all the way through permafrost are widespread under rivers and sites of recent rivers in the discontinuous zone of permafrost and under major, deep rivers in the far north. Under the wide floodplains of rivers in the sub-Arctic, the permafrost is sporadically distributed both laterally and vertically. Small, shallow lakes that freeze to the bottom each winter are underlain by a zone of thawed material, but the thawed zone does not completely penetrate permafrost except near the southern border of permafrost.

Inasmuch as south-facing hillslopes receive more incoming solar energy per unit area than other slopes, they are warmer; permafrost is generally absent on these in the discontinuous zone (Figure 5) and is thinner in the continuous zone.

The main role of vegetation in permafrost areas is to shield perennially frozen ground from solar energy. Vegetation is an excellent insulating medium and removal or disturbance of it, either by natural processes or by man, causes thawing of the underlying permafrost. In the continuous zone the permafrost table may merely be lowered by the disturbance of vegetation, but in a discontinuous zone permafrost may be completely destroyed in certain areas.

Snow cover also influences heat flow between the ground and the atmosphere and therefore affects the distribution

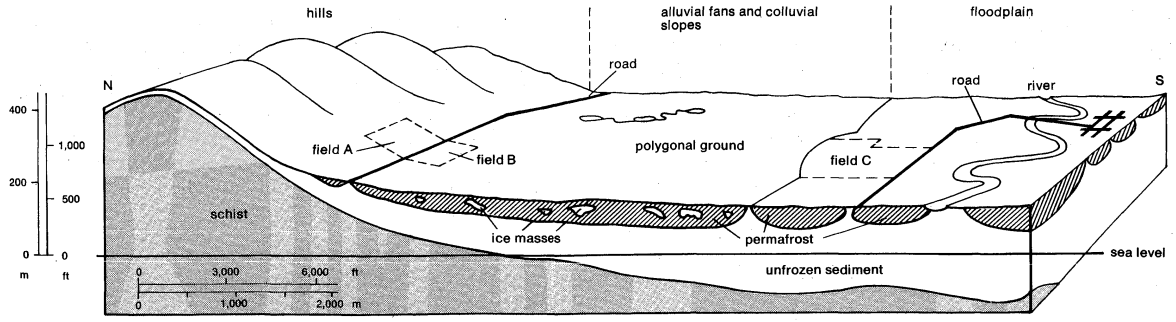


Figure 5: Character and distribution of permafrost in Fairbanks, Alaska, area.
From T. Pewe, *Permafrost and its Effect on Life in the North* (1966); Oregon State University Press

of permafrost. If the net effect of timely snowfalls is to prevent heat from leaving the ground in the cold winter, permafrost becomes warmer. Actually, local differences in vegetation and snowfall in areas of thin and warm permafrost are critical for the formation and existence of the perennially frozen ground. Permafrost is not present in areas of the world where great snow thicknesses persist throughout most of the winter.

Types of ground ice. The ice content of permafrost is probably the most important feature of permafrost affecting human life in the north. Ice in the perennially frozen ground exists in various sizes and shapes and has definite distribution characteristics. The forms of ground ice can be grouped into five main types: (1) pore ice; (2) segregated, or Taber, ice; (3) foliated, or wedge, ice; (4) pingo ice; and (5) buried ice.

1. Pore ice, which fills or partially fills pore spaces in the ground, is formed by pore water freezing *in situ* with no addition of water. The ground contains no more water in the solid state than the ground could hold if the water were in liquid state.
2. Segregated, or Taber, ice includes ice films, seams, lenses, pods, or layers generally 1/16 inch to 5 inches (0.16 to 12.7 centimetres) thick that grow in the ground by drawing in water as the ground freezes. Small ice segregations are the least spectacular but one of the most extensive types of ground ice, and engineers and geologists interested in ice growth and its effect on engineering structures have studied them considerably. Such observers generally accept the principle of bringing water to a growing ice crystal, but they do not completely agree as to the mechanics of the processes. Pore ice and Taber ice occur both in seasonally frozen ground and in permafrost.
3. Foliated ground ice, or wedge ice, is the term for large masses of ice growing in thermal contraction cracks in permafrost.
4. Pingo ice is clear, or relatively clear, and occurs in permafrost more or less horizontally or in lens-shaped masses. Such ice originates from groundwater under hydrostatic pressure.
5. Buried ice in permafrost includes buried sea, lake, and river ice and recrystallized snow, as well as buried blocks of glacier ice in permafrost climate.

Ice content

World estimates of the amount of ice in permafrost vary from 48,000 to 120,000 cubic miles (200,000 to 500,000 cubic kilometres), or less than 1 percent of the total volume of the earth. Based on an examination of ice in the ground in many bore holes near Barrow, Alaska, and extrapolation of this bore hole information to the rest of the Coastal Plain by use of aerial photographs and geologic maps, it has been estimated that 10 percent by volume of the upper 10 ft (3.05 m) of permafrost on the Coastal Plain of Alaska is composed of foliated ground ice (ice wedges). Taber ice is the most extensive type of ground ice, and in places it represents 75 percent of the ground by volume. It is calculated that the pore and Taber ice content in the depth between 1.5 and 10 ft (0.46 and 3.05 m; surface to 1.5 feet is seasonally thawed) is 61 percent by volume, and between 10 and 30 ft (3.05 and 9.15 m) it is 41 percent. The total amount of pingo ice is less than 0.1 percent of the permafrost. The total ice content in the permafrost of the Arctic Coastal Plain of Alaska is estimated to be 360 cu mi (1,500 cu km), and below 30 ft (9.15 m) most of that is present as pore ice.

Ice wedges. The most conspicuous and controversial type of ground ice in permafrost is that formed in large ice wedges or masses with parallel or subparallel foliation structures. Most foliated ice masses occur as wedge-shaped, vertical, or inclined sheets or dikes 1 in. to 10 ft (2.54 cm to 3.05 m) wide and 1–30 ft (0.305 to 9.15 m) high when viewed in transverse cross section (Figure 6).

Troy L. Pewe

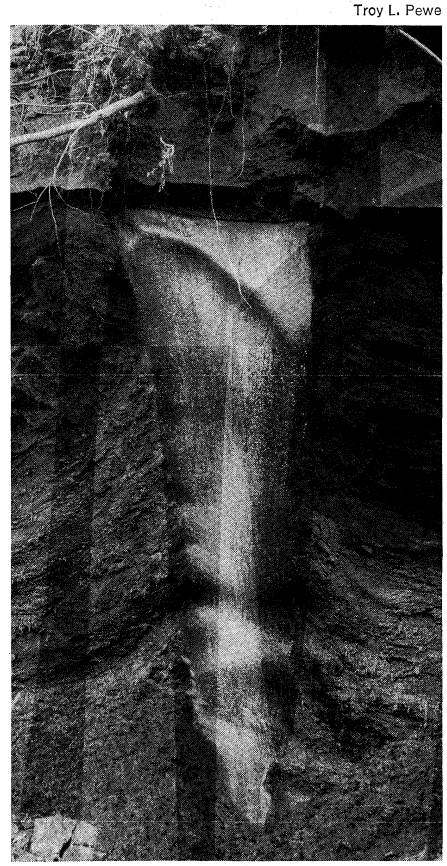


Figure 6: Foliated ground-ice mass (ice wedge) in organic rich silt exposed by gold-mining operations on Wilber Creek, near Livengood, Alaska.

Some masses seen on the face of frozen cliffs may appear as horizontal bodies a few inches to 10 ft (3 m) in thickness and 1–45 ft (0.305 to 13.7 m) long, but the true shape of these ice wedges can be seen only in three dimensions. Ice wedges are parts of polygonal networks of ice enclosing cells of frozen ground 10–100 ft (3–30 m) or more in diameter.

Thermal contraction theory

The origin of ground ice was first studied in Siberia, and discussions in print of the origin of large ground-ice masses in perennially frozen ground of North America have gone on since Otto von Kotzebue recorded ground ice in 1816 at a spot now called Elephant's Point in Eschscholtz Bay of Seward Peninsula. The theory for the origin of ice wedges now generally accepted is the thermal

contraction theory that, during the cold winter, polygonal thermal contraction cracks, about an inch wide and several feet deep, form in the frozen ground; then when, in early spring, water from the melting snow runs down these tension cracks and freezes, a vertical vein of ice is produced that penetrates into permafrost; when the permafrost warms and re-expands during the following summer, horizontal compression produces upturning of the frozen sediment by plastic deformation; then during the next winter, renewed thermal tension reopens the vertical ice-cemented crack, which may be a zone of weakness; another increment of ice is added in the spring when meltwater again enters and freezes. Over the years the vertical wedge-shaped mass of ice is produced (Figure 6).

Ice wedges may be classified as active, inactive, and ice-wedge casts. Active ice wedges are those that are actively growing. The wedge may not crack every year, but during many or most years cracking does occur, and an increment of ice is added. Ice wedges require a much more rigorous climate to grow than does permafrost. The permafrost table must be chilled to −15° C to −20° C for contraction cracks to form. On the average, it is assumed that ice wedges generally grow in a climate where the mean annual air temperature is −7° or −8° C or colder.

The area of active ice wedges appears to roughly coincide with the continuous permafrost zone. From north to south across the permafrost area in North America, a decreasing number of wedges crack frequently. The line dividing zones of active and inactive ice wedges is arbitrarily placed at the position where it is thought most wedges do not frequently crack.

Inactive ice wedges are those that are no longer growing. The wedge does not crack in winter and, therefore, no new ice is added. A gradation between active ice wedges and inactive ice wedges occurs in those wedges that crack rarely. Inactive ice wedges have no ice seam or crack extending from the wedge upward to the surface in the spring. The wedge top may be flat (Figure 6), especially if thawing has lowered the upper surface of the wedge at some time in the past.

Indication of former temperature

Ice wedges in the world are of several ages, but none appear older than the outset of the last major cold period, about 70,000 years ago. Wedges dated by radiocarbon analyses range from 3,000 to 32,000 years in age.

In many places in the now temperate latitudes of the world, in areas of past permafrost, ice wedges have melted, and resulting voids have been filled with sediments collapsing from above and the sides. These ice-wedge casts, or fossil ice wedges, are important as paleoclimatic indicators and indicate a climate of the past with at least a mean annual air temperature of −7° or −8° C or colder.

SURFACE MANIFESTATIONS OF PERMAFROST AND SEASONALLY FROZEN GROUND

Many distinctive surface manifestations of permafrost exist in the Arctic and sub-Arctic, including such geomorphic features as polygonal ground, thermokarst phenomena, and pingos. In addition to the above, there are many features caused in large part by frost action that are common in but not restricted to permafrost areas, such as solifluction (soil flowage) and frost-sorted patterned ground.

Areas underlain by permafrost. One of the most widespread geomorphic features associated with permafrost is the microrelief pattern on the surface of the ground generally called polygonal ground, or tundra polygons (Figure 7). This pattern, which covers thousands of square miles of the Arctic and less in the sub-Arctic, is caused by an intersecting network of shallow troughs (delineating polygons 10–100 ft (3.05 to 30.5 m) in diameter. The troughs are underlain by more or less vertical ice wedges (Figure 6) 2–10 ft (0.71 to 3.05 m) across on the top that are joined together in a honeycomb network. These large-scale polygons should not be confused with the small-scale polygons or patterned ground produced by frost sorting.

Ice-wedge polygons

The ice-wedge polygons may be low centred or high centred. Upturning of strata adjacent to the ice wedge may make a ridge of ground on the surface on each side of the wedge, thus enclosing the polygons. Such polygons are lower in the centre and are called low-centre polygons or raised-edge polygons and may contain a pond in the centre. Low-centre, or raised-edge, polygons indicate that ice wedges are actually growing and that the sediments are being actively upturned. If erosion, deposition, or thawing is more prevalent than the up-pushing of the sediments along the side of the wedge or if the material being pushed up cannot maintain itself in a low ridge, the low ridges will be absent, and there may be either no polygons at the surface or the polygons may be higher in the centre than the troughs over the ice wedges that enclose them. Both high-centre and low-centre tundra polygons are widespread in the polar areas and are good indicators of the presence of foliated ice masses; care must be taken, however, to demonstrate that the pattern is not a relic and an indication of "fossil ice" wedges.

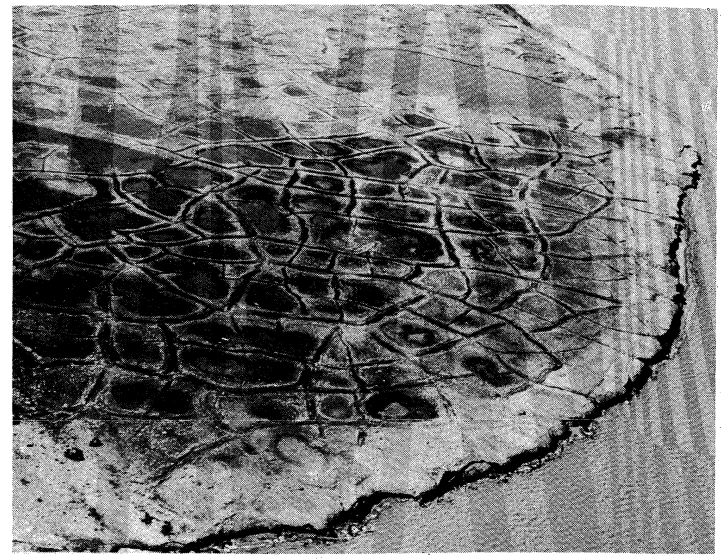


Figure 7: Raised-edge ice wedge polygons on the seacoast near Barrow, Alaska, in summer. The polygons are from seven to 15 metres in diameter.
By courtesy of the U.S. Army (CRREL); photograph, Robert Lewellen

In many parts of the temperate latitudes of Asia, Europe, and North America, incompletely developed or poorly developed polygonal ground occurs on the same scale as in the Arctic. These large-scale polygons in the nonpermafrost areas are excellent evidence of the former extent of permafrost and ice wedges in the last glacial period.

In many areas of the continuous permafrost zone surface, drainage follows the troughs of the polygons (tops of the ice wedges); and at ice wedge junctions, or elsewhere, melting may occur to form small pools. The joining of these small pools by a stream causes the pools to resemble beads on a string, a type of stream form called beaded drainage. Such drainage indicates the presence of perennially frozen, fine-grained sediments cut by ice wedges.

Thermokarst topography and pingos

The thawing of permafrost creates thermokarst topography, an uneven surface that contains mounds, sinkholes, tunnels, caverns, and steep-walled ravines caused by melting of ground ice. The hummocky ground surface resembles karst topography found in limestone areas. Thawing may result from artificial or natural removal of the vegetation or from warming of the climate (Figure 8).

Thawed depressions filled with water (thaw lakes, thermokarst lakes, cave-in lakes) are widespread in permafrost areas, especially in those underlain with perennially frozen silt. They may occur on hillsides or even on hill tops and are good indicators of ice-rich permafrost. Locally, deep thermokarst pits 20 ft (6.1 m) deep and 30 ft (9.15 m) across may form as ground ice melts. These openings may exist as undetected caverns for many years before the roof collapses. Such collapses in agricultural or construction areas are real dangers. Thermokarst mounds are polygonal or circular hummocks 10–50 ft (3.05 to 15.25 m) in diameter and 1–8 ft (0.305 to 2.84 m) high

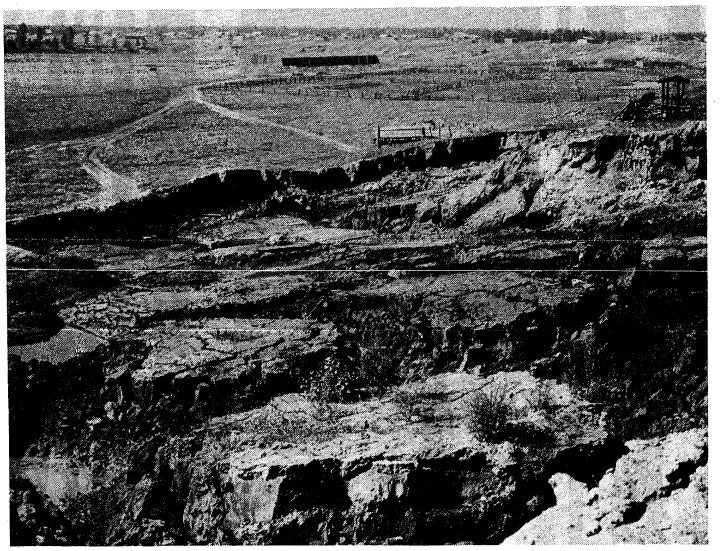


Figure 8: Modern thermokarst formation at Maya, near Yakutsk, Siberia.
Troy L. Pewe

that are formed as a polygonal network of ice melts and leaves the inner-ice areas as mounds.

The most spectacular landforms associated with permafrost are pingos, small ice-cored circular or elliptical hills of frozen sediments or even bedrock, 10 to more than 200 ft (3.05 to 61 m) high and 50 to 1,500 ft (15.25 to 457.5 m) in diameter. Pingos are widespread in the continuous permafrost zone and are quite conspicuous because they rise above the tundra. They are much less conspicuous in the forested area of the discontinuous permafrost zone. They are generally cracked on top with summit craters formed by melting ice. There are two types of pingos, based on origin. The closed-system type forms in level areas when unfrozen groundwater in a thawed zone becomes confined on all sides by permafrost, freezes, and heaves the frozen overburden to form a mound. This type is larger and occurs mainly in tundra areas of continuous permafrost. The open-system type is generally smaller and forms on slopes when water beneath or within the permafrost penetrates the permafrost under hydrostatic pressure. A hydrolaccolith (water mound) forms and freezes, heaving the overlying frozen and unfrozen ground to produce a mound.

Present pingos are apparently the result of postglacial climate and are less than 4,000–7,000 years old. Pingos were present in now temperate latitudes during the latest glacial epoch and are now represented by low circular ridges enclosing bogs or lowlands.

Near the southern border of permafrost occur palsas, low hills and knobs of perennially frozen peat about 5–20 ft (1½ to 6 m) high, evidently forming with accumulation of peat and segregation of ice.

Features related to seasonal frost. Many microgeomorphic features common to the periglacial environment may or may not be associated with permafrost. Intense seasonal frost action, repeated freezing and thawing throughout the year, produces small-scale patterned ground. Repetitive freezing and thawing tends to stir and sort granular sediments, thus forming circles, stone nets, and polygons a few inches to 20 ft (6 m) in diameter. The coarse cobbles and boulders form the outside of the ring and the finer sediments occur in the centre. The features require a rigorous climate with some fine-grained sediments and soil moisture, but they do not necessarily need underlying permafrost. Permafrost, however, forms an impermeable substratum that keeps the soil moisture available for frost action. On gentle slopes the stone nets may be distorted into garlands by downslope movement or, if the slope is steep, into stone stripes 1–2 ft (0.305 to 0.61 m) wide and 100 ft (30.5 m) long.

Stone stripes, stone nets, patterned ground, and solifluction

In areas underlain by an impermeable layer (seasonally frozen ground or perennially frozen ground), the active layer is often saturated with moisture and is quite mobile. The progressive downslope movement of saturated detrital material under the action of gravity and working in conjunction with frost action is called solifluction. This material moves in a semifluid condition and is manifested by lobelike and sheetlike flows of soil on slopes. The lobes are up to 100 ft (30.5 m) wide and have a steep front 1–5 ft (0.305 to 1.52 m) high. An outstanding feature of solifluction is the mass transport of material over low-angle slopes. Solifluction deposits are widespread in polar areas and consist of a blanket 1–6 ft (0.305 to 1.83 m) thick of unstratified or poorly stratified, unsorted, heterogeneous, till-like detrital material of local origin. In many areas the terrain is characterized by relatively smooth, round hills and slopes with well-defined to poorly defined solifluction lobes or terraces. If the debris is blocky and angular and fine material is absent, the lobes are poorly developed or absent. Areas in which solifluction lobes are well formed lie almost entirely above or beyond the forest limit.

In many areas the frost-rived debris contains few fine materials, little water, and consists of angular fragments of well-jointed, resistant rock. Under such circumstances, solifluction lobes do not often occur, but instead striking sheets or streams of angular rubble form. These are called rock streams or rubble sheets.

PROBLEMS POSED BY PERMAFROST

Development of the north demands an understanding of and the ability to cope with problems of the environment dictated by permafrost. Although the frozen ground hinders agricultural and mining activities, the most dramatic, widespread, and economically important examples of the influence of permafrost on life in the north involve construction and maintenance of roads, railroads, airfields, bridges, buildings, dams, sewers, and communication lines. Engineering problems are of four fundamental types: (1) those involving thawing of ice-rich permafrost and subsequent subsidence of the surface under unheated structures such as roads and airfields; (2) those involving subsidence under heated structures; (3) those resulting from frost action, generally intensified by poor drainage caused by permafrost; and (4) those involved only with the temperature of permafrost that causes buried sewer, water, and oil lines to freeze.

A thorough study of the frozen ground should be part of the planning of any engineering project in the north. It is generally best to attempt to disturb the permafrost as little as possible in order to maintain a stable foundation for engineering structures, unless the permafrost is thin; then, it may be possible to destroy the permafrost. The method of construction preserving the permafrost has been called the passive method; alternately, the destroying of permafrost is the active method.

Because thawing of permafrost and frost action are involved in almost all engineering problems in polar areas, it is advisable to consider these phenomena generally. The delicate thermal equilibrium of permafrost is disrupted when the vegetation, snow cover, or the active layer is compacted. The permafrost table is lowered, the active layer is thickened, and considerable ice is melted. This process lowers the surface and provides (in summer) a wetter active layer with less bearing strength. Such disturbance permits a greater penetration of summer warming. It is common procedure to place a fill, or pad, of gravel under engineering works. Such a fill generally is a good conductor of heat and, if thin, may cause additional thawing of permafrost. The fill must be made thick enough to contain the entire amplitude of seasonal temperature variation—in other words, thick enough to restrict the annual seasonal freezing and thawing to the fill and the compacted active layer. Under these conditions no permafrost will thaw. Such a procedure is quite feasible in the Arctic, but in the warmer sub-Arctic it is impractical because of the enormous amounts of fill needed. Under a heated building, profound thawing may occur more rapidly than under roads and airfields.

Frost action, the freezing and thawing of moisture in the ground, has long been known to seriously disrupt and destroy structures in both polar and temperate latitudes.

Frost action and man-made structures

In the winter the freezing of ground moisture produces upward displacement of the ground (frost heaving), and in the summer excessive moisture in the ground brought in during the freezing operation causes loss of bearing strength. The intensity of frost action is regulated by temperature, the texture of the ground, and the amount of soil moisture. Frost action is best developed in silt-sized and silty clay-sized sediments in areas of rigorous climate and poor drainage. Polar latitudes are ideal for maximum frost action because most lowland areas are covered by fine-grained sediments, and the underlying permafrost causes poor drainage.

Piles are used to support many, if not most, structures built on ice-rich permafrost. In regions of cold winters, many pile foundations are in ground subject to seasonal freezing and, therefore, possibly subject to the damaging effect of frost heaving, which tends to displace the pile upward and thus to disturb the foundation of the structure. The displacement of piling is not limited to the far north, though maximum disturbance probably is encountered most widely in the sub-Arctic. Expensive maintenance and sometimes complete destruction of bridges, school buildings, military installations, and other structures have resulted from failure to understand the principles of frost heaving of piling.

The need for careful geological location of highways, railroads, airfields, and pipelines is very important anywhere in the world; but it is absolutely essential in permafrost areas in order to avoid constant, costly maintenance—if not complete abandonment. Such location studies were not frequently considered in construction work before the 1950s in North America and perhaps before the 1930s in Asia. Even late in the 20th century, many construction projects proceeded with inadequate location studies.

Highways in polar areas are relatively few and mainly unpaved. They are subject to subsidence by thawing of permafrost in summer, frost heaving in winter, and loss of bearing strength on fine-grained sediments in summer. Constant grading of gravel roads permits maintenance of a relatively smooth highway. Where the road is paved over ice-rich permafrost, the roadway becomes rough and is much more costly to maintain than are unpaved roads. Many of the paved roads in polar areas have required resurfacing two or three times in a ten-year period.

Railroads particularly have serious construction problems and require costly upkeep in permafrost areas because of the necessity of maintaining a relatively low gradient and the subsequent location of the roadbed in ice-rich lowlands that are underlain with perennially frozen ground. The Trans-Siberian Railroad, the Alaska Railroad, and some Canadian railroads in the north are locally underlain by permafrost with considerable ground ice. As the large masses of ice melt each summer, constant maintenance is required to level these tracks. In winter, extensive maintenance is also required to combat frost heaving when local displacements of 1–14 in. (2.54–35.6 cm) occur in roadbeds and bridges.

Heated buildings in polar areas can have a conventional basement if they are built in areas of no permafrost or of bedrock and gravel with little or no ice. Construction of such buildings on ice-rich permafrost, however, requires a technique that enables the cold winter air to circulate under the building to counteract the heat generated in the building. The most common way to build structures above the ground with large spaces under the foundation for airflow is to construct on piles, a technique perfected both in the U.S.S.R. and in North America. The piles must be erected so that they do not move. If the piles are jacked up in the winter by frost action, the building will be disturbed and can be destroyed. It is therefore important that the pile be adequately anchored in permafrost or it be isolated from the heaving of the ground by a movable collar.

Groundwater and agriculture

Both surface and subsurface water supplies in the north are beset by problems uncommon in temperate latitudes. Groundwater sources are above the permafrost, in unfrozen zones within the permafrost, and below the permafrost. Sources above permafrost may be seasonal and subject to contamination from cesspools. Sources from unfrozen zones within the permafrost are common in the discontinuous zones where water also can be obtained below permafrost. In the far north, however, little water is available in the permafrost, deep drilling (1,000–2,000 ft [305 to 610 m]) is expensive, and the sediments generally lack sufficient groundwater.

Distribution of water and sewage presents unique problems in permafrost areas. In addition to requiring a heated distribution system to prevent freezing in permafrost or seasonal frost areas, utility distribution lines are subject to problems because of frost action and thawing of permafrost. The most successful construction method generally used in such situations is to place all services in heated distribution boxes called utilidors, which enclose water, steam, gas, sewage, and other utilities. The utilidor may be either under or above ground.

Permafrost affects agricultural developments in many parts of the discontinuous permafrost zone. Its destructive effect on cultivated fields in both the U.S.S.R. and North America results from the thawing of large masses of ice in the permafrost. If care is not exercised in selecting areas to be cleared for cultivation, thawing of the permafrost may necessitate abandonment of fields or their reduction to pasturage. As illustrated in Figure 5, Field A, on the southward-facing hillside, is not affected by permafrost; Field B, on the alluvial fan, however, is underlain with permafrost containing large masses of ground ice. When Field B is cleared, thermokarst topography will form. Field C, on the floodplain, is underlain with permafrost without large ice masses, and so no thermokarst pits and mounds will form when this ground is cleared.

In summary, it can be said that in construction and maintenance of railroads, buildings, water and sewer lines, dams, roads, bridges, and airfields, normal techniques must often be modified at additional costs because of permafrost. But, despite special problems unique to polar areas, development of the permafrost regions will continue at an ever-increasing rate. Man has already learned to cope with the problems, and improvements of scientific and engineering approaches, with careful geological site selection and study of the permafrost problem, will allow successful expansion into polar areas.

BIBLIOGRAPHY. R.J.E. BROWN, *Permafrost in Canada* (1970), the only comprehensive book on permafrost in Canada; O.J. FERRIANS, JR., *et al.*, "Permafrost and Related Engineering Problems in Alaska," *Prof. Pap. U.S. Geol. Surv. 678* (1969), brief, up-to-date, well-illustrated treatment of permafrost in Alaska; D.M. HOPKINS *et al.*, "Permafrost and Ground Water in Alaska," *Prof. Pap. U.S. Geol. Surv. 264-F*, pp. 113–146 (1955), a good review of permafrost effect on ground water and methods of determining permafrost from aerial photographs; A.H. LACHENBRUCH, "Mechanics of Thermal Contraction Cracks and Ice Wedge Polygons in Permafrost," *Spec. Pap. Geol. Soc. Am. 70* (1962), classic paper on quantitative interpretation of the formation of ice-wedge polygons, and "Thermal Considerations in Permafrost," in *Geological Seminar on the North Slope of Alaska* (AAPGP, Pacific sect., 1970), with discussions by ROBERT STONELY, T.C. MATHEWS, and S.W. MULLER; NATIONAL ACADEMY OF SCIENCES-NATIONAL RESEARCH COUNCIL, *Proc. Intern. Permafrost Conf., 11–15 Nov., 1963*, Pub. No. 1287 (1966), an extensive collection of up-to-date papers on permafrost presented at the First International Permafrost Conference; T.L. PEWE (ed.), *The Periglacial Environment: Past and Present* (1969), an up-to-date approach to discussion of present and past permafrost throughout the world, and with R.E. CHURCH and M.J. ANDERSON, "Origin and Paleoclimatic Significance of Large Scale Polygons in the Donnelly Dome Area, Alaska," *Spec. Pap. Geol. Soc. Am. 109* (1969), a thorough up-to-date discussion of "fossil" ice-wedge polygons and relict permafrost in central Alaska with paleoclimatic interpretations; A.I. POPOV, *Podzemnyïled*, 3 vol. (1965), an extensive review with up-to-date papers dealing with ground ice and permafrost (in Russian, with preface and table of contents also in English); M.I. SUMGIN *et al.*, *Obshchee merzlotovedenie* (1940), classical summary of permafrost investigations in the USSR (in Russian); I.A. TYUTYUNOV, *An Introduction to the Theory of the Formation of Frozen Rocks*, trans. from the Russian by J.O.H. MUHLHAUS (1964), a comprehensive summary of the mechanics and chemistry of the freezing process and permafrost formation.

(T.L.Pe.)

Permian Period

Identification and definition

The Permian Period is a geological interval of time from about 280,000,000 to 225,000,000 years ago, during which the rocks of the Permian System were formed. This system was next to the last to be defined and is the only one not based on a type area in central or western Europe. In 1841 the Czar of Russia invited Sir Roderick Murchison, renowned British geologist, to carry out a comprehensive reconnaissance of the vast territories of European Russia, which at that time were little known geologically. Along the western slopes of the Ural Mountains, Murchison recognized a sequence of beds that proved to be younger than rocks of the Carboniferous Period (345,000,000 to 280,000,000 years ago) in England, which he knew well. Murchison assigned these rocks to a new system that he called Permian, after the district of Perm, U.S.S.R., in which they typically occurred. Murchison also recognized the equivalence of these rocks with sequences that had been described in Germany as Rotliegendes and Zechstein by earlier observers. In his original (1845) description, Murchison included the lower part of the Buntsandstein, which is now regarded as Triassic (225,000,000 to 190,000,000 years ago) in age.

Although the concept of the Permian Period was accepted by most geologists well before the end of the 19th century, proposals to discontinue its use were discussed as late as 1948, and the U.S. Geological Survey did not officially recognize it until 1941.

The most important Permian economic deposits are rock salt and potash salts, coal, oil, and gas. Economically important evaporites (salts formed by chemical precipitation under generally arid conditions) of Permian age are found in the United States, especially in west Texas; in Europe, in the Upper Permian Zechstein Basin of northern Germany, from where they extend in the subsurface into The Netherlands, England, Denmark, the North Sea, and Poland. The largest area of Permian salt deposition was in eastern European Russia, extending in a broad arch from the Dnepr River to the Caspian Sea and northward along the length of the Ural Mountains. In the Southern Hemisphere, Permian salt deposits are known only from South America (northeastern Brazil, Peru), where they are not of economic importance.

Permian coals (*q.v.*) occur in great quantity and excellent quality in many parts of the world. The most important deposits are in central and northern Siberia, northern and eastern China, India, Australia, South Africa, Rhodesia, and the Congo. No important coal deposits of Permian age are known in the Western Hemisphere.

Oil and gas are produced from Permian rocks in Oklahoma, Texas, eastern European Russia, and, increasingly, in The Netherlands, Germany, and the North Sea.

Phosphorites (*q.v.*) of Permian age crop out in Montana, Idaho, Wyoming, and Utah and also occur in the U.S.S.R. west of the Ural Mountains.

A copper-bearing deposit, known as *Kupferschiefer* (copper shale), occurs in Germany, where it has long been mined. It is a rare deposit formed under anaerobic (oxygen deficient) conditions in a basin with restricted circulation. Copper also is associated with sedimentary rocks of Permian age in Texas and Oklahoma.

Standard subdivisions

The rocks of the Permian Period generally are subdivided into a Lower Permian Series and an Upper Permian Series. The boundary between the two series is placed either at the base or the middle of the Guadalupian Stage of marine deposits in west Texas. Some authors have adopted a threefold division and use Middle Permian in their stratigraphic discussions, which is not recognized internationally. When so used, it generally includes the Artinskian and the lower half of the Guadalupian (Table 1).

Investigations during the 19th and especially the 20th century indicated that in its type locality the Permian was represented by marine and continental sediments. Marine stages are the Asselian, Sakmarian, and Artinskian. The Kungurian and Kazanian stages are of restricted marine type, containing a fauna that is not comparable with faunas in other parts of the world, and the uppermost stage, the Tartarian, comprises sediments of continental origin. Marine deposits of Late Permian age occur in other parts of the world, however. The question of worldwide correlation of the youngest Permian rocks is as yet unresolved.

Table 1: Marine Stages of the Permian System

	name	typical locality
Upper Permian	Dzhulfian	Armenian S.S.R.
	Chhidruan	Salt Range, Pakistan
	Guadalupian	West Texas
Lower Permian	Roadian	West Texas
	Artinskian, Sakmarian, Asselian	Perm Province, Ural Mountains

In parts of the world where relationships of strata of local significance with the standard reference sequences are not clear, independent stage classifications have been proposed (Table 2).

This article treats the rocks, life, and environment of Permian time. For coverage of adjacent intervals of geological time, see CARBONIFEROUS PERIOD, UPPER; and TRIASSIC PERIOD; an overview including the Permian Period is presented in PALEOZOIC ERA, UPPER. See also, FOSSIL RECORD; STRATIGRAPHIC BOUNDARIES; and MOUNTAIN BUILDING PROCESSES for further details on Permian life and environments; CLIMATIC CHANGE and CONTINENTAL DRIFT for the problem of Permian paleography; and PETROLEUM for oil and gas deposits.

PERMIAN ROCKS

Permian rocks are widely distributed in the geosynclinal belts (primarily, the modern mountain belts) of the world, although complete sections are rare because of much subsequent folding and metamorphism (changes due to temperature and pressure).

Major sequences

Thick Permian sequences occur in western North America and are best known in Texas, Nevada, and Utah. In the east–west geosynclinal belt of Europe and Asia, generally referred to as Tethys, important Permian occurrences are found from Sicily in the west to the island of Timor in the east, although in no single locality is the system represented in its entirety. Thick sequences of Permian rocks, mostly carbonates, are known in Anatolia, the Middle East, Iran, Transcaucasia, the Salt Range of Pakistan, Siberia, Central Asia, and Kashmir. In the circum-Pacific geosynclinal belt, Permian rocks are best known from China, Japan, western North America (including northern Mexico), Peru, and from New Zealand.

From the geosynclines, shallow seas spread over adjacent continental areas, especially in Early and early Late Permian times. Deposits in these seas, usually sandstone and shale and partly carbonate rocks, typically occur in the midcontinent of North America and in the Russian platform. Thick accumulations of Permian sediments are found in the Sverdrup Basin of the Canadian Arctic Archipelago and in East Greenland.

A special case is that of the Ural Geosyncline, which through much of Paleozoic time separated Europe from Asia. Permian sediments accumulated in great thickness but were folded during the Late Permian when the two continents were welded into one land mass.

Continental and restricted basin sediments are characteristic of western and central Europe north of the Tethys, peninsular India, and most continents of the Southern Hemisphere. In Australia, marine sediments occur in several large sedimentary basins along the west coast and in a broad belt in the east, extending from Tasmania to northern Queensland. Continental sediments of Permian age are rare in the west, but in the eastern belt they generally form the upper part of the Permian sequence. The predominantly continental development of the Permian in the Southern Hemisphere and peninsular India has been called the Gondwana Facies.

Igneous rocks commonly are associated with sedimentary deposits in the geosynclinal areas and less commonly in epicontinental or shallow-shelf areas.

Like other geological systems the Permian is represented

Table 2: Standard Stage Sequences of the Permian in Some Important Areas

	Ural Mountains	Central and Southwestern Asia		Texas		New Zealand
Upper Permian	Tartarian	Pamirian	Dzhulfian	Ochoan		Makarewan
			Chhidruan			Waiitian
	Kazanian	Murghabian		Capitanian	Guadalupian	Purnhanan
	Kungurian	Kubergandian		Wordian		
Lower Permian	Artinskian	Darvarian	Artinskian	Roadian		Braxtonian
				Leonardian		Mangapirian
	Sakmarian / Asselian	Karachatyrian		Wolfcampian		Telfordian ?

Sedimentation and facies

by a great variety of sedimentary rocks. Limestone occurs widely in the miogeosynclines (shallow, near-continent, depositional troughs) and shallow shelves of the Northern Hemisphere. It includes ancient coral reef-like formations in the Guadalupe Mountains of New Mexico and Texas and in the subsurface of west Texas. Elsewhere, Permian reef formations are known from Germany, the eastern Alps, the western Urals, Caucasus, and from Chile.

The world's eugeosynclines (deep-sea depositional troughs associated with volcanic deposits), such as those situated in western North America, the Urals, and New Zealand, contain huge accumulations of clastic rocks mixed with products of extensive igneous activity, both intrusive and extrusive. Thick limestone sequences are also present locally. Areas of maximum deposition include west Texas (5,000 metres or 16,000 feet), the miogeosyncline of Nevada and Utah (4,000 metres or 13,000 feet), and parts of Central Asia: Tien Shan (3,900 metres or 12,800 feet); Pamir (3,700 metres or 12,100 feet).

Clastic sediments prevail almost entirely in the great basins on the eastern and western side of Australia. In the eastern basins, especially in New South Wales and Queensland, much igneous activity accompanied sedimentation; and basalt, andesite, and other volcanic rocks were formed. In the western basins, great columns of sedimentary rocks accumulated—*e.g.*, 4,600 metres or 15,000 feet of Permian marine rocks in the Carnarvon Basin of Australia. Clastic marine rocks of Permian age are also widely distributed in Southwest Africa, western South Africa, and eastern Argentina.

Tillite, a name applied to ancient glacial tills, is widely distributed in the Permian record. The term also has been applied to other glacial and nonglacial deposits, however, and this has caused some confusion.

Rocks of continental origin are mostly sandstone but include conglomerate and shale, and these are widely distributed in the southwestern United States (Supai Formation, Coconino Sandstone of Arizona, Yeso Formation of New Mexico and Texas) and in Europe (Rotliegendes of Germany, New Red of England, Verrucano of Switzerland). In peninsular India, the Permian is represented by more than 3,000 metres or 9,800 feet of continental beds, mostly sandstone and shale (lower part of Gondwana System), containing coal and some marine intercalations. A similar facies prevails in Africa south of the Sahara, most completely developed in South Africa (lower part of Karroo System), and in southern Brazil (Paraná Basin) and western Argentina (Precordillera).

PERMIAN LIFE

No break exists between Carboniferous and Permian life in the fossil record (*q.v.*). This statement is true for invertebrates, vertebrates, and, with some qualifications, plants. Evolutionary lineages of many marine invertebrate groups, such as fusulines, corals, brachiopods, mollusks, and ostracods, extend from the Carboniferous into the Permian. All families of the Fusulinacea present in the Carboniferous, for example, continued into the Permian. The same is probably true for corals. Of 66 families of brachiopods that existed in the Carboniferous, 46 survived into the Permian, and only three new families were added during that period. Lineages of nautiloid and ammonoid cephalopods were continuous across the boundary. Of the six trilobite families of the Carboniferous, four continued into the Permian.

A paleontological census published in 1958 revealed that 8,560 species and subspecies of invertebrate fossils had been described from Permian rocks up to that time. Nearly 30 percent of these are brachiopods, and more than 25 percent cephalopods.

Marine fishes of Permian age are not well known. Shark-like forms, which predominated, were survivors from Carboniferous time. Freshwater fishes are known in much greater variety, and most lineages extended from the Carboniferous Period into the Permian Period. There is no significant break between Pennsylvanian and Permian amphibian faunas.

Reptile evolution

Although reptiles arose in the Carboniferous, their first major evolution took place in the Permian. Three principal groups were present in the Permian: cotylosaurs, the stem group ancestors of Mesozoic dinosaurs and modern reptiles and birds; pelycosaurs, the bizarre sail-backed reptiles leading to the therapsids; and the therapsids, the mammal-like reptiles that lived in great numbers and variety in South Africa and European Russia.

The Lower Permian is characterized by a plant assemblage in which the descendants of Carboniferous ferns (pteridophytes) and seed ferns (pteridosperms) predominate. A sharp change in vegetation occurred at the transition from Early to Late Permian when conifers increased in numbers to lend the flora a decidedly Mesozoic aspect. It may be said that a "Mesophytic" era began in the Late Permian.

In eastern Asia and western North America (Texas), the predominant plant was the fern *Gigantopteris*. The southern continents and peninsular India are characterized by *Glossopteris*, a genus recently shown to have been related to the seed ferns.

Guide and zone fossils

The most useful guide and index fossils of the Permian are fusulinaceans and ammonoids. Where either one of these groups occurs in reasonable numbers, correlation of beds with international standard divisions is greatly facilitated. Both groups exhibit easily recognizable evolutionary trends—the shell size and shape and internal structure of fusulines and the shell shape and suture patterns of ammonoids.

Based on the occurrence of fusulines, Permian strata are divided into five biostratigraphic zones (see STRATIGRAPHIC BOUNDARIES). In a similar manner, six ammonoid zones may be distinguished (Table 3). The most abundant fossils in the Permian record, the brachiopods, were a slowly developing group and have limited use for long-range stratigraphic correlation. Vertebrate fossils are too rare and serve as guide fossils only in the Lower Beaufort Beds, South Africa.

Although the Rotliegende in Europe and the Dunkard Group in the United States are zoned by plants, the *Glossopteris* flora of the Gondwana area and eastern Siberia is too uniform in composition throughout the time of its

Table 3: Fusuline and Ammonoid Zones

stage	substage	fusulines	ammonoids
Dzhulfian		*Yabeina*	*Cyclolobus*
Chhidruan		*Verbeekina*	
Guadalupian	Capitanian	*Polydiexodina*	*Timorites*
	Wordian	*Parafusulina*	*Waagenoceras*
Artinskian			*Perrinites* *Metaperrinites*
Sakmarian		*Pseudoschwagerina*	*Properrinites*
Asselian			

existence for zonal sequence recognition. Plants are generally of limited use for zonation purposes in Permian rocks.

During the transition from the Carboniferous to the Permian Period, the evolution of animal and plant life proceeded evenly, without startling changes. An exception was the sudden appearance of the *Glossopteris* flora in the Gondwana area, where it replaced a Carboniferous flora of widespread type. Evolutionary rates in fusulines and ammonoids were rapid, making genera and species of these groups excellent zone fossils for this system. Trilobites continued their decline, which had begun in the Late Devonian, and became extinct before the end of Permian time. The corals showed a decline in diversity and numbers from their peak in the Devonian and Carboniferous.

At the end of Permian time, however, more spectacular evolutionary events occurred; and many forms of life, especially among invertebrates, became extinct, whereas others survived into Triassic and later times. The extinctions affected the faunas worldwide and the floras of the Gondwana area, where the *Glossopteris* flora gave way to the Triassic *Dicroidium* flora. In the Northern Hemisphere, as already stated, the principal change in floral evolution had taken place in the beginning of the Late Permian, and the conifer flora that had achieved predominance survived the Permian with little change.

Many major animal groups, entire orders and suborders, disappeared at the end of the Permian or survived only for a very short time thereafter. Most spectacular was the extinction rate of brachiopods. Of a total of between 130 and 140 genera present in Upper Permian rocks, only two survived into the Triassic (one of them only questionably represented in the Permian). Fewer than 30 new genera arose in the Early Triassic. In addition, a handful of genera survived locally and for a very short time only.

Among Foraminiferida, only the fusulines, of which 40 genera existed in the Late Permian, became extinct at the end of the period. Other foraminiferal suborders contain many genera that ranged virtually unchanged from the Paleozoic into the Mesozoic and even Recent.

Out of ten families of ammonoid cephalopods existing at the end of the Permian, only two survived in the Triassic for a short time, and these were constituted by different genera. On the other hand, out of eight families of non-ammonoid cephalopods, five survived into the Triassic and remained well established for varying times. At least three nautiloid genera are common to the Permian and Triassic. The Permian and Triassic representatives of many invertebrate groups such as sponges, bryozoans, and gastropods are not sufficiently well known to allow evaluation of their evolutionary behaviour.

The survival of the vertebrates is similar. How the irregularity of these extinction patterns can be explained is not entirely clear. Certainly no one single cause can be invoked, although suggestions have not been lacking: changes in ocean salinity, climatic change, neuroendocrinal reactions, geomagnetic polarity reversals, cosmic radiation, and supernovae have been suggested as causes. The most probable explanation is related to unusual combinations of factors that continuously influence the course of the evolution of life, such as transgressions and regressions (advance and retreat of seas), nonsedimentation, climatic changes, orogenies (mountain building), changes in ocean current patterns, cosmic radiation, mutations, genetic shifts, changing supplies of nutrients, and the appearance of predators, viruses, parasites, and pests of all kinds.

STRATIGRAPHIC BOUNDARY PROBLEMS

Stratigraphic boundaries (*q.v.*) within and between geologic systems are determined on the basis of zone fossils. Where rocks are formed in the marine environment, marine fossils, usually invertebrates, are used. In continental rock sequences, plant fossils usually serve this purpose.

Lower Permian boundary

In areas where Lower Permian marine rocks conformably overlie Upper Carboniferous (Pennsylvanian) rocks and where sedimentation was continuous from one period to the next, location of the boundary between the two systems is arbitrary. The midcontinent of the United States is such an area and, consequently, in the course of time numerous proposals for the placement of this boundary have been made. This boundary is also transitional in places in the western American geosyncline (*e.g.*, Oquirrh Formation of Utah) and in the subsurface of west Texas (*e.g.*, Horseshoe Atoll). Continuous Carboniferous–Permian sections exist, probably over wide areas, also in China and in the Soviet Union (western Urals).

Because of widespread crustal movement in the latter part of the Carboniferous and subsequent Early Permian transgressions of the sea, the lower boundary of the Permian is well marked in most places.

If fusulines or ammonoids are present in the strata near the boundary, its position can be determined with great confidence. Other groups such as sponges, brachiopods, and gastropods are of highly dubious value for zonation and correlation; where such groups prevail, the position of the boundary commonly is in doubt.

Because widespread emergence of land occurred during the Late Permian, rocks of this age are less widely distributed than Lower Permian rocks, or they are represented by strata of nonmarine facies. An upper boundary of the Permian System that is of worldwide validity can be determined only in stratigraphic sections in which uppermost Permian rocks are conformably overlain by lowermost Triassic rocks and where sedimentation was apparently continuous from the Permian into the Triassic Period. Few such sections are known (East Greenland, Armenia-Azerbaijan, central Iran, Pakistan, Kashmir, southern China). It is becoming apparent that in some of these areas a small vertical interval exists in which Permian-type and Triassic-type faunal elements are mixed. Ideas on the placement of the boundary therefore differ. Viewed on a worldwide scale, however, the inconsistencies are of minor nature and are being eliminated.

PERMIAN PALEOGEOGRAPHY

Permian paleogeography is one of the most vigorously disputed topics in contemporary geology. A majority of earth scientists today favour the hypothesis of continental drift (*q.v.*) in one form or another. Most of these authors reassemble the southern continents and peninsular India into one supercontinent (Pangaea, Gondwanaland), which is believed to have existed in Paleozoic times. Many different versions exist, however, that agree with each other in only one point, namely the fit of the opposing coasts of South America and Africa. Great differences in interpretation of the relative positions of all other continents exist.

Criticisms of the drift hypothesis that are relevant to the Permian record include the following:

1. On a continent that is larger in size than Asia at present, not enough moisture can exist in the atmosphere to allow major ice accumulations. The coldest place on earth today outside Antarctica is Verkhoyansk, in eastern Siberia, which is not glaciated now and was not glaciated during the Pleistocene.

2. Formation of coal deposits would suffer from the same moisture deficiency. The extensive distribution of Permian coal in Siberia, on the other hand, is in harmony with widespread marine transgressions in the same general area.

3. Paleozoic and Triassic rocks of marine character are widespread on the Gondwana continents, but their origin and presence are rarely mentioned and never explained.

Examples from the Permian are: beds with *Eurydesma* in eastern Argentina and Southwest Africa; a *Productus* (brachiopod) limestone fauna in Madagascar; marine intercalations with Tethys affinities in the Gondwana Series of India; thick sequences (up to 4,000 metres) of predominantly marine Permian rocks containing a Tethys fauna in the great sedimentary basins of Western Australia (overlain by marine Triassic rocks).

4. Vertebrates in Australia similar to those occurring in the Permian of South Africa and India are absent.

5. Distribution of the Cathaysian flora in east Asia, Sumatra, Turkey, and western North America is scattered.

6. A widespread Carboniferous flora changed to a provincial (restricted) Permian one.

These facts are reconcilable with modified continental relocation and with the assumption of stable continents. It is not possible to compile a Permian paleogeography that would satisfy every specialist and be compatible with all known data. Future work may resolve this difficulty.

One of the outstanding features of Permian paleogeography was the Tethys, a persistent marine belt that extended roughly between 30 and 40° N lat. from the western Mediterranean eastward and southeastward as far as Indonesia. A southern branch extended along the west coast of Australia as far as 30° S lat. The Tethys was complex, including both shallow- and deep-water regions.

Towards the northeast, Tethys connected with the epicontinental seas covering China and with the Japanese geosyncline of complex structure and history. Connected with Tethys to the north was the Uralian Geosyncline.

In central and western Europe and eastern North America continental and restricted marine conditions prevailed. Western North America was the site of a complex geosyncline, consisting of an eastern miogeosyncline and western eugeosyncline characterized by volcanic activity resembling that of island arcs (*q.v.*) of the East Asian type. This may have extended southward into Peru.

Toward the end of Permian time the sea withdrew from large areas which it had covered in the Early and early Late Permian. Desert conditions must have been widespread in North America and in Siberia during this latest Permian time.

Orogeny and volcanism

The Late Paleozoic was a time of widespread mountain building and volcanism, although in some parts of the world exact dating of these events is difficult. In eastern North America, the Appalachian Mountains were probably most intensely folded in the Permian; this Appalachian Revolution came to an end late in the period. In the west, there were probably continuous crustal movements in the Cordilleran Eugeosyncline involving intense igneous activity.

In Europe, orogenic and igneous activities were on a much smaller scale than they had been in the Carboniferous. Here, two minor orogenies (episodes of mountain building) are generally named Pfälzian (post-Early Permian) and Saalian (Late Permian). Volcanism was widespread in the Early Permian but died down in the Late Permian.

The Ural Mountains were folded in the late Permian and Europe became welded to Asia. The mobile belt of central Asia went through a complex history of foldings, metamorphism, and intensive and extensive igneous activity, which included granite intrusions in the Pamir, Kazakhstan, and elsewhere, and the eruption of acid and intermediate lavas, ignimbrites, and tuffs.

Similarly, in the Pacific mobile belt, diabase, diorites, and gabbros were formed and were penetrated by granite intrusions. In Japan, two orogenies occurred—one in the early Artinskian, one in the early part of Late Permian time. Mountain building also took place in Peru.

In Australia, folding and igneous activity were essentially restricted to Queensland and northern New South Wales, where the large New England Batholith was emplaced. In New Zealand, the Permian was characterized by intrusive and extrusive volcanic activity.

PERMIAN CLIMATES

The Permian Period began with widespread glaciations and ended in aridity. Sedimentary rocks of glacial origin, such as true tillites and fluvio-glacial deposits, laid down either in the sea or in lakes, are known from Argentina, southern Brazil, Uruguay, and the Falkland Islands; from many places in South Africa, including Southwest Africa, and from central Africa as far north as Lake Kivu and Gabon; from Madagascar; from peninsular India and from the Salt Range of Pakistan; in many places from the Australian continent and Tasmania; and from Antarctica. New Zealand is the only southern country with a near-complete record of Permian sedimentary rocks in which glacial sediments are absent.

Although the glacial series are found below or at the base of sedimentary sequences that are demonstrably of Permian age, the glacial rocks themselves are either unfossiliferous or contain specialized faunas. These consist of cold-water forms (*e.g., Eurydesma*) that are difficult to compare and correlate with faunas of the standard sections. It is, therefore, not known with certainty whether all the above-mentioned occurrences are in fact the results of one strictly contemporaneous event. The assumption that they were has been one argument set forth in favour of continental drift. By rearranging the southern continents and peninsular India in a manner first proposed by Afred Wegener, it is possible to reduce the area covered by glacial deposits to an allegedly credible size. The difficulty with this theory, as previously mentioned, is that on a very large continent (Gondwanaland) there would be insufficient moisture to allow formation of a huge ice cap.

It is now almost certain that not all these glacial beds were formed at the same time. Glacial deposits in Uruguay have been shown to be of mid-Pennsylvanian (Upper Carboniferous) age; those in the Salt Range of Pakistan are very Early Permian; and some occurrences in Australia (New South Wales, Tasmania) are of Artinskian, Kungurian, and even Tartarian age. The earliest glaciation in Australia was mountain glaciation of Middle Carboniferous age. Further, although many glacial deposits undoubtedly were laid down in water, others were not and are the result of highland and mountain glaciation rather than continental glaciation. This kind of origin is well established for the Congo, India, and eastern Australia. Lastly, the glacial origin of at least some of the deposits is not undisputed, and an interpretation of mud-flow origin is favoured by some investigators.

In summary, evidence furnished by the distribution and nature of glacial deposits does not necessarily favour the hypothesis that the southern continents and peninsular India formed one supercontinent in Permian time. It is possible that world temperatures in the Early Permian were comparatively low on a global scale. This postulate is supported by the conclusion that the climate of Central Asia at that time was cool temperate to cold, and it would explain occurrence of glacial deposits outside peninsular India—in the Salt Range, for example. In general, climatic conditions improved somewhat during the Permian; a warm, dry climate is indicated by extensive evaporite deposits in the Northern Hemisphere, and temperate conditions by widespread coal formation in the Southern Hemisphere.

BIBLIOGRAPHY. J.J. BIGARELLA, R.D. BECKER, and J.D. PINTO (eds.), "Problems in Brazilian Gondwana Geology," *Int. Symp. Gondwana Stratigraphy and Paleontology, Curitiba* (1967), a symposium containing two important articles on Permian stratigraphy in Brazil; P.J. DARLINGTON, *Biogeography of the Southern End of the World* (1965), contains a balanced discussion of continental drift and a reconstruction of Gondwanaland that avoids some of the problems posed by assuming one supercontinent; J.M. DICKENS (ed.), "The Permian of Australia," *Spec. Publ. Geol. Soc. Australia,* no. 2, pp. 1–116 (1969), a symposium of 12 articles bringing knowledge of the Permian of Australia up-to-date; C.O. DUNBAR *et al.*, "Correlation of the Permian Formations of North America," *Bull. Geol. Soc. Am.,* 71:1763–1806 (1960), authoritative correlation chart of all important Permian formations in North America; H. FLUEGEL, "Die Entwicklung des vorderasiatischen Paläozoikums," *Geotekt. Forsch.*, no. 18 (1964), includes data on Permian for Turkey, Iraq, Iran, and Saudi Arabia; B. KUMMEL and C. TEICHERT, "Stratigraphy and Paleontology of the Permian Boundary Beds, Salt Range and Trans-Indus Ranges, West Pakistan," in *Strati-*

graphic Boundary Problems: Permian and Triassic of West Pakistan, pp. 1–109 (1970), a study of relationships between the Permian and Triassic in a classical area in Pakistan; E.D. MCKEE *et al.*, "Paleotectonic Maps of the Permian System," *Misc. Geol. Investigations Map 1–450, U.S. Geol. Sur.* (1967), "Paleotectonic Investigations of the Permian System in the United States," *Prof. Pap. U.S. Geol. Surv. 515* (1967), these two works together offer a major monographic description of the Permian System in the U.S.; C. TEICHERT (ed.), "Symposium sur les séries de Gondwana," *Int. Geol. Cong., Rept.*, 19th sess., Algeria (1952), a symposium that is still authoritative for Africa, but out-of-date for Australia, Antarctica, and in part South America; J.B. WATERHOUSE, "Permian Stratigraphy and Faunas of New Zealand," *Bull. Geol. Surv. N.Z.*, n.s. (1967), up-to-date discussion of Permian stratigraphy and faunas in New Zealand.

(C.T.)

Pernambuco

Pernambuco, a state (*estado*) of Brazil, is located in the Northeast region of the country, near the eastern tip of South America's bulge into the Atlantic. It is bounded on the east by the ocean, on the south by the states of Alagoas and Bahia, on the west by Piauí, and on the north by Ceará and Paraíba. The region has long been the most economically depressed in Brazil. The area of the state is 37,947 square miles (98,281 square kilometres) and its population in 1970 was 5,252,590. The state capital is Recife. (See also the article RECIFE.)

History. In 1526 Cristóvão Jacques founded a trading post on the mainland near the Rio Iguaraçú. The place was abandoned in 1530 after a devastating attack from French interlopers, but it was restored by the Portuguese two years later. The first permanent settlement of Pernambuco was at Olinda in 1535 by Duarte Coelho, who the previous year had been granted a captaincy extending from the mouth of the São Francisco River northward to that of the Iguaraçú. The place where the present-day state capital of Recife is located offered a sheltered anchorage but was at first used only for loading and unloading cargo. During the Dutch occupation (1630–54), a well-planned town was built there that became, under the name Mauritzstad, the administrative capital. Adjacent to Duarte Coelho's grant, on the north, was the captaincy of Itamaracá, which covered the remainder of the present state. Pernambuco was ably governed and prospered with the production of sugar in the plantations established in the rich alluvial soil along the coast. It took an active part in the expulsion of the French from the trading posts northward to Maranhão. It was one of the only two captaincies that met financial success; the other was São Vicente, in the modern state of São Paulo. The Dutch were attracted to Pernambuco mainly by its valuable sugarcane lands. Although an active guerrilla warfare was waged against their administration during large part of their occupation, the Dutch did much to promote the commercial interests of the colony, especially under the able governorship of Count John Maurice of Nassau-Siegen.

A bitter rivalry developed between Olinda, the administrative capital of the captaincy and residence of the rich aristocratic plantation owners, and Recife, inhabited by traders, ship chandlers, and warehouse workers, mostly Portuguese. This rivalry came to a head in 1709, when the growing importance of Recife was recognized by the royal government, and it was given a separate municipal organization (*vila*). There followed some bloody clashes, known as the Peddlers' War, between the two factions, which lasted until 1711 when a new, and more conciliating, governor for the captaincy arrived. Recife continued to prosper while Olinda declined. In 1823 Recife was elevated to the category of city (*cidade*) and in 1827 was made the capital of the province, including Olinda in its municipal district.

19th-century history

In 1817 Pernambuco was the scene of a revolutionary outbreak that resulted in the temporary separation of the present states of Alagoas and Rio Grande do Norte, Ceará and Paraíba having been detached in 1799. There was another insurrection in 1822, when the Portuguese were expelled; and in 1824 dissatisfaction with the arbitrary proceedings of Dom Pedro I at Rio de Janeiro led to a separatist revolution for the formation of a new state, to be called the Confederação do Equador. There was another outbreak in 1831 and frequent disorders to 1848, which culminated in another unsuccessful revolution. The population of the Pernambuco *sertão* (the inland plateau) became noted for its turbulent, lawless character, due partly to the distance of the *sertão* from the populated coast, partly to poor communications, and partly to the fact that this region became the refuge of criminals from coast towns.

The natural environment. Pernambuco's Atlantic coastline, which extends approximately 116 miles (187 kilometres) from south to north, is relatively short when compared to the state's east–west extent of 447 miles (720 kilometres). The coastal zone is also comparatively narrow and is separated from the high inland plateau by an intermediate zone formed by the terraces and slopes between the two. Its surface is much broken by the remains of the ancient plateau, which has been worn down by erosion, leaving escarpments and ranges of flat-topped mountains, called *chapadas*, capped in places by horizontal layers of sandstone. The coastal zone is low, well wooded, and fertile. The coastal zone has a hot, humid climate, relieved to some extent by the southeast trade winds. This region is known locally as the *zona da mata* ("forest") though most of the forest is gone. The middle zone, called the *agreste* region, has a drier climate and lighter vegetation. The *sertão*, sometimes called *zona de caatinga*, is high, stony, and dry and frequently is devastated by prolonged droughts (*sêcas*). The climate of the *sertão* is characterized by hot days and cool nights. For the state as a whole there are two clearly defined seasons, rainy from March to July and dry for the remainder.

The rivers of the state include a number of small plateau streams flowing southward to the São Francisco River and several large streams in the eastern part flowing eastward to the Atlantic. The former are dry the greater part of the year. The largest of the coastal rivers are the Goiana, which drains a rich agricultural region in the northeast part of the state, the Capibaribe and the Beberibe, which flow eastward to the Atlantic at Recife, the Ipojuca, which reaches the coast south of Recife, the Sirinhaém, and the Una; a large tributary of the Una—the Rio Jacuípe—forms part of the boundary with Alagoas.

Population. Beginning in the 16th century and continuing through the colonial period, large numbers of Negro slaves were brought by the Portuguese from Africa to work the sugar plantations in Pernambuco. Thus, like the other former sugar-plantation districts, the population of the state includes a high percentage of Negroes and mulattoes, especially in the coastal region. As with most of the other northeastern states, the greater part of the population lives within 200 miles or less of the coast.

Cities and towns

Recife (*q.v.*), the capital, with a population (1970) of 1,046,454 is the largest city in the state. There are a number of other large towns, the more important being Olinda, the old colonial capital, which has a population of almost 190,000, and Caruaru, the largest city in the interior, with a population of 101,000.

The population density of about 140 persons per square mile (54 persons per square kilometre) makes it one of the six most densely populated of Brazil's states.

Administration and social conditions. The seat of the state government is at Recife, the capital, which is also the site of the principal government offices. The state has its own constitution and laws; the governor and members of the legislature are elected.

Ilha Fernando de Noronha, off the coast, for many years has been attached to the state for administrative purposes. Like the other states of the Northeast, Pernambuco has received considerable financial assistance from the federal government. A national department created to combat the drought in the nine northeastern states and parts of the interior state of Minas Gerais has built in those states a number of reservoirs to alleviate the effects of periodic drought. Direct federal assistance is channeled primarily through the Superintendência do Desenvolvimento do Nordeste (Sudene; Northeast Development Agency), with its main offices in Recife.

Living conditions are in general better in the coastal region than in the interior, as are public health and education. Elementary education is compulsory. Higher education in law, medicine, engineering, commerce, the fine arts, agriculture, and teacher-training is available in Recife at the Federal University of Pernambuco, the Catholic University of Pernambuco, and the Rural University of Pernambuco.

Economy. Pernambuco's resources are varied. Agriculture is the leading activity, although grazing, commerce, manufacturing, and mining also contribute to the economy.

Since the 16th century the coastal plain has been devoted to the production of sugarcane, largely for export abroad and to other areas of the state and nation. Tropical and semi-tropical fruits (bananas, oranges, coconuts) and food crops (sweet potatoes, maize, manioc, rice) are also grown commercially in the coastal zone. The principal crops of the *agreste* region are cotton, coffee, beans, manioc, onions, tomatoes, and maize. Tobacco, formerly an important export commodity, is now grown chiefly for local consumption. Livestock raising is the chief activity in the *sertão*, but in areas of adequate rainfall or irrigation, cotton, sisal, castor beans, and food crops are also produced.

Pernambuco's large herds of cattle and goats are the source of meat, hides, skins, and leather. The traditional manufactures are refined sugar, cotton cloth, and cigars. Sugar processing continues to be the major industry, although a broad range of other industries, such as steel manufacture and the processing of non-ferrous metals, have been introduced in response to tax incentives offered Brazilian and foreign firms.

Transportation. Although it serves a chronically depressed area, Pernambuco's port of Recife is one of the country's most important because of its location on the continent's bulge into the Atlantic. Its two-mile long breakwater is sheltered by reefs and breakwaters. It is equipped with cranes, conveyors, and warehouses serving the ships engaged in the import and export trade. The railways of the area converge on the port, as does the system of surfaced roads. Recife also is the site of an airport served by national and international airlines and has national and international cable, telegraph, telephone, and radio communications.

Location of the port of Recife

Cultural life and institutions. The state's cultural life and institutions also are concentrated in Recife. The major universities and other educational institutions are located there, as are the state museum and public library. It is also the site of the church of São Francisco de Assis, as well as other churches, the Santa Isabel theatre and many historic landmarks, including the old Dutch fort of Cinco Pontas and Restoration Square, which commemorates the end of Dutch domination. Olinda, a national monument, is of great touristic interest because of its old churches and houses, as well as the quaint streets that have preserved their colonial character.

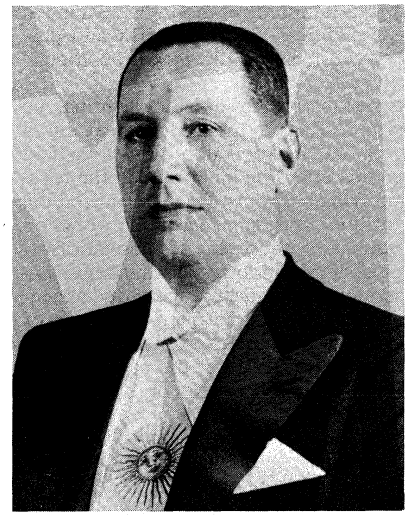
BIBLIOGRAPHY. The literature in English on Pernambuco is restricted largely to general works on Brazil and specialized monographs that deal with the Brazilian Northeast. Pernambuco's leading scholar, GILBERTO FREYRE, has written by far the most important study of the evolution of the traditional sugar plantation society and culture in *Casa-Grande y Senzala*, 12th ed., 2 vol. (1964; Eng. trans., *The Masters and the Slaves*, 2nd ed. rev., 1956; abridged ed., 1964). Significant episodes of the colonial history of Pernambuco may be found in C.R. BOXER, *The Dutch in Brazil, 1624–1654* (1957) and *The Golden Age of Brazil, 1695–1750* (1962). Twentieth-century developments are surveyed in T. LYNN SMITH, *Brazil: People and Institutions*, 4th ed. (1972), while alternative approaches to the contemporary problems of Pernambuco are examined in STEFAN H. ROBOCK, *Brazil's Developing Northeast* (1963); and JOSUE DE CASTRO, *Death in the Northeast* (1966).

(R.E.P./R.d'E.)

Perón, Juan

Whether in office as president of Argentina (1946–55), or in his subsequent long exile, Juan Domingo Perón stands out as one of the most remarkable Latin American political leaders of the 20th century. For good and for evil, he altered the society of his country, established himself as one of the earliest populist–nationalist–military leaders of the developing nations, and proved in exile that he could remain a significant political factor in his homeland. Perón's strength as a leader derived from his acute political opportunism, his sensitive attunement to the needs of Argentine social sectors (especially labour), and his personal magnetism.

By courtesy of the OAS



Perón, 1954.

Born on October 8, 1895, in a town on the Pampas of the province of Buenos Aires, Perón had a typical Argentine inheritance. He was a creole—a descendant of families that have been settled in Argentina for several generations—yet he also had non-Spanish European ancestors (Italian and French) as do most Argentines, for they live in a land dominated by immigrants and their descendants.

Early career

His early career was in many ways typical of the upwardly mobile, lower-middle class, male youth. He entered military school at 16 and made somewhat better than average progress through the officer ranks. But he had superior attributes. A strongly built six-foot-tall youth, Perón was the champion fencer of the army and a fine skier and boxer. He served in Italy during the late 1930s as a military attaché, and observed the successes of the Fascists and Nazis. He had a bent for history and political philosophy and published in those fields.

In 1943 he joined a clique of military plotters that overthrew the ineffective civilian government of Argentina, a government that had failed to solve the pressing problems of social and economic justice for a people ideologically split by the issues of World War II.

The military regimes of the following three years came increasingly under the influence of Perón, who had shrewdly requested for himself only the minor post of secretary of labour and social welfare. By 1945, he had also become vice president and minister of war. Clearly, he was bidding for undisputed power, based on the support of the underprivileged labourers (the descamisados, or shirtless ones) and on his popularity and authority in the army.

In early October 1945, Perón was ousted from his positions by a coup of constitutionally-minded civilians and officers. But his beautiful and dynamic mistress, Eva Duarte, and associates in the labour unions rallied the workers of greater Buenos Aires, and Perón was released from custody on October 17, 1945. That night, from the balcony of the presidential palace, he addressed 300,000 people, and his address was broadcast to the nation on radio. He promised to lead the people to victory in the pending presidential election and to build with them a strong and just nation. A few days later he married Eva, or Evita, as she was popularly called, who would help him rule Argentina in the years ahead.

After a campaign marked by repression of the liberal

opposition by the federal police and by strong-arm squads, Perón was elected president in February 1946 with 56 percent of the popular vote.

Perón set Argentina on a course of industrialization and state intervention in the economy, calculated to provide greater economic and social benefits for the working class. He also adopted a strong anti-United States and anti-British position, preaching the virtues of his Third Position—between Communism and capitalism—both domestically and in international relations. He extended Argentina's influence widely in Latin America and particularly sought to establish its political and ideological hegemony over South America.

If he did not structurally revolutionize Argentina, he did reshape the nation. Basing his government on a foggy doctrine he called Justicialismo, the President showered needed benefits upon the labourers, in the form of wage increases and fringe benefits. He nationalized the railroads and other utilities and financed public works on a large scale. The funds for those costly innovations—and for the graft that early began to corrode his regime—came from the foreign exchange accumulated by Argentine exports during World War II and from the profits of the state agency that set the prices for agricultural products. Perón dictated the political life of the nation by his command of the armed forces. He severely restricted and in some areas eliminated constitutional liberties.

Re-elected by a somewhat larger margin in 1951, Perón modified some of his policies. But he was overthrown and fled to Paraguay on September 19, 1955, after an army–navy revolt led by democratically inspired officers who reflected growing popular discontent with inflation, corruption, demagoguery, and oppression.

Exile to Paraguay

Perón finally settled in Madrid. There in 1961 he married for the third time (his first wife had died of cancer, as had Evita in 1952); his new wife was the former María Estela Martínez, an Argentine and a dancer whom he met in 1956. In his villa in the most fashionable suburb of the Spanish capital, the former dictator tended his roses, though he also followed a rigorous schedule of writing, interviews, and conferences. These were calculated to ensure, if not his return to Argentina, at least the eventual assumption of power by the millions of Peronista followers, whose memory of his regime improved with time and with the incapacity of the Argentine governments that succeeded Perón's decade of power.

In election after election the Peronistas emerged as a large, indigestible mass in the Argentine body politic. Neither the civilian nor the military regimes that precariously ruled in Argentina after 1955 were able to solve the relatively rich nation's condition of "dynamic stagnation," in part, because they refused to give political office to the Peronistas.

The military regime of General Alejandro Lanusse, which took power in March 1971, proclaimed its intention to restore constitutional democracy by the end of 1973 and allowed the re-establishment of political parties, including the Peronista Party. The long-hidden remains of Eva Perón—still a political issue because of the cult she had inspired—were exhumed by the government's orders and sent to Perón. Upon invitation from the military government, Perón returned to Argentina for a short time in November 1972. In the elections of March 1973, Peronista candidates captured the presidency and majorities in the legislature, and in June Perón was welcomed back to Argentina with wild excitement. In October, in a special election, he was elected president and, at his insistence, his wife—whom the Argentines disliked and, because of her obvious efforts to emulate Evita, resented—became vice president.

While in exile Perón had wooed the left-wing Peronistas and had supported the most belligerent labour unions. Once returned to power, however, he formed close links with the armed forces and other previously opposition right-wing groups. When he died—on July 1, 1974—he left his widow and successor as president an untenable situation. Divisions in the Peronista movement and in the government that had been present when Perón was alive were brought into the open. Mrs. Perón failed to obtain the firm support of any power group, not even the labour unions. Terrorist activity and political violence increased. By this time, too, inflation was running at an annual rate of more than 600 percent. On March 24, 1976, the armed forces took power, removed Mrs. Perón from office, and set up a military junta.

BIBLIOGRAPHY. A.P. WHITAKER, *Argentine Upheaval: Perón's Fall and the New Regime* (1956), subjective but still useful; T.F. MCGANN, *Argentina: The Divided Land* (1966), emphasizes psychological and geographical factors; J.R. BARAGER (ed.), *Why Perón Came to Power* (1968), varied views of the man and regime; H.S. FERNS, *Argentina* (1969), a penetrating British view of Perón's economic and political failures.

(T.F.McG./Ed.)

Perpetual Motion

Perpetual motion describes the action of a device that once set in motion would theoretically continue in motion indefinitely without receiving any additional energy from an external source. The perpetual motion machines most frequently proposed by inventors are not designed simply to stay in motion but are meant as well to be capable of delivering useful work. Since, in theory, such machines require no external energy source, they hold out the hope of delivering virtually free and limitless quantities of power. No perpetual motion machine that has ever actually been built has ever run or produced power for any significant period of time; no perpetual motion machine that has ever been proposed has successfully withstood careful scientific scrutiny. It must be noted, however, that over the past five centuries a number of promoters have built devices that gave the illusion of perpetual motion, which they used to enhance either their reputations or their pocketbooks.

Perpetual motion machines of the first kind. The reason perpetual motion machines fail to work is that their principles must inevitably violate either the first or second law of thermodynamics. The first law of thermodynamics states that energy is neither created nor destroyed, but is always conserved. It was proposed about the middle of the 19th century by three scientists working independently along different lines of attack. Julius Robert von Mayer (1814–78) of Germany based his reasoning on observations of the human body, having observed that the body takes in fuel (food) and converts it to work and heat. James Prescott Joule (1818–89) of England was able to show the experimental equivalence of work and heat by way of a cleverly designed apparatus that converted mechanical work into heat. And lastly, Hermann von Helmholtz, also of Germany, convinced the international scientific community that the principle of energy conservation was valid by first assuming the impossibility of a perpetual motion machine.

In 1847, when he was 26, Helmholtz presented his formulation of the first law before the Physical Society of Berlin. Beginning his analysis by declaring that perpetual motion machines were axiomatically impossible, Helmholtz pointed out that no one had yet built a successful perpetual motion machine, and, proceeding from that axiom, next showed that the failure of perpetual motion machines had to lead logically to the conclusion that energy is conserved. He went on to demonstrate that both heat and work were forms of energy and that what was conserved was the total of the two forms rather than either heat or work taken separately. Furthermore, Helmholtz showed that the findings of Joule's experiments were in general agreement with calculations of the kind made by Mayer.

Helmholtz' formulation

A perpetual motion machine that purports to deliver more energy from a falling or turning body than is required to restore the device to its original state is called a perpetual motion machine of the first kind. Most of the devices most often proposed by inventors through the centuries have been of this kind, and the oldest recorded scheme for achieving perpetual motion is such a device. It was suggested by Villard de Honnecourt, a 13th-century French architect, who so far as is known was the first man to suggest the idea of an overbalanced wheel (Figure 1). The inventors of overbalanced wheels believe that they can ex-

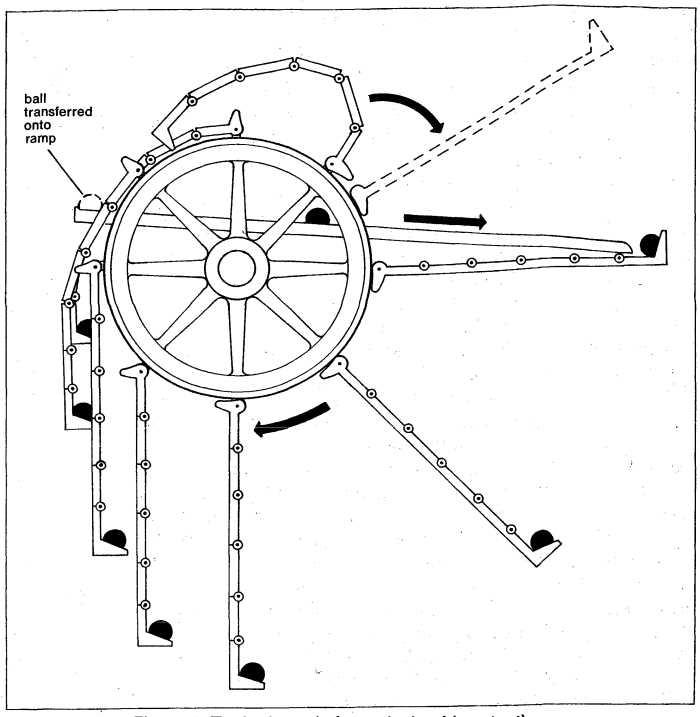


Figure 1: Typical overbalanced wheel (see text).
From S. Angrist, *Perpetual Motion Machines*, Copyright © 1968 by Scientific American, Inc.; all rights reserved

tract more energy from falling weights on one side of the wheel than is required to raise those weights on the other. This belief is based on the assumption that such weights deliver more effort if they are farther removed from the centre of rotation than if they are closer.

Two of the earliest and best known inventors of would-be perpetual motion machines based on overbalanced wheels were Edward Somerset, 2nd marquis of Worcester (1601–67), and Jean Ernest Elie Bessler (1680–1745), who was generally known as Orffyreus. Each of these men built rather large wheels that were capable of giving impressive demonstrations of apparent perpetual motion due probably to their large inertia which kept them turning for long periods of time once set into motion. Within a few years of 1640, probably around 1638 or 1639, Somerset operated his device for Charles I and his court. A Dutch physicist, W.J. Gravesande (1688–1742), inspected Orffyreus' last wheel and in many ways was impressed with its construction. In a long and detailed letter to Sir Isaac Newton, he described the externals of the wheel. On grounds that he did not want to be robbed of the secret of perpetual motion that he had discovered, Orffyreus refused to allow Gravesande to inspect its interior.

Before the end of the 14th century almost every English manor that was situated on a stream, roughly, that is, a third of all the manors in the Domesday Book, had its own water mill to grind flour, saw wood and full woollens. Villagers and townspeople who had no access to running water naturally sought alternative sources of power. The solution to this problem most commonly suggested was the closed-cycle water mill such as that proposed by the English physician Robert Fludd in 1618 (Figure 2). If the water that turns a mill wheel could be collected from the race at the foot of the wheel and somehow put back into the reservoir above the wheel, the need for a source of running water would disappear. It was reasoned that the wheel could drive a pump that would recycle the mill's water supply. At the time Fludd proposed this device and for more than 200 years after, there was no scientific reason to believe that it would not work, though, of course, all practical efforts failed to achieve perpetual motion by this route. All of the countless closed-cycle mills that have been proposed down through the centuries were supposed to work because their inventors believed that the falling water would give up more energy than was required to return the water to the race above the wheel. Such devices are another variation of the overbalanced wheel and thus constitute perpetual motion machines of the first kind.

John Wilkins, bishop of Chester and an early official of the Royal Society, envisioned in the 1670s three natural power sources that might be harnessed to provide perpetual motion. These in his words, were "Chemical Extractions," "Magnetical Virtues," and "the Natural Affection of Gravity." Wilkins' third power source embraces the entire family of overbalanced wheels. He specifically mentioned only one formula for chemical extraction; its underlying concept may have arisen from a misunderstood observation of the ceaseless motion of small particles visible in a fluid that is known today as Brownian movement. Wilkins also designed, but almost certainly never tried to build, a machine to utilize magnetic attraction. A lodestone on top of a pedestal was supposed to pull a bullet up an inclined ramp where it would fall through a hole and roll back to its starting position. Of course, any magnet strong enough to pull the bullet up the ramp would not let it drop through the hole.

Wilkins' magnetic device was probably the first in a long series of devices designed to use electric or magnetic forces to produce perpetual motion. Machines that have been proposed more recently usually have some mechanical arrangement designed to interpose a shield to isolate one portion of the machine temporarily from an existing magnetic or electric field. Invariably inventors of such devices fail to account for the work required to push the isolator through the field and into place in order that isolation may be achieved.

Would-be perpetual-motion-machine inventors have also directed their attention toward machines that would supposedly make use of hydrodynamic and capillary principles. In 1686 the Abbé de la Roque, editor of the *Journal des sçavans*, suggested as a source of perpetual motion the use of a beaker-like vessel that gives its liquid contents a large surface area on top and narrows to a small cross section on the bottom. A small tube running from the bottom of the beaker along the outside was permitted to discharge on to the top surface of the liquid in the beaker. De la Roque hoped (in vain) that the water in the wide section at the top would force liquid to flow continuously out of the bottom of the beaker and be returned to the top by means of the narrow tube. Sir William Congreve, a member of Parliament in the early 19th century, also proposed to use water to achieve an eternal source of free power. His scheme was based on an inclined plane that had a continuous chain of sponges looped around it; the sponges were forced to pass through a water bath. On top of the sponges a chain of heavy weights was fastened. The plan was that while one section of the sponges was filling with water thereby making that section heavier, another portion of the sponges was being squeezed dry by the heavy weights thereby making them lighter. Congreve's hope was that since one section of the sponge chain would be perpetually heavier than the section that was being squeezed dry, the chain would be kept in motion around

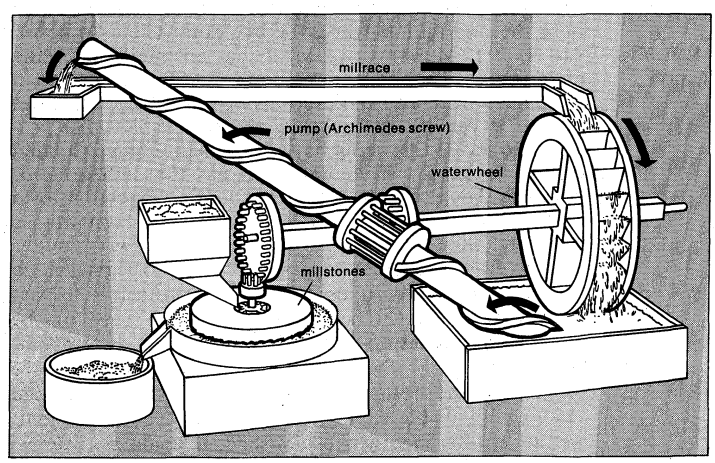


Figure 2: Closed-cycle water mill.

some appropriately arranged pulleys. The plan obviously failed to allow for the work that would have to be exerted on the sponges to remove the water from them.

Perpetual motion machines of the second kind. There is, however, one class of perpetual motion machines that is fundamentally different from the ones just described; they do not violate the first law in that energy in these machines is always conserved. They are impossibilities nonetheless because they attempt instead to circumvent the second law of thermodynamics. The foundations of the second law were laid down by French physicist Sadi Carnot (1796–1832) who wanted to find out whether improved design could eliminate all energy losses in steam engines. Although the first law states that a fixed amount of mechanical work can always be converted into the equivalent amount of heat, the most casual observation of a heat engine in operation makes plain that the reverse of the first law's axiom is not precisely true: a fixed amount of heat cannot be completely converted into the same amount of work. Some of the initial energy is unavoidably wasted. After much thought Carnot realized that a complete conversion of heat into work was impossible. It is this law that inventors of perpetual motion machines of the second kind try to violate.

Gamgee's zeromotor

One attempt to construct such a device was made by John Gamgee, a professor from London working in Washington, D.C. during the 1880s. He developed a heat engine that was much like an ordinary steam engine; he called it a zeromotor because its normal operating temperature was zero degrees centigrade. This engine used ammonia instead of water as the working fluid; liquid ammonia vapourizes into a gas at a low temperature, and at 0° C (32° F) the gas exerts a pressure of four atmospheres. Gamgee reasoned that the transfer of heat from the environment, rather than the energy supplied by the combustion of fuel, would be enough to transform the ammonia working fluid from a liquid to a gas. He reasoned further that the ammonia gas on driving the piston back and expanding, would cool, condense, and drain into a reservoir, whereupon the cycle could begin again. Gamgee and his supporters, though undoubtedly sincere, had either incorrectly calculated or had failed to calculate at all the zeromotor's temperature requirements. The heat transfer from the environment was indeed sufficient to convert ammonia from a liquid to a gas, but this advantage is nullified in the system as a whole by the cooling of the gas on expansion. Starting at 0° C and a pressure of four atmospheres, the temperature of the gas falls to −33° by the time its volume has quadrupled. If the gas is to condense into a liquid, both the condenser and the reservoir must be at a temperature lower than −33°. Gamgee had not provided for this cooling, and if he had, the cooling process would of course have required more energy than the zeromotor could produce.

The Gamgee zeromotor attempted to get work from thermal energy already at the temperature of the surroundings, which the second law expressly forbids. That statement of the second law that denies the possibility of such devices has been assaulted many times but no assault has been more persistent nor none taken more seriously than the one initiated in 1871 by James Clerk Maxwell (1831–79). Maxwell suggested that an intelligent being endowed with superordinary powers and who was small enough to deal with the motions of individual molecules would be capable of violating the second law. He proposed to locate this being he called a demon in an isolated enclosure that initially contained a gas at a uniform temperature. The enclosure was divided into two sections by a partition with a small door in it and the demon was to open and close the door to permit molecules with greater than average velocities to pass from left to right and to permit molecules with less than average velocities to pass from right to left. The higher the mean speed of the molecules making up a gas, the higher the gas temperature. The gas in the right-hand half of the enclosure would become hotter and the gas in the left-hand side colder without the expenditure of energy. Once the temperature difference was established, it could be used to drive a heat engine that would deliver useful work.

Maxwell's demon

Maxwell's demon bothered physicists for almost 100 years until Louis M. Brillouin, a French physicist, pointed out that the demon needed information about the individual molecules if he was going to open and close the door at the appropriate time. As the demon peers into either side of the isolated enclosure at uniform temperature, the uniformity of radiation throughout does not permit him to see anything, unless he is supplied with a source of radiation to illuminate the molecules. Brillouin showed that the energy that must be expended in the radiation source to illuminate the molecules is greater than any work that could be derived from a heat engine that could be made to operate off the temperature difference that the demon produced by means of his sorting abilities; thus Maxwell's demon does not violate the second law and does not constitute a perpetual motion machine.

Perpetual motion machines of the third kind. Sometimes the term "perpetual motion of the third kind" may be used to refer to the continuous motion that a device would have if it were completely freed from the action of all nonconservative forces, that is, forces such as electrical resistance or friction; the forces that always irreversibly remove energy from a system. Experience has shown that such forces can be reduced, for example, by lubrication in the case of friction, but can never be eliminated without the expenditure of additional energy. Studies of materials at very low temperatures have revealed that certain dissipative phenomena like friction or electrical resistance can be made to disappear for as long as the material is maintained below certain very low transition temperatures. For example, 24 pure metals and many alloys and intermetallic compounds are superconductors—that is, their electrical resistance disappears completely at their transition temperatures, usually below about 20° K. Zero electrical resistance can be demonstrated by inducing a current in a closed ring of superconductive metal; by placing a compass needle near the ring and noting its deflection one can demonstrate that this current will flow in the ring as long as the metal is kept below its transition temperature. Such persistent currents have been maintained for more than a year without diminishing in strength.

Persistent currents

A similar phenomenon has been observed with liquid helium below a temperature of 2.18° K, called its lambda point, or transition temperature. At any temperature below this point liquid helium loses all resistance to flow through capillaries of less than 10^{-4} centimetres in diameter. A current allowed to flow continuously in a superconducting ring or helium in the superfluidic state allowed to flow through a capillary loop does not constitute perpetual motion because considerable energy must be expended to maintain such systems at the extremely low temperatures required to cause the material to lose its resistance to flow. The energy expended on cooling is always greater than any work that might be derived from the superconducting or superfluidic flows.

Fictitious perpetual motion machines. A number of schemes have been proposed (and many of them realized) to utilize the radiant energy of the sun to produce systematic mechanical motion. For example, clocks that wind themselves by making use of changes in atmospheric temperature or pressure are really using some of the energy that the sun delivers to the earth each day. Because such devices do receive energy from an external source they do not constitute perpetual motion machines. Early in the 20th century R.J. Strutt (later the fourth baron Rayleigh) caused a sensation when he constructed an apparatus he called a radium clock. As the radium contained within the glass structure decayed by emitting alpha particles, it caused thin gold leaves also located within the structure to become charged. After a while the leaves discharged and moved to the collapsed position; the process repeated indefinitely. The half-life of radium is 1,620 years, so the motion of the clock would continue a long time. For a time it was thought that such a clock was really a perpetual motion machine but as the atom became better understood it was realized that the energy for operating the leaves is released as the radium nucleus is transformed from one unstable chemical element to another; thus the

radium clock has an energy source, though it is considerably different from the mainspring found in most watches and clocks.

As early as 1775 the French Academy of Sciences passed a resolution refusing to entertain any future communications concerning perpetual motion. The British Patent Office has also had a long-standing policy against reviewing patent applications for perpetual motion machines. The U.S. Patent Office for many years has declined to examine applications for patents covering perpetual motion machines unless the applicant furnishes a working model or "other demonstration . . . of the operativeness of the invention," a ruling that has produced much hostile correspondence but no working models.

In spite of such official opposition public sophistication regarding the possibility of building perpetual motion machines was slow to develop. One of the most ingenious and certainly the longest-lived swindles involving a supposed perpetual motion machine began in 1874, when

John E.W. Keely

John E.W. Keely unveiled a combined "generator" and engine at his home in Philadelphia. Keely's generator was an elaborate combination of metal globes, tubes, petcocks, nozzles, valves, and gauges. For almost 25 years Keely brought a steady stream of investors into his laboratory to demonstrate the "power" his generator could extract from water. "People have no idea of the power in water," Keely would say. "A bucket of water has enough of this vapor to produce a power sufficient to move the world out of its course." With time his demonstrations with his apparatus became more polished and indeed were so successful that Keely and his associates were able to capitalize the Keely Motor Company for over $1,000,000, with much of this amount coming from New York businessmen. Keely died in 1898. The son of one of his major backers promptly rented his Philadelphia house and explored the premises in the company of reputable witnesses. They found concealed beneath the floor a giant metal tank evidently used as a reservoir for compressed air. In the walls was located a network of brass tubing that conducted the compressed air to the generator. In the course of his long career Keely may have broken a number of laws, but he left the first and second laws of thermodynamics, which forbid perpetual motion machines, inviolate.

BIBLIOGRAPHY. S.W. ANGRIST, "Perpetual Motion Machines," *Scient. Am.*, 218:114–122 (1968), and with L.G. HEPLER, *Order and Chaos: Laws of Energy and Entropy* (1967), an introduction to the laws of thermodynamics, written for the interested layman; M.W. ZEMANSKY, *Heat and Thermodynamics*, 4th ed. (1957), a college-level introductory text in thermodynamics; H. DIRCKS, *Perpetuum Mobile; or Search for Self-Motive Power*, 2 vol. (1861–70), a historical review of the attempts to achieve perpetual motion during ancient times; S.R. SMEDILE, *Perpetual Motion and Modern Research for Cheap Power* (1962).

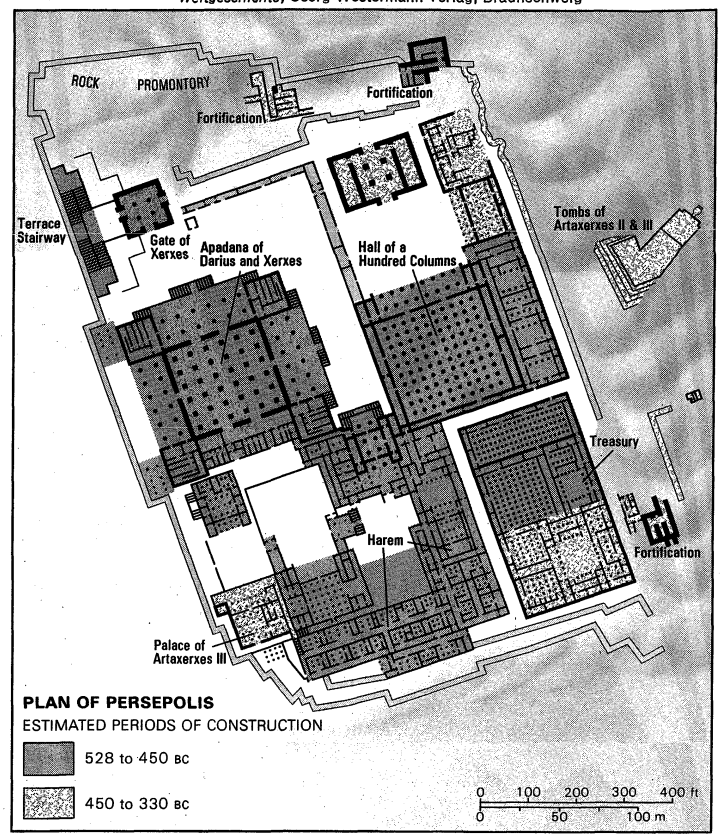
(S.A.)

Persepolis

Persepolis (Old Persian Parsa) was an ancient capital of the Achaemenian kings of Iran. Its site, now called Takht-e Jamshīd (the Throne of Jamshīd, a mythical Persian hero), is located about 32 miles northeast of Shīrāz in the *ostān* (province) of Fars, southwest Iran, near the confluence of the small river Pulvār (Rūdkhāneh-ye Sīvan) with the Rūd-e Kor.

History. Though archaeologists have discovered evidence of prehistoric settlement, inscriptions indicate that the construction of the city began under Darius I the Great (reigned 522–486 BC), who, as a member of a new branch of the royal house, made Persepolis the capital of Persia proper, replacing Pasargadae, the burial place of Cyrus the Great. Built in a remote and mountainous region, Persepolis was an inconvenient royal residence and was visited mainly in the spring. Effective administration of the Achaemenian empire was carried on from Susa, Babylon, or Ecbatana. This accounts for the fact that the Greeks were not acquainted with Persepolis until Alexander the Great's invasion of Asia. In 330 BC Alexander plundered the city and burned the palace of Xerxes, probably to symbolize the end of his Panhellenic war of revenge. In 316 BC Persepolis was still the capital of Persis as a province of the Macedonian empire. The city gradually declined in the Seleucid period and after, its ruins attesting its ancient glory. In the 3rd century AD the nearby city of Istakhr became the centre of the Sāsānian Empire.

Adapted from *Westermann Grosser Atlas zur Weltgeschichte*, Georg Westermann Verlag, Braunschweig



City of Persepolis.

Archaeology. The site is marked by a large terrace with its east side leaning on Kūh-e Raḥmat (the Mount of Mercy). The other three sides are formed by a retaining wall, varying in height with the slope of the ground from 13 to 41 feet; on the west side a magnificent double stair in two flights of 111 easy stone steps leads to the top. On this terrace are the ruins of a number of colossal buildings, all constructed of a dark gray stone, often polished to the consistency of marble, from the adjacent mountain. The stones, of great size, cut with the utmost precision, were laid without mortar, and many of them are still in place. Especially striking are the huge columns, 13 of which still

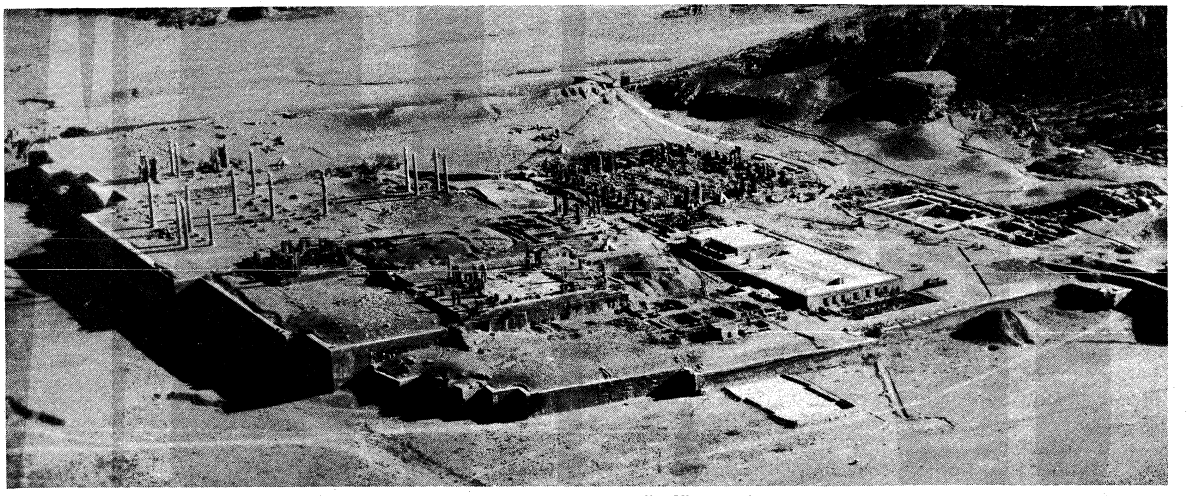
The 100 Columns

stand in Darius the Great's audience hall, known as the apadana, the name given to a similar hall built by Darius at Susa (there are two more columns in the propylaeum, or entrance hall, of his son, Xerxes). These ruins have been named "Sad-Sotūn" (the 100 Columns).

In 1933 two sets of gold and silver plates recording in the three forms of cuneiform, Ancient Persian, Elamite, and Babylonian, the boundaries of the Persian Empire were discovered in the foundations of Darius' hall of audience. A number of inscriptions, cut in stone, of Darius I, Xerxes I, and Artaxerxes III indicate to which monarch the various buildings are to be attributed. The oldest of these on the south retaining wall gives Darius' famous prayer for his people: "God protect this country from foe, famine and falsehood." There are numerous reliefs of Persian, Median, and Elamite officials, and 23 scenes separated by cypress trees depict representatives from the remote parts of the empire who, led by a Persian or a Mede, made appropriate offerings to the king at the national festival of the vernal equinox.

Cliff tombs

Behind Takht-e Jamshīd are three sepulchres hewn out of the mountainside; the facades, of which one is incom-



The ruins of Persepolis from the south, showing the terrace and retaining walls. The apadana of Darius I, of which 13 stone pillars are still standing, and the propylaeum of Xerxes, marked by two pillars, are at the top left.
By courtesy of the Oriental Institute, University of Chicago

plete, are richly ornamented with reliefs. About eight miles north by northeast, on the opposite side of the Pulvār, rises a perpendicular wall of rock in which four similar tombs are cut at a considerable height from the bottom of the valley. This place is called Naqsh-e Rostam (the Picture of Rostam) from the Sāsānian carvings below the tombs, which were thought to represent the mythical hero Rostam. That the occupants of these seven tombs were Achaemenian kings might be inferred from the sculptures, and one of those at Naqsh-e Rostam is expressly declared in its inscriptions to be the tomb of Darius I, son of Hystaspes, whose grave, according to the Greek historian Ctesias, was in a cliff face that could be reached only by means of an apparatus of ropes. An inscription beside the entrance to Darius' tomb describes his character and states that God had given him two special qualities, wisdom and activity. The three other tombs at Naqsh-e Rostam, besides that of Darius I, are probably those of Xerxes I, Artaxerxes I, and Darius II. Xerxes II, who preceded Darius II, reigned for a very short time and could not have obtained so splendid a monument; still less could Sogdianus (Secydianus), who usurped his brother's throne and reigned for less than a year. The two completed graves behind Takht-e Jamshīd probably belong to Artaxerxes II and Artaxerxes III. The unfinished one might be that of Arses, who reigned at the longest two years, but is more likely to be that of Darius III, the last of the Achaemenian line, who was overthrown by Alexander the Great.

BIBLIOGRAPHY. E. FLANDIN and P. COSTE, *Voyage en Perse*, 2 vol. (1851); F. STOLZE, *Persepolis: Die Achaemenidischen und Sasanidischen Denkmäler und Inschriften von Persepolis* (1882); G. PERROT and C. CHIPIEZ, *Histoire de l'art dans l'antiquité*, vol. 5 (1890; Eng. trans., *History of Art in Persia*, 1892), with a description of the platform and the Propylaeum of Xerxes at Persepolis; G.N. CURZON, "Persepolis and Other Ruins," in *Persia and the Persian Question*, vol. 2 (1892); F. SARRE and E. HERZFELD, *Iranische Felsreliefs* (1910); A. SAMI, *Persepolis*, 6th ed. (Eng. trans. 1970); D.N. WILBER, *Persepolis: The Archaeology of Parsa Seat of the Persian Kings* (1969), with full bibliography of works on Persepolis.

(R.N.S.)

Persian Gulf

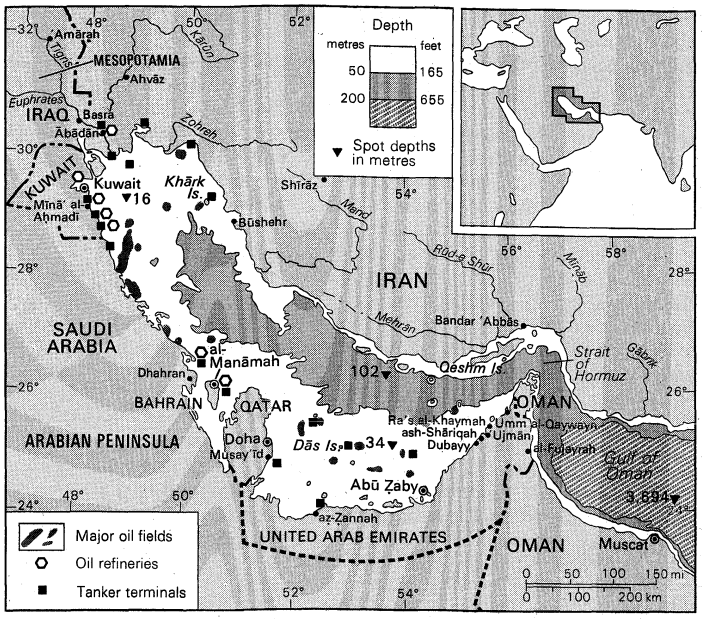
The Persian Gulf (known to the Arabs as the Arabian Gulf) is the shallow marginal sea of the Indian Ocean that lies between the Arabian Peninsula and southeast Iran. The sea has an area of 92,500 square miles (240,000 square kilometres). Its length is 615 miles, and its width varies from a maximum of 210 miles to a minimum of 35 miles in the Strait of Hormuz. It is bordered on the north, northeast, and east by Iran, on the northwest by Iraq and Kuwait, on the west and southwest by Saudi Arabia, Bahrain, and Qatar, and on the south and southeast by the United Arab Emirates and part of Oman. The term Persian Gulf is often used to refer not only to the Persian Gulf proper but also to its outlets, the Strait of Hormuz and the Gulf of Oman, which open into the Arabian Sea. This article deals with the Persian Gulf alone (see also TIGRIS–EUPHRATES RIVER SYSTEM).

Physiography. The Iranian shore is mountainous, and there are often cliffs; elsewhere a narrow coastal plain with beaches, intertidal flats, and small estuaries borders the gulf. The coastal plain widens north of Būshehr on the eastern shore of the gulf and passes into the broad deltaic plain of the Tigris, Euphrates, and Kārūn rivers. Cliffs are rare on the Arabian shore of the gulf, except around the base of the Qatar Peninsula and in the extreme southeast around the Strait of Hormuz. Most of the Arabian shore is bordered by sandy beaches, with many small islands enclosing small lagoons.

The gulf is very shallow, rarely deeper than 300 feet, although depths exceeding 360 feet are found at its entrance and at isolated localities in its southeastern part. It is noticeably asymmetrical in profile, with the deepest water occurring along the Iranian coast and a broad shallow area, which is usually less than 120 feet deep, along the Arabian coast. There are numerous islands, some of which are salt plugs or domes and others merely accumulations of coral and skeletal debris.

The Persian Gulf receives very small amounts of river-borne sediment except in the northwest, where the Tigris, Euphrates, and Kārūn, together with other small streams, empty into the gulf. The rivers reach their peak flow in spring and early summer, when the snow melts in the mountains; disastrous floods sometimes result. There are some ephemeral streams on the Iranian coast south of Būshehr, but virtually no freshwater flows into the gulf on its southwest side. Large quantities of fine dust are, however, blown into the sea by predominant northwest winds from the desert areas of the surrounding lands. Biological, biochemical, and chemical processes lead to the production of considerable calcium carbonate in the form of skeletal debris and fine mud. The deeper parts of the Persian Gulf adjacent to the Iranian coast and the area around the Tigris–Euphrates Delta are mainly floored with gray-green muds rich in calcium carbonate. The shallower areas to the southwest are covered with whitish-gray or speckled skeletal sands and fine carbonate muds. Often the sea floor has been hardened and turned to rock by the deposition of calcium carbonate from the warm, salty waters. Chemical precipitation is abundant in the coastal waters, and sands and muds are produced that mix with the skeletal debris of the local sea life. These sediments are thrown up by the waves to form coastal islands that enclose lagoons. The high salinities and temperatures result in the precipitation of calcium sulfate and sodium chloride to form extensive coastal salt flats.

Sea floor



The Persian Gulf.

Geology. The present-day Persian Gulf, together with its northwestern continuation now infilled by the deposits of the Mesopotamian rivers, is thought by some to be the remains of a once much larger northwest-to-southeast-aligned basin of deposition that existed throughout a large part of geological history. In this basin many thousands of feet of sediments accumulated, consisting mostly of limestone and marls (loose, crumbling earthy deposits containing calcium carbonate), together with evaporites and organic matter, which ultimately produced the area's vast oil resources.

Climate. The gulf has a notoriously bad climate. Temperatures are high, though winters may be quite cool at the northwestern extremities. The sparse rainfall occurs mainly as sharp downpours between November and April and is higher in the northeast. Humidity is high. The little cloud cover is more prevalent in winter than in summer. Thunderstorms and fog are rare, but dust storms and haze occur frequently in summer. The wind blows predominantly from a north–northwest direction —the so-called *shamāl* (Arabic: "north wind")—is seldom strong, and rarely reaches gale force. Squalls and waterspouts are common in autumn, when winds sometimes reach speeds of 95 miles per hour in as little as five minutes. Intense heating of the land adjacent to the coasts leads to gentle offshore winds in the mornings and strong onshore winds in the afternoons and evenings.

Hydrography. The small freshwater inflow into the gulf is mostly from the Tigris–Euphrates–Kārūn rivers. The water temperatures range from 75° to 90° F (24° to 32° C) in the Strait of Hormuz to 60° to 90° F (16° to 32° C) in the extreme northwest. These high water temperatures and a low influx of freshwater result in evaporation in excess of freshwater inflow; high salinities result, ranging from 37 to 38 parts per thousand in the entrance to 38 to 41 parts per thousand in the extreme northwest. Even greater salinities and temperatures are found in the waters of the lagoons on the Arabian shore.

Tides

The tidal range varies from about four to five feet around Qatar and increases to ten to 11 feet in the northwest and nine to ten feet in the extreme southeast. When onshore winds are strong, the level of the coastal waters, particularly in the southern gulf, may rise by as much as eight feet, causing extensive flooding of the low coastal plains. Tidal currents are strong (five miles per hour) in the entrance of the gulf but elsewhere—except between islands or in estuaries and lagoon entrances—rarely exceed one to two miles per hour. The wind affects local currents and may sometimes even reverse them.

Waves rarely exceed ten feet in height and are largest in the southern gulf. The swell from the Indian Ocean affects only the water at the entrance of the gulf; when it is opposed by wind, very turbulent conditions result. The general circulation pattern appears to be caused by water entering from the Indian Ocean, evaporating, becoming denser, and sinking to flow out into the Indian Ocean beneath the inflowing open ocean water.

The waters of the area support many plants and animals, but the high temperatures and salinities lead to a diminution in the variety of forms; many Indian Ocean forms penetrate only a small way into the gulf.

Economic resources. Until the discovery of oil in Iran in 1908, the Persian Gulf area was important mainly for fishing, pearling, the building of dhows (Arab lateen-rigged boats), sailcloth making, camel breeding, reed-mat making, date growing, and the production of other minor products, such as red ochre from the islands in the south. The arid lands surrounding the gulf produced little else and, except for the rich alluvial lands of the Mesopotamian plain, supported a small population of fishermen, date growers, and nomads. Today these traditional industries have declined, and the economy of the region is dominated by the production of oil. The Persian Gulf and the surrounding countries produce approximately 31 percent of the world's total oil production and have 63 percent of the world's proved reserves. The area has been explored only in a preliminary way, and there are still fewer drill holes in the area than are put down in a single month in the United States. Offshore exploration below the shallow waters of the gulf has revealed the presence of large reserves of oil and gas. These discoveries have led to numerous legal wrangles between states about exact territorial limits. Large quantities of oil are refined locally at Ābādān, in Iran, in Bahrain, and elsewhere, but most of it is exported to northwestern Europe and other parts of the world as crude oil.

Oil

Other exploitable mineral deposits appear to be rare, but only cursory surveys have been made; exploration was actively being pursued in the area in the early 1970s.

Fishing is becoming highly commercialized. The traditional pearl-fishing industry has declined since the introduction of the Japanese cultivated pearls in the 1930s. Large fishing industries have been set up in Kuwait, Qatar, and Bahrain; in 1968, 10,000 tons of prawns were exported by Kuwait.

There has always been a considerable sea trade carried on by local craft between the Persian Gulf and Africa and India; this is now completely dominated by an incessant flow of large tankers that carry oil from the large marine terminals at Khārk Island, Kuwait, Dhahran, Bahrain, Musay'īd, aẓ-Ẓannah, and Dās Island to all parts of the world. The heavy traffic and the offshore oil installations have produced many hazards, despite the use of a system of radio-navigational stations.

Prospects for the future. The Persian Gulf area will probably remain an important source of world oil for a long period. The abundance of cheap power should make it possible to set up local extractive industries using raw materials that are too low grade to be exploited elsewhere: a large aluminum-smelting plant is already being constructed in Bahrain. This abundance of cheap power should also allow the exploitation of the highly saline coastal lagoon waters for chemicals such as magnesium, sodium, and chlorine. Although large underground reserves of freshwater have been found in the area, the general lack of freshwater and the unfavourable climate make agriculture difficult. The vast oil wealth, however, should make it possible to develop solar desalination plants to overcome this problem.

BIBLIOGRAPHY. G. EVANS, "Persian Gulf," in R.W. FAIRBRIDGE (ed.), *Encyclopedia of Oceanography*, pp. 689–695 (1966), a general review of the geology and physiography of the gulf; "The Recent Sedimentary Facies of the Persian Gulf," *Proc. R. Soc.*, Series A, 259:291–298 (1966), a general review of the oceanography and marine geology; N.L. FALCON, "The Geology of the North-East Margin of the Arabian Basement Shield," *Advmt. Sci.*, 24:31–42 (1967), an excellent general review of the geological development of the area, with references; GREAT BRITAIN, HYDROGRAPHIC OFFICE, *Persian Gulf Pilot* (rev. periodically), an invaluable source of meteorological, oceanographical, navigational, and local physiographic detail.

(G.Ev.)

Personality, Measurement of

The study of personality as it is treated here is an effort to identify and to describe distinctive human characteristics and to apply these in predicting the behaviour of people. Descriptions of personality made in ordinary conversation are often highly personal, informal, and of unknown reliability and validity. A statement that someone "has a lot of personality" or "has personality problems" is too vague and imprecise for careful study. Personality descriptions toward which psychologists strive contain objective statements about one's personal attributes that are predictive of his future behaviour and that have theoretical significance in scientific research.

The measurement of personal characteristics is the subject matter of the field of psychology today called personality assessment. Assessment is an end result of gathering information intended to advance psychological theory and research, and to increase the probability that wise decisions will be made in applied settings (*e.g.*, in selecting the most promising people from a group of job applicants). The approach taken by the specialist in personality assessment is based on the assumption that much of the observable variability in behaviour from one person to another results from differences in the extent to which individuals possess particular underlying personal characteristics (traits). The assessment specialist seeks to define these traits, to measure them objectively, and to relate them to socially significant aspects of behaviour.

Personality assessment

A distinctive feature of the scientific approach to personality measurement is the effort, wherever possible, to describe human characteristics in quantitative terms. How *much* of a trait manifests itself in an individual? How *many* traits are present? Quantitative personality measurement is especially useful in comparing groups of people as well as individuals. Do groups of persons from different cultural and economic backgrounds differ when considered in the light of their particular personality attributes or traits? How large are the group differences?

Overt behaviour is a reflection of interactions among a wide range of underlying factors, including the bodily state of the individual and the effects of his past personal experiences on him. Hence, a narrowly focussed approach is inadequate to do justice to the complex human behaviour that occurs under the constantly changing set of challenges, pleasures, demands, and stresses of everyday life. The sophisticated measurement of human personality inescapably depends on the use of a variety of concepts to provide trait definitions and entails the application of a diversity of methods of observation and evaluation. Personality theorists and researchers seek to define and to understand the diversity of human traits, the many ways people have of thinking and perceiving and learning and emoting. Such human dimensions, types, and attributes are constructs—in this case, inferences drawn from observed behaviour. Widely studied personality constructs include: anxiety, hostility, emotionality, motivation, and introversion-extroversion. Anxiety, for example, is a concept, or construct, inferred from such acts of behaviour as what a person says, the expression on his face, and the manner in which he moves his body.

Efforts to measure personality constructs stem from a variety of sources. Frequently they grow out of theories of personality; anxiety and repression (the forgetting of unpleasant experiences), for example, are among the central concepts of the theory of psychoanalysis. It is understandable that efforts would be made to quantify one's degree of anxiety, for example, and to use the score thus obtained in the assessment of and in the prediction of his future behaviour. Among the major issues and sources of controversy that characterize the study of personality measurement is the question of which of the many personality constructs that have been quantified are basic or fundamental and which can be expected to involve wasted effort in their measurement because they represent poorly defined combinations of more elemental constructs; which measurement techniques are most effective and convenient for the purpose of assessment; whether it is better to interview people in measuring personality, or to ask them to say, for example, what an inkblot or a cloud in the sky reminds them of.

Efforts to measure any given personality construct can fail, coming to naught because of many combinations of inadequacies in formulating or defining the trait to be measured and weaknesses in the assessment methods employed. An investigator might desire to specify quantitatively the degree to which individuals are submissive in social and competitive situations. His effectiveness will depend on the particular theory of submissiveness he brings to bear on the problem; on the actual procedures he selects or devises to measure submissiveness; and on the adequacy of the research he performs to demonstrate the usefulness of the measure. Each of these tasks must be considered carefully in evaluating efforts to measure personality attributes.

The methods used in personality description and measurement fall into several categories that differ with regard to the type of information gathered and the methods by which it is obtained. While all should rely on data that come from direct observations of human behaviour if they are to have at least the semblance of scientific value, all may vary with regard to underlying assumptions, validity, and reliability (consistency, in this case).

PERSONALITY EVALUATION WITHOUT TESTS

The interview. In an interview the individual under assessment must be given considerable latitude in "telling his story." Interviews have both verbal and nonverbal (*e.g.*, gestural) components. The aim of the interview is to gather information, and the adequacy of the data gathered depends in large part on the questions asked by the interviewer. In an employment or personnel interview the focus of the interviewer reasonably will be on the job candidate's work experiences, on his general and specific attitudes, and on his occupational goals. In a diagnostic medical or psychiatric interview considerable attention would be paid to the history of the patient's bodily health and to any symptoms of behavioral disorder he may have displayed over the years.

Two broad types of interview may be delineated. In the interview designed for use in research, face-to-face contact between an interviewer and interviewee is directed toward eliciting information that may be relevant to particular practical applications under general study or to those personality theories (or hypotheses) being investigated. Another type, the clinical interview, is focussed on assessing the status of a particular individual (*e.g.*, a psychiatric patient); such an interview is action oriented (*i.e.*, it may indicate appropriate treatment). Both research and clinical interviews frequently may be conducted to obtain an individual's life history and biographical information (*e.g.*, identifying facts, family relationships), but they differ in the uses to which the information is put.

Although it is not feasible to quantify all of the events occurring in an interview, personality researchers have devised ways of categorizing many aspects of the content of what a person has said. In this approach, called content analysis, the particular categories used depend upon the researchers' interests and ingenuity, but the method of content analysis is quite general and involves the construction of a system of categories that hopefully can be used reliably by an analyst or scorer. The categories may be straightforward (*e.g.*, the number of words uttered by the interviewee during designated time periods), or they may rest on inferences (*e.g.*, the degree of personal unhappiness the interviewee appears to express). The value of content analysis is that it provides the possibility of using frequencies of uttered response to describe verbal behaviour and defines behavioral variables for more-or-less precise study in experimental research. Content analysis has been used, for example, to gauge attitudinal changes as they occur within a person with the passage of time. Changes in the frequency of hostile references a neurotic makes toward his parents during a sequence of psychotherapeutic interviews, for example, may be detected and assessed as may the changing self-evaluations of psychiatric hospital inmates in relation to the length of their hospitalization.

Erroneous conclusions

Sources of erroneous conclusions that may be drawn from face-to-face encounters stem from the complexity of the interview situation, the attitudes, fears, and expectations of the interviewee, and the interviewer's manner and training. Research has been conducted to identify, control, and, if possible, eliminate these sources of interview invalidity and unreliability. By conducting more than one interview with the same interviewee and by using more than one interviewer to evaluate his behaviour, light can be shed on the reliability of the information derived and may reveal differences in influence among individual interviewers. Standardization of interview format tends to increase the reliability of the information gathered; for example, all interviewers may use the same set of questions. Such standardization, however, may restrict the scope of information elicited; and even a perfectly reliable (consistent) interview technique can lead to incorrect inferences.

Behavioral observation. In every type of assessment situation behaviour is sampled in some way; *i.e.*, not every bit of the individual's behaviour (past and future, for example) can be observed. The behaviour sampled in the interview may be that of self-description in which the individual indicates how he perceives himself. Other behavioral observations (*e.g.*, of facial expression) are also made in interviews as well as in specially devised test and everyday situations. A situational test permits observation of specific aspects of one's behaviour, which may be predictive of his future behaviour. The specially devised test takes the form of a miniature life situation; during World War II, for example, situational tests were used by a number of governments to select individuals who were judged capable of withstanding the inquisition and grilling to which a captured spy might be subjected. The tests typically took the form of stressful situations in which the individual was required to attain specified goals (*e.g.*, completing an assigned project under trying, frustrating conditions).

Another approach to behavioral observations involves the establishment of quantitative base lines against which to compare the future behaviour of individuals. Does the frequency with which a child seeks out other children, for example, increase with the number of play-therapy sessions in which he participates? Such observations permit quantitative description of someone's response repertory at a given point in time and can provide a basis for measuring therapeutic effectiveness.

Behavioral observation. In every type of assessment ways. One of these is to keep track of the frequency with which people make designated responses during a given period of time (*e.g.*, the number of times a psychiatric patient makes his own bed or the number of times a child asks for help in a novel situation). Another approach involves asking raters to support their judgments of others by citing specific behaviour (critical incidents); a shop foreman, for example, may rate a worker as depressed by citing incidents when the worker burst into tears. Critical incidents not only add validity to ordinary ratings; they also suggest behavioral details that might be promising predictors of success on the job, response to psychiatric treatment, or level of academic achievement.

Behavioral observations are widely made in interviews and in a variety of workaday settings. Employers, supervisors, and teachers—either formally or informally—make use of behavioral observations in making decisions about people for whom they have responsibility. Unfortunately the subject may know he is being studied or evaluated and, therefore, may behave atypically (*e.g.*, by working harder than usual, or by growing tense). The observer may be a source of error by being biassed in favour of or against the subject. Disinterested observers clearly are to be preferred (other things being equal) for research and clinical purposes. The greater the care taken to control for such contributions to error, the greater the likelihood that observations will prove to be reliable.

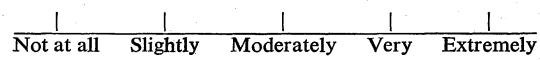
Rating scales. Some observations do not lend themselves to quantification as readily as do simple counts of motor behaviour (such as the number of times a worker leaves his lathe to go to the bathroom). It is difficult, for example, to quantify how charming an office receptionist is. In such cases, one may fall back on relatively subjective judgments, inferences, and relatively imprecise estimates, as in deciding how disrespectful a child is.

One approach to securing such judgments of the behaviour of others is through the use of rating scales, which have much in common with self-report questionnaires described below. The rating scales present an observer with scalar dimensions along which those who are observed are to be placed. A teacher, for example, might be asked to rate his students on the degree to which the behaviour of each reflects leadership capacity, shyness, or creativity. Peers might rate each other along dimensions such as friendliness, trustworthiness, and social skills. Several standardized, printed rating scales are available for describing the behaviour of psychiatric hospital patients. Relatively objective rating scales have also been devised for use with other groups; for example, the Fels Parent Behavior Scales are available for assessing parental behaviour toward their children in the home situation.

Psychiatric ratings

Rating scales are among the most venerable and versatile of the paper-and-pencil methods of personality measurement. They often take a graphic form:

To what degree is John shy?



A number of requirements should be met to maximize the usefulness of rating scales. One is that they be reliable: the ratings of the same person by different observers should be consistent. Other requirements are reduction of sources of inaccuracy in personality measurement; the so-called halo effect results in an observer's rating someone favourably on a specific characteristic because the observer has a generally favourable reaction to the person he is rating. One's tendency to say only nice things about others or his proneness to think of all people as average (to use the midrange of scales) represent other methodological problems that arise when rating scales are used.

Factual information: unobtrusive and physiological measurement. In an interview there is an opportunity for direct observations of expressive behaviour (*e.g.*, speech patterns, poise, gestures) and for gathering the subject's life history and other types of information. While an interviewee usually is aware that he is an object of interest and inquiry, behavioral observations may be recorded whether the subject is or is not aware that his behaviour is under study. If he is aware, his behaviour may be influenced significantly by that fact, and he may strive to create a good impression. There are, however, ways of gathering data that are resistant to the biases this creates.

One type of information that is sometimes overlooked because of its very simplicity consists of the subject's life history and present status.

Much of this information may be gathered through direct interviews with a subject or with an informant, through questionnaires, and through searches of records and archives. The information might also be gathered by examining the subject's personal documents (*e.g.*, letters, autobiographies) and medical, educational, or psychiatric case histories. The information might concern the individual's social and occupational history, his cultural background, his present economic status, and his past and present physical characteristics. Life-history data can provide clues to the precursors and correlates of present behaviour. This information may help the investigator avoid needlessly speculative or complex hypotheses about the causation of personality traits when simple explanations might be superior. Failure on the part of a personality evaluator to be aware of the fact that someone had spent two years during World War II in a concentration camp could result in misleading inferences and conjectures about the subject's present behaviour.

Searches of records and archives

Even if an individual knows he is being studied or evaluated, he may not know which of his personality attributes

are of special interest to the assessor. One often unwittingly or uncontrollably reveals feelings and thoughts that he desires to keep private. Experienced personality assessors (*e.g.*, psychiatric clinicians) come to be adept at observing expressive nuances (subtleties in behaviour) that an individual either believes are being successfully suppressed or that he is unaware of himself. Slight variations in gestures, intonations, and mannerisms illustrate these expressive aspects of behaviour.

Physiological measurement

Along with expressive behaviour, measures of bodily functioning may reveal data that are related to the motivation and behaviour of the individual. Aspects of physique, the levels of hormones or sugar in the blood, tests of neurological functioning (*e.g.*, the knee jerk), and characteristics such as alterations in the resistance of the skin to electric current (the galvanic skin reaction) and changes in the dilation of the pupils of the eyes are among the bodily measures that have been used in personality assessment. With technological advances, it is becoming increasingly feasible and convenient to gauge unobtrusively the internal physical state of the individual. The activity of the sweat glands, the heartbeat, one's total bodily blood volume and blood pressure, and the amount of different substances in his bloodstream can be recorded and used to predict his social behaviour, the intensity of his drives (motivation), and the presence or absence of specified emotionally significant states such as frustration and stress.

One of the challenges confronting assessment psychologists is the fact that personal characteristics are not only not constant but may vary even from moment to moment with changes in the conditions of life. Electronic and mechanical devices so adapted as continuously to monitor changes in bodily functioning that are related to periodic ratings of mood and to other physiological states provide measures of stability and change within the individual.

PERSONALITY TESTS

The aim of personality assessment is to secure accurate and useful (valid) statements about persons. While there may be general agreement about the theoretical and practical value of characterizing individuals, it is not immediately obvious how this can be accomplished; disagreements about methodological matters in personality assessment thus have been widespread. The success that has attended the use of convenient intelligence tests in providing reliable, quantitative (numerical) indices of individual ability has stimulated interest in the possibility of devising similar tests for measuring personality. Procedures now available vary in the degree to which they achieve scorability and convenience. These desirable attributes can be partly achieved by restricting in designated ways the kinds of responses a subject is free to make. A test that restricts the subject to true–false answers, for example, is likely to be convenient to give and easy to score. So-called personality inventories tend to have these characteristics, in that they are relatively restrictive, objectively scorable, and convenient to administer. Other techniques (such as inkblot tests) for evaluating personality possess these characteristics to a lesser degree.

Personality inventories. In the early history of personality measurement, most tests were constructed on the basis of so-called face validity; that is, they simply appeared to be valid. Items were included simply because, in the fallible judgment of the person who constructed or devised the test, they were indicative of certain personality attributes. In other words, face validity need not be defined by careful, quantitative study; rather, it typically reflects one's more-or-less imprecise, possibly erroneous, impressions. Personal judgment, even the personal judgment of an expert, is no guarantee that a particular collection of test items will prove to be reliable and meaningful in actual practice.

Self-report inventories

A widely used early self-report inventory, the so-called Woodworth Personal Data Sheet, was developed during World War I to detect soldiers who were emotionally unsuited for combat. Among its ostensibly face-valid items were these: Does the sight of blood make you sick or dizzy? Are you happy most of the time? Do you sometimes wish you had never been born? Recruits who answered these kinds of questions in a way that could be taken to mean that they suffered psychiatric disturbance were detained for further questioning and evaluation. Clearly, however, symptoms revealed by such answers are exhibited by many people who are relatively free of emotional disorder.

Present-day self-report measures that take the form of personality inventories consist of many and varied statements or questions to which the subject is asked to respond in a particular way. His task might be to decide whether or not each of a series of statements is characteristic of his own feelings or behaviour; or he might respond true or false to a series of questions about the world in which he lives. Several inventories require that each of a series of statements be placed on a rating scale in terms of the frequency or adequacy with which the statements are judged by the individual to reflect his tendencies and attitudes. Regardless of the way in which the subject responds, most inventories yield several scores each intended to identify a distinctive aspect of personality.

One of these, the *Minnesota Multiphasic Personality Inventory* (MMPI), is probably the personality inventory in widest use in the English-speaking world at the present time. Also available in other languages, it consists in one version of 550 items (*e.g.*, "I like tall women.") to which subjects are to respond "true," "false," or "cannot say." Work on this inventory began in the 1930s, when its construction was motivated by the need for a practical, economical means of describing and predicting the behaviour of psychiatric patients. In its development efforts were made to achieve convenience in administration and scoring and to overcome many of the known defects of earlier personality inventories. Varied types of items were included and emphasis was placed on making these printed statements (presented either on small cards or in a booklet) intelligible even to persons with limited reading ability.

Most earlier inventories lacked subtlety; many people were able to fake or bias their answers since the items presented were easily seen to reflect gross disturbances; indeed, in many of these inventories maladaptive tendencies would be reflected in either all true or all false answers. Perhaps the most significant methodological advance to be found in the MMPI was the attempt on the part of its developers to measure tendencies to respond, rather than actual behaviour, and to rely but little on assumptions of face validity. The true–false item "I hear strange voices all the time" has face validity for most people in that to answer "true" to it seems to provide a strong indication of abnormal hallucinatory experiences. But some psychiatric patients who "hear strange voices" can still appreciate the socially undesirable implications of a "true" answer and may therefore actually try to conceal their abnormality by answering "false." A major difficulty in placing great reliance on face validity in test construction is that the subject may be as aware of the significance of his responses as is the test constructor and thus may be able to put his "best foot" forward and mislead the tester. Nevertheless, the person who hears strange voices and yet answers the item "false" clearly is responding to something—his answer still is a reflection of his personality, even though it may not be the aspect of personality to which the item seems to refer; thus, careful study of responses beyond their mere face validity often proves to be profitable.

Reactions to the testing environment

Much study has been given to the ways in which response sets and test-taking attitudes influence behaviour on the MMPI and other personality measures. The response set called acquiescence, for example, refers to one's tendency to respond with "true" or "yes" answers to questionnaire items regardless of what the item content is. It is conceivable that two people might be quite similar in all respects except for their tendency toward acquiescence. This difference in response set can lead to misleadingly different scores on personality tests. One person might be a "yea-sayer" (someone who tends to answer true to test items); another might be a "nay-

sayer"; a third individual might not have a pronounced acquiescence tendency in either direction.

Acquiescence is not the only response set; there are other test-taking attitudes that are capable of influencing personality profiles. One of these, already suggested by the example of the person who hears strange voices, is social desirability. A person who has convulsions might say "false" to the item "I have convulsions" because he believes that others will think less of him if they know he has convulsions. The intrusive potentially deceiving effects of the subjects' response sets and test-taking attitudes on scores derived from personality measures can sometimes be circumvented by varying the content and wording of test items. Nevertheless, users of questionnaires have not yet completely solved problems of bias such as those arising from response sets. Indeed, many of these problems first received widespread attention in research on the MMPI and research on this and similar inventories has significantly advanced understanding of the whole discipline of personality testing.

The MMPI as originally published consists of nine clinical scales (or sets of items), each scale having been found in practice to discriminate a particular clinical group, such as people suffering from schizophrenia, depression, or paranoia (see PSYCHOSES). Each of these scales (or others produced later) was developed by determining patterns of response to the inventory that were observed to be distinctive of groups of individuals who had been psychiatrically classified by other means (*e.g.*, by long-term observation). The responses of apparently normal subjects were compared with those of hospital patients with a particular psychiatric diagnosis; for example, with symptoms of schizophrenia. Items to which the greatest percentage of "normals" gave answers that differed from those more typically given by patients came to constitute each clinical scale.

In addition to the nine clinical scales and many specially developed scales, there are four so-called control scales on the inventory. One of these is simply the number of items placed by the subject in the "cannot say" category. The L (or lie) scale was devised to measure the tendency of the test taker to attribute socially desirable attributes to himself. In response to "I never tell a lie," such a person is apt to answer true; for "I get angry sometimes" he should tend to mark false; extreme L scorers in the other direction appear to be too good, too virtuous. Another so-called F scale was included to provide a reflection of the subjects' carelessness and confusion in taking the inventory (*e.g.*, "Everything tastes the same" tends to be answered true by careless or confused people). More subtle than either the L or F scales is what is called the K scale. Its construction was based on the observation that some persons tend to exaggerate their symptoms because of excessive openness and frankness and may obtain high scores on the clinical scales; others may exhibit unusually low scores because of defensiveness. On the K-scale item "I think nearly anyone would tell a lie to keep out of trouble," the defensive person is apt to answer false, giving the same response to "I certainly feel useless at times." The K scale was designed to reduce these biassing factors; by weighting clinical-scale scores with K scores, the distorting effect of test-taking defensiveness may be reduced.

In general, it has been found that the greater the number and magnitude of one's unusually high scores on the MMPI, the more likely it is that he is in need of psychiatric attention. Most professionals who use the device refuse to make assumptions about the factualness of the subject's answers and about his personal interpretations of the meanings of the items. Their approach does not depend heavily on theoretical predilections and hypotheses. For this reason the inventory has proved particularly popular with those who have strong doubts about the eventual validity that many theoretical formulations will show in connection with personality measurement after they have been tested through painstaking research. The MMPI also appeals to those who demand firm experimental evidence that any personality assessment method can make valid discriminations among individuals.

In recent years there has been growing interest in actuarial personality description; that is, in personality description based on traits shared in common by groups of people. Actuarial description studies yield rules by which persons may be classified according to their personal attributes as revealed by their behaviour (on tests, for example). Computer programs are now available for diagnosing such disorders as hysteria, schizophrenia, and paranoia on the basis of typical group profiles of MMPI responses. Computerized methods for integrating large amounts of personal data are not limited to this inventory and are applicable to other inventories, personality tests (*e.g.*, inkblots), and life-history information. Computerized classification of MMPI profiles, however, has been explored most intensively.

Diagnosis with computers

The MMPI has been considered in some detail here because of its wide usage and because it illustrates a number of important problems confronting those who attempt to assess personality characteristics. There are many other personality inventories and paper-and-pencil assessment techniques that are used in applied settings and in research. The *California Psychological Inventory* (CPI), for example, is keyed for several personality variables that include sociability, self-control, flexibility, and tolerance. Unlike the MMPI, it was developed specifically for use with "normal" groups of people. Whereas the judgments of experts (usually psychiatric workers) were used in categorizing subjects given the MMPI during the early item-writing phase of its development, nominations by peers (one's equals or associates) were relied upon in work with the *California Psychological Inventory*. Its technical development has been evaluated by test authorities to be of high order, in part because its developers profited from lessons learned in the construction and use of the MMPI. It also provides measures of response sets and has been subjected to considerable research study.

Another instrument, the *Eysenck Personality Inventory*, was created in England and provides measures of neuroticism and extroversion. Its use has been closely linked with English psychologist Hans J. Eysenck's theories of personality. The device is short and easily administered, the neuroticism factor being designed to measure emotional instability and neurotic predispositions to break down under stress (see PSYCHONEUROSES). The extroversion items on this inventory are held to sample one's tendency to exhibit sociable, out-going behaviour.

The *Edwards' Personality Preference Schedule* is a so-called forced-choice personality inventory. For each item the subject is required to characterize himself by choosing between paired statements of equal apparent desirability or undesirability. Such a task can be distasteful to many test takers; it may require, for example, being forced to describe oneself as either "stupid" or "ugly." Among others, the *Edwards' Personality Preference Schedule* yields measures of one's needs for achievement, heterosexuality, and social dominance. The strength of each need is not expressed in absolute terms but in relation to the intensity of the individual's other needs.

Beyond personality inventories, there are other self-report approaches to personality measurement available for research and applied purposes. Mention was made earlier of the use of rating scales filled out by observers. An individual also can use rating scales as a vehicle for his own self-description. The so-called semantic differential is a rating-scale technique that permits quantification of an individual's reactions to himself, to others, and, in fact, to any object or concept in terms of a standard set of semantic (word) polarities such as "hot–cold" or "good–bad." It is a general method for assessing the meanings of these semantic concepts to individuals.

Self-description techniques

Another method called the Q-sort is devised for problems similar to those for which rating scales are used. In a Q-sort a person is given a set of sentences, phrases, or words (usually presented individually on cards) and is asked to use them to describe himself (as he thinks he is or as he would like to be) or someone else. This description is carried out by having the subject sort the items on the cards in terms of their degree of relevance so that they can be distributed along what amounts to a rating scale.

Examples of descriptive items that might be included in a Q-sort are "worries a lot," "works hard," and "is cheerful."

Typical paper-and-pencil instruments such as personality inventories involve verbal stimuli (words) intended to call forth designated types of responses from the individual. There are clearly stated ground rules under which he makes his responses. Paper-and-pencil devices are relatively easy and economical to administer and can be scored accurately and reliably by relatively inexperienced clerical workers. They are generally regarded by professional personality evaluators as especially valuable assessment tools in screening large numbers of people, as in military or industrial personnel selection. Assessment specialists do not assume that self-reports are accurate indicators of personality traits. They are accepted, rather, as samples of behaviour for which validity in predicting one's everyday activities or traits must be established empirically (*i.e.*, by direct observation or experiment). Paper-and-pencil techniques have moved from their early stage of assumed (face) validity to more advanced notions in which improvements in conceptualization and methodology are clearly recognized as basic to the determination of empirical validity.

Projective techniques. One group of assessment specialists holds that the most meaningful path to personality description and classification requires that the person who is assessed be given the opportunity to respond as freely as possible so as to reflect his unique (idiosyncratic) ways of thinking and behaving. Inkblots and other ambiguous stimuli comprise a number of tests that have been used widely in clinical situations (*e.g.*, in psychiatric hospitals) in this type of assessment. Such ambiguous stimuli are used to provide so-called projective techniques (since they allow one relative freedom to project his own interests and feelings into them, reacting in any way that seems appropriate).

Personality inventories and projective techniques do have some elements in common; inkblots, for example, are ambiguous, but so also are many of the statements on inventories such as the MMPI. These techniques differ in that the subject is given substantially free rein in responding to projective stimuli rather than merely answering true or false, for example. Another similarity between projective and questionnaire or inventory approaches is that all involve the use of relatively standardized testing situations.

While projective techniques are often lumped together as one general methodology, in actual practice there are several approaches to assessment from a projective point of view. Although projective techniques share the common characteristic that they permit the subject wide latitude in responding, they still may be distinguished broadly as follows: (1) *associative techniques*, in which the subject is asked to react to words, to inkblots, or to other stimuli with the first associated thoughts that come to mind; (2) *construction techniques*, in which the subject is asked to create something—for example, make up a story or draw a picture of himself; (3) *completion techniques*, in which the subject is asked to finish a partially developed stimulus, such as adding his own last words to an incomplete sentence; (4) *choice or ordering techniques*, in which the subject is asked to choose from among or to give some orderly sequence to stimuli—for example, to choose from or arrange a set of pictures or inkblots; (5) *expressive techniques*, in which the subject is asked to express himself freely in some manner, as in finger painting.

Types of projective techniques

These techniques allow the subject wide latitude in making his responses. Because of this freedom, projective techniques are held by some theorists to be sensitive to unconscious dimensions of personality. Hidden personality defense mechanisms, latent emotional impulses, and inner anxieties all have been attributed to test takers by making theoretical inferences from data gathered as they responded in projective situations. While projective stimuli are ambiguous, they are usually administered under fairly standardized conditions. Quantitative (numerical) measures can be derived from subjects' responses to them. These include the number of responses one makes to a series of inkblots and the number of responses to the blots in which the subject perceives what seem to him to be moving nonhuman animals.

The Rorschach Inkblot Test. The Rorschach inkblots were developed by a Swiss psychiatrist, Hermann Rorschach, in an effort to reduce the time required in psychiatric diagnosis. His test consists of ten cards, half of which are in colour and half in black and white. The test is administered by showing the subject the ten blots one at a time; the subject's task is to describe what he sees in the blots or what they remind him of. The subject is usually told that the inkblots are not a test of the kind he took in school and that there are no right or wrong answers.

Rorschach's work was stimulated by his interest in the relationship between perception and personality. He held that a person's perceptual responses to inkblots could serve as clues to basic personality tendencies. Despite Rorschach's original claims for the validity of his test, subsequent negative research findings have led many users of projective techniques to become dubious about the role assigned the inkblots in delineating relationships between perception and personality. In recent years, emphasis has tended to shift to the analysis of nuances of the subject's social behaviour during the test and to the content of his verbal responses to the examiner; whether, for example, he seeks to obtain the assistance of the examiner in "solving" the inkblots presented to him, sees "angry lions" or "meek lambs" in the inkblots, is apologetic or combative about his responses.

Evaluation of Rorschach validity

Over the years, considerable research has been carried out on Rorschach's inkblots; important statistical problems in analyzing data gathered with projective techniques have been identified, and researchers have continued in their largely unsuccessful efforts to overcome them. There is a vast experimental literature to suggest that the Rorschach technique lacks empirical validity.

A similar method, the *Holtzman Inkblot Test*, has been developed in an effort to eliminate some of the statistical problems that beset the Rorschach test. It involves the administration of a series of 45 inkblots, the subject being permitted to make only one response per card. The Holtzman has the desirable feature that it provides an alternate series of 45 additional cards for use in retesting the same person.

Research with the Rorschach and Holtzman has proceeded in a number of directions; many studies have compared psychiatric patients and other groups of special interest (delinquents, underachieving students) with ostensibly normal people. Some investigators have sought to derive indices or predictions of future behaviour from responses to inkblots, and have checked, for example, to see if anxiety and hostility (as inferred from content analyses of verbal responses) are related to favourable or unfavourable response to psychotherapy. A sizable area of exploration concerns the effects of special conditions (*e.g.*, experimentally induced anxiety or hostility) on the inkblot perceptions reported by the subject and the content of his speech.

Thematic Apperception Test (TAT). There are other personality assessment devices, which, like the Rorschach, are based on the idea that an individual will project something of himself into his description of an ambiguous stimulus.

The *Thematic Apperception Test*, for example, presents the subject with pictures of persons engaged in a variety of activities (*e.g.*, someone with a violin). While the pictures leave much to one's imagination, they are more highly specific, organized visual stimuli than are inkblots. The test consists of 30 black and white pictures and one blank card (to test imagination under very limited stimulation). The cards are presented to the subject one at a time and he is asked to make up a story that describes each picture and that indicates the events that led to the scene and the events that will grow out of it. He is also asked to describe the thoughts and feelings of the persons in his story.

Although some content-analysis scoring systems have been developed for the TAT, attempts to score it in a stan-

dardized quantitative fashion tend to be limited to research and have been fewer than has been the case for the Rorschach. This is especially the state of affairs in applied settings in which the test is often used as a basis for conducting a kind of clinical interview; the pictures are used to elicit a sample of verbal behaviour on the basis of which inferences are drawn by the clinician.

In one popular approach, interpretation of a TAT story usually begins with an effort to determine who is the hero (*i.e.*, to identify the character with whom the subject seems to have identified himself). The content of the stories is often analyzed in terms of a so-called need-press system. Needs are defined as the internal motivations of the hero. Press refers to environmental forces that may facilitate or interfere with the satisfaction of needs (*e.g.*, in the story the hero may be physically attacked, frustrated by poverty, or suffer the effects of rumours being spread about him). In assessing the importance or strength of a particular inferred need or press for the individual who takes the test, special attention is given to signs of its pervasiveness and consistency in different stories. Analysis of the test may depend considerably on the subjective, personal characteristics of the evaluator, who usually seeks to interpret the subjects' behaviour in the testing situation; the characteristics of his utterances; the emotional tone of the stories; the kinds of fantasies he offers; the outcomes of the stories; and the conscious and unconscious needs speculatively inferred from the stories.

Word-association techniques. The list of projective approaches to personality assessment is long, one of the most venerable being the so-called word-association test. The Swiss psychiatrist Carl Jung used associations to groups of related words as a basis for inferring personality traits (*e.g.*, the inferiority "complex"). Administering a word-association test is relatively uncomplicated; a list of words is presented one at a time to the subject who is asked to respond with the first word or idea that comes to mind. Many of the stimulus words may appear to be emotionally neutral (*e.g.*, building, first, tree); of special interest are words that tend to elicit personalized reactions (*e.g.*, mother, hit, love). The amount of time the subject takes before he begins each response and the response itself are used in efforts to analyze a word association test. The idiosyncratic, or unusual, nature of one's word-association responses may be gauged by comparing them to standard published tables of the specific associations given by large groups of other people.

Word stimulus responses

Sentence-completion techniques. The sentence-completion technique is a logical extension of word-association methods. In administering a sentence-completion test, the subject is presented with a series of partial sentences that he is asked to finish in his own words (*e.g.*, "I feel upset when . . . "; "What burns me up is . . . "). Users of sentence-completion methods in assessing personality typically analyze them in terms of what they judge to be recurring attitudes, conflicts, and motives reflected in them.

ASSESSING ASSESSMENT TECHNIQUES

Assessment, whether it is carried out with interviews, behavioral observations, physiological measures, or tests, is intended to permit the evaluator to make meaningful, valid, and reliable statements about individuals. What makes John Doe tick? What makes Mary Doe the unique individual that she is? Whether these questions can be answered depends upon the reliability and validity of the assessment methods used. The fact that a test is *intended* to measure a particular attribute is in no way a guarantee that it really *accomplishes* this goal. Assessment techniques must themselves be assessed.

An objective of research in personality measurement is to delineate the conditions under which the methods do or do not make trustworthy descriptive and predictive contributions. One approach to this problem is to compare groups of people known through careful observation to differ in a particular way. It is helpful to consider, for example, whether the MMPI or TAT discriminates significantly between those who show progress in psychotherapy and those who do not, whether they distinguish between law violators of record and apparent nonviolators. Experimental investigations in which the conditions under which subjects perform are varied systematically also make contributions.

Another area of research on assessment has to do with the role of the assessor himself as an evaluator and predictor of the behaviour of others. In most applied settings he subjectively (even intuitively) weighs, evaluates, and interprets the various assessment data that are available. How successful he is in carrying out his interpretive task is critical as is knowledge of the kinds of conditions under which he is effective in processing such diverse data as impressions gathered in an interview, test scores, and life-history data. The typical clinician usually does not use a statistical formula that weighs and combines test scores and other data at his disposal. Rather, he integrates the data using impressions and hunches based on his past clinical experience and on his understanding of psychological theory and research. The result of this interpretive process usually includes some form of personality description of the person under study and specific predictions or advice for that person; *e.g.*, perhaps a recommendation that the person consult a brain surgeon or a physician who treats syphilis.

Evaluating the evaluator

The degree of success an assessor has when he responds to the diverse information that may be available about a particular person is the subject of research that has been carried out on the issue of clinical versus statistical prediction. It is reasonable to ask whether a clinician will do as good a job in predicting behaviour as does a statistical formula or "cookbook"; *i.e.*, a manual that provides the empirical, statistically predictive aspects of test responses or scores based on the study of large numbers of people. An example would be a book or table of intelligence test norms (typical scores) used to predict how well children perform in school. Another book might offer specific personality diagnoses (*e.g.*, neurotic or psychotic) based on scores such as those yielded by the different scales of the MMPI. Many issues must be settled before the deceptively simple question of clinical versus statistical prediction can be answered definitively.

When statistical prediction formulas (well-supported by research) are available, for combining clinical information, however, experimental evidence clearly indicates that they will be more valid and less time-consuming than will a clinician (who may be subject to human error in trying to simultaneously consider and weigh all of the factors in a given case). The clinician's chief contributions to diagnosis and prediction are in situations for which satisfactory formulas and quantified information (*e.g.*, test scores) are *not* available. Clinicians are especially to be thought of when what are being evaluated are rare and idiosyncratic personality characteristics that have escaped rigorous, systematic empirical study. The greatest confidence obtains when both statistical and subjective clinical methods simultaneously converge (agree) in the solution of specific clinical problems; unfortunately, they often disagree.

Although much progress has been made in efforts to measure personality, all available instruments and methods have defects and limitations that must be borne in mind when using them; responses to tests or interview questions, for example, often are easily controlled or manipulated by the subject and thus are readily "fakeable." Some tests, while useful as group screening devices, empirically exhibit only limited predictive value in individual cases, yielding frequent (sometimes tragic) errors. These caveats are especially poignant when significant decisions about a person are made on the basis of his personality measures. Psychiatric institutionalization or discharge, and hiring or firing, are weighty personal matters indeed and can wreak great injustice when based on faulty assessment. In addition, many personality assessment techniques require the probing of highly private areas of the individual's thought and action. Those who seek to measure personality for descriptive and predictive reasons must concern themselves deeply with the ethical and legal implications of their work.

A major methodological stumbling block in the way of

establishing the validity of any method of personality measurement is that there always is an element of subjective judgment in selecting or formulating criteria against which measures may be validated. This is not so serious a problem when popular, socially valued, fairly obvious criteria are available that permit ready comparisons between such groups as convicted criminals and ostensible noncriminals, or psychiatric hospital patients and noninstitutionalized individuals. Many personality characteristics, however, cannot be validated in such directly observable ways (*e.g.*, one's inner, private experiences such as anxiety or depression). When such straightforward empirical validation of an untested measure hopefully designed to measure any personality attribute is not possible, efforts at establishing a less impressive kind of validity (so-called construct validity) may be pursued. A construct is a theoretical statement concerning some underlying, unobservable aspect of an individual's characteristics or of his internal state. ("Intelligence," for example, is a construct; one cannot hold "it" in his hand, or weigh "it," or put "it" in a bag, or even look at "it.") Constructs thus refer to private events inferred or imagined to contribute to the shaping of specific public events (observed behaviour). The explanatory value of any construct has been considered by some theorists to represent its validity. Construct validity, therefore, refers to evidence that endorses the usefulness of a theoretical conception of personality. A test designed to measure an unobservable construct (such as "intelligence" or "need to achieve") is said to accrue construct validity if it usefully predicts the kinds of empirical criteria one would expect it to; *e.g.*, achievement in academic subjects.

The degree to which a measure of personality is empirically related to or predictive of any aspect of behaviour observed independently of that measure contributes to its validity in general. A most desirable step in establishing the usefulness of a measure is called cross-validation. The mere fact that one research study yields positive evidence of validity is no guarantee that the measure will work as well the next time; indeed, often it does not. It is thus important to conduct additional, cross-validation studies to establish the stability of the results obtained in the first investigation. Failure to cross-validate is viewed by most testing authorities as a serious omission in the validation process. Evidence for the validity of a full test should not be sought from the same sample of people that was used for the initial selection of individual test items. Clearly this will tend to exaggerate the effect of traits that are unique to that particular sample of people and can lead to spuriously high (unrealistic) estimates of validity that will not be borne out when other people are studied. Cross-validation studies permit assessment of the amount of "shrinkage" in empirical effectiveness when a new sample of subjects is employed. When evidence of validity holds up under cross-validation, confidence in the general usefulness of test norms and research findings is enhanced. Establishment of reliability, validity, and cross-validation are major steps in determining the usefulness of any psychological test (including personality measures).

Progress in personality measurement depends upon answers to both methodological and theoretical questions. Scholarly interest in theories of personality (see PERSONALITY, THEORIES OF), pressing practical problems (such as the need to screen or diagnose patients efficiently), and the intriguing results of incomplete research provide the impetus for those who seek to devise objective, meaningful measures of personal characteristics. In devising these measures a number of frustrating, but crucial, questions remain to be answered. The pursuit of empirical answers to such questions has been partially rewarding and provides hope that progress will continue to be made in the relatively youthful field of personality measurement.

BIBLIOGRAPHY. C.S. HALL and G. LINDZEY, *Theories of Personality*, 2nd ed. (1970), surveys the major theories of personality. E.F. BORGATTA and W.W. LAMBERT (eds.), *Handbook of Personality Theory and Research* (1968), integrates several of these theories and the research methods that characterize the discipline. Examples and reviews of personality research are contained in I.G. SARASON, *Personality: An Objective Approach* (1966), and (ed.), *Contemporary Research in Personality*, 2nd ed. (1969). L.J. CRONBACH, *Essentials of Psychological Testing*, 3rd ed. (1970), surveys general approaches of assessment; while O.K. BUROS (ed.), *Personality: Tests and Reviews* (1970), provides a valuable compendium of personality tests and measurement devices. D.N. JACKSON and S. MESSICK (ed.), *Problems in Human Assessment* (1970), deals with selected aspects of personality assessment.

(I.G.S.)

Personality, Theories of

Theories of human personality presumably have existed in some form ever since men began to speculate about why people act as they do. The term personality has been given many meanings, but at the core of all of them the reference is to the organization of traits or characteristics within the individual. There are two emphases in this identification of the concept of personality: the individual shares his characteristics (*e.g.*, happiness or ugliness) in common with other people; and each individual is distinguished from all others on the basis of his unique combination of these characteristics and of the degree to which he manifests them.

To suggest that there is a single widely accepted definition of the term personality, however, is misleading. A definition commonly encountered in many textbooks on the subject is comprehensive, defining personality as the sum total of the characteristics of the individual. In contrast, an alternative definition is implied in the everyday practice of suggesting that a person "has a lot of personality," in which usage the reference is to positive, desirable characteristics that give a person social competence and attractiveness. The latter popular definition, which is evaluative, has received relatively little use in the scholarly consideration of theories of personality (see also PERSONALITY, MEASUREMENT OF).

The fullest range of ideas about personality pervades the literary, legal, religious, and philosophical literature about man, just as there are myriad assumptions about what constitutes the nature of man. The notion that whole peoples (*e.g.*, ethnic or racial groups) differ in basic personality characteristics is common; the resulting study of culture and personality in modern times has received attention especially among anthropologists. Fundamental definitions and principles underlying this type of scientific pursuit, however, were clearly enunciated in the writings of the German psychologist Wilhelm Wundt in the mid-19th century. More journalistic forms of ethnic-group personality description are found in earlier writings and essays dating as far back as known history; the ancient Romans, for example, recorded a great variety of personality traits held to be characteristic of the barbarian peoples who lived in the borderlands of the military empire of the Caesars. In more recent journalistic and fictional literature, there are still to be found assertions such as that the French (as an entire people) are volatile, that the Italians are hot-blooded, that the Scandinavians are phlegmatic, and so forth; such statements bear implications that alleged group differences in personality are hereditary rather than being based on culture. Indeed, in the discussion of scientific efforts to establish personality types (which follows below), it should be noted that some human types are defined in terms of heritable bodily characteristics. Since such bodily traits are found to be associated with geographical location, it may be that assertions of ethnic or national tendencies to manifest specific personality traits are not necessarily farfetched.

Early studies of personality traits

Assumptions about the nature of man that are commonly found as aspects of religious, moral, and philosophical writings often assume a model of man based on selected qualities; these may be drawn from myths of human origins or from the characteristics attributed to local gods. Thus, in some cultures, humans may be said to have the characteristics of "mother earth," from whom they are believed to spring. Or humans may be portrayed as having both good and evil in them, corresponding to good spirits and bad spirits held to exist in some religions.

Often in the religions of technologically advanced peoples, these characteristics are outlined in great detail in sacred writings, and inferences about the nature of men may be drawn from examples and incidents recorded in these writings. Thus, the personality model of a saviour (*e.g.*, the Messiah) may imply both the good example and the bad contrast, with emphasis being placed on the socially important aspects of behaviour. Distinctions among unobservable internal "entities" (*e.g.*, "mind," "spirit," "conscience," "true self") may be contrasted with manifest (observable) behaviour, and there may be considerable controversy about problems of moral and behavioral conflict.

Historically, the precursors of the 19th-century study of personality theory were often concerned with the ultimate or essential nature of man. Drawing from concepts outlined by the ancient Greeks, and from other sources, philosophers of the 1600s and 1700s speculated about human personality and went into theoretical detail at times about human perception and personality traits or dispositions. In the writings of the 18th-century economist Adam Smith, philosophical notions of human personality are postulated with great specificity. Smith's theory of man as an economic creature ("economic man") required him to assert the primacy of particular characteristics of personality (*e.g.*, greed) for his analysis to have any basis.

In the paragraphs that follow, some emphasis is placed on formulations of human personality that have been advanced as theories (that is, as more or less speculative guesses). This does not mean, however, that concern with the objective, nontheoretical aspects of personality is not pervasive in society. The behaviour of every person involves some guessing about associates and family and acting according to these hypotheses. Legal rules, drawn to regulate social behaviour, imply definitions and expectations about "normal" personality, about what may happen in an undesirable ("abnormal") way, and are designed to limit such classes of unwanted behaviour. The nature of man thus is the urgent concern of lawmakers, and in relatively sophisticated legal systems there is considerable concern with defining the personality of a legally accountable, responsible person.

TYPE THEORIES OF PERSONALITIES

Classical biochemical theories. Theories of personality invariably involve some notions of types (or classes) of people, if only because particular aspects of the person (*e.g.*, his appearance or his mood) are stressed in describing his unique structure and function. Any rigid classification of people into specific personality types might seem to suggest that a given person must fall into one class only, and that he is distinctly, qualitatively unlike others who exhibit different personality types. Such a rigid point of view is an oversimplification of the observable evidence. Though the degree of specificity with which personality types are advanced may vary considerably, the convenience that derives from thinking of persons as being of one distinct personality type or another has been appealing to many theorists. Since there are so many descriptive terms that relate to personality in any given language, the potential variety of type theories is enormous. A classical theory of human types, offered by ancient Greek physicians, focussed on the emotional (temperamental) attributes of human personality and was based on a relatively primitive understanding of human biology (specifically, bodily physiology), which prevailed at that time. According to these early theorists, emotional equilibrium (indeed, general health) depended on an appropriate balance among four fluids (humours) within the body. It was held that physiological imbalance (produced by an excess of one of the humours, for example) would be reflected in bodily illness and in exaggerated personality traits. Thus, if a person had an excess of blood (one of the four humours), he was expected to have a sanguine temperament; that is, to be optimistic, enthusiastic, and excitable. The modern term hot-blooded may be a survival of this Greek theory of human personality; the notion was so influential that for many centuries physicians throughout the Western world continued the practice of bleeding people who suffered from medical and psychiatric disorders. Such people are still said to be in "bad humour."

The four humours

Too much of a humour called black bile (congealed blood from the spleen) was believed to produce a melancholic temperament. The term melancholia literally means black bile, and there are literary allusions to venting one's spleen. When someone was oversupplied with yellow bile (the yellow-green gall secreted by the liver and stored in the gall bladder), he was held to become choleric; that is, to be angry, irritable, and to view his world with a jaundiced eye. Jaundice remains in modern medical language as a disease of the liver or gall bladder in which bile is present in the body to such a degree that the eyeballs and body may turn yellow. Finally, with an abundance of the humour called phlegm (as secreted in the throat), people were supposed to become stolid, apathetic, and undemonstrative; that is, to grow phlegmatic.

As biological science has progressed, these primitive concepts of body chemistry have been replaced by more subtle and complex biological theories of personality. Thus, the chemical factors associated with given psychological dispositions are now more likely to be understood in terms of hormones (as from the thyroid gland), nerve impulses, and so-called psychotropic drugs such as tranquillizers.

Morphological theories. Somewhat less subtle than the biochemical theories of personality are those that distinguish different classes of people through relatively gross characteristics of bodily build (shape or morphology). A morphological theory developed by a German psychiatrist, Ernst Kretschmer (1888–1964), is possibly the most widely known. Kretschmer distinguished among his patients three types of bodily structure (somatotypes). What he called the pyknic type was characterized by a thick trunk, short limbs, and a tendency to put on weight, especially with increasing age. According to his observations in the psychiatric hospital where he worked, sufferers of major fluctuations in mood (so-called manic-depressive psychotics, alternatively called extreme cyclothymic personalities) were more likely to show the pyknic somatotype than were other inmates. By contrast, Kretschmer delineated an asthenic (leptosomic) body type, characteristically displaying a slim trunk, long legs, and a tendency to be tall and lean. These leptosomes tended to appear temperamentally withdrawn or introverted and seemed more prone than other people to develop symptoms of a group of psychiatric disorders now labelled schizophrenia (see PSYCHOSES). The so-called athletic somatotype, with broad shoulders and slim hips, while no less susceptible to psychiatric disturbance, apparently favoured neither schizophrenia nor the manic-depressive psychosis.

Kretschmer's body types

Despite early hopes that somatotypes might be useful in diagnosing personality characteristics or in identifying psychiatric difficulties, the relationships observed by Kretschmer were not found to be sufficiently clear-cut for practical application. Subsequent and much more elaborate studies reported in the 1940s in the United States also yielded three somatotypes: endomorphy (roughly pyknic with similar temperament called viscerotonia); ectomorphy (crudely equivalent to Kretschmer's asthenics, and temperamentally characterized as cerebrotonic); and mesomorphy (comparable to the athletic body type, and called somatotonic in temperament).

This system did not serve practical needs any more effectively than had Kretschmer's, even though it had been based on much more painstaking work. In this context, other physical characteristics, such as head shape (physiognomy) and bumps on the skull (phrenology), have been found so wanting as bases for personality theories that they have not received much serious attention in the current century. The all-encompassing, exaggerated claims made for these theories of physical morphology first drew enthusiasm and finally generated the reaction of disappointment. Nevertheless, structural differences in the body do seem to have some significance in pointing to such aspects of personality as intelligence and emotion.

In cases of a form of mental deficiency called mongolism, for example, distinctive "Asian" facial features and squat bodily structure are readily recognizable even to untrained observers.

ANALYTIC AND SOCIAL THEORIES

Psychoanalytic theories. *Freud.* Freud's psychoanalytic theories of personality have been among the most influential in the modern era of psychological study. The interest they attract stems in part from Freud's efforts to address himself to the practical problems of dealing with psychiatric disturbances, an orientation quite different from that typically found among academic psychologists. Freud wrote on so many different questions in attempting to explain the origins of psychiatric disorders that it is still difficult to put him in proper historical perspective. Freud's contributions have received many divergent interpretations, and many of his ideas are now viewed as models, suggestions, and intellectual metaphors.

Id, ego, and superego

Freud's key view of human personality as a system involved three concepts: the id, the ego, and the superego, each of which was held to constitute a portion of the total personality, and each of which was said to have a unity of its own. The three segments, however, were supposed to operate in interaction, and the behaviour of the individual was said to be the consequence of this interaction. Separately, the id was seen as the original "structure" of the personality system and as consisting of inborn, primitive, inherited characteristics or tendencies. It was depicted as the source of psychological energy initially, and the ego and the superego were held to develop from this source. The id was associated with what Freud called the "pleasure principle," reflected in the tendency of the person to maintain a pleasant, tolerable energy level through the relief of inner tension. The "pleasure principle" may be described in more mundane language as applying to the activities of the individual in achieving pleasure and avoiding pain. Thus, the loud crying of a hungry infant was called an id activity, it being assumed that the baby was responding to his inborn tendency to avoid the discomfort of an empty stomach.

The ego, associated with what Freud called the "reality principle," was held to develop from the id so that the individual could negotiate realistically with the objective world. The infant who experienced pain on touching a hot object was said to be developing his ego (reality contact) by learning to avoid the object. Thinking realistically provides one with means for the satisfaction of his primitive needs and thus serves as a secondary means for tension reduction. The ego was portrayed as controlling intellectual and perceptual functions (the so-called higher mental processes) involved in the control of behaviour. The ego was held to be coordinated to satisfy the inborn instincts (implicit in the id) but primarily to be oriented in a direction that serves to maintain the individual comfortably in relation to his environment.

The superego was theorized to be the third segment of the personality to develop and to represent the person's adoption (internalization) of social values; *e.g.*, the table manners he learns from his parents. Morality is said to be incorporated into the superego, which is supposed to evaluate the correctness of one's behaviour from the point of view of the standards of society. What is called conscience was held to be manifested by part of the superego. This theoretical division of the personality was seen to function in controlling sexual and aggressive impulses that derived from the id. With regard to the ego, the superego was held to impose "higher" moral goals as substitutes for "lower" realistic (ego) aims. In general, the superego might be said to represent the behavioral ideal.

Freud's theory of *Triebe*

Freud's view of personality rested firmly on his theory of *Triebe* (German: "needs," "drives," "urges," or "instincts"), with biological factors conceived as giving the driving force to the personality through the id. This source of psychic energy was said to be manifested as the person's wishes (corresponding to specific needs). *Triebe* were said to be of two types: (1) so-called life instincts, including the primary drives of sex, hunger, and thirst, were identified with Freud's term libido (or life energy). The orientation of the personality system was theorized to be directed toward satiation of these instincts (see also MOTIVATION). The psychic (libidinal) energy involved could also be dissipated by alternate means; expending energy in this substitute way is known as sublimation or displacement. (2) Correspondingly, there were said to be death instincts, of which an important group is labelled the aggressive motives. These id instincts were said to be controlled and harnessed by the ego, which drew psychic energy from them. The ego, as the "manager" of the personality, became crucial in psychoanalytic theory to the understanding of the learning processes involved in identification and displacement.

There are many aspects of Freud's theory, including his analysis of stages of human sexual development (see HUMAN BEHAVIOUR, DEVELOPMENT OF: *Personality Theories*). Concepts such as these have been among the most controversial in Freud's theory. Critics of the theory have emphasized its nonexperimental, empirically unverifiable aspects; id, ego, and superego are not observable entities. Some Freudian writings, including a description of the origin of the incest tabu, have been taken by some social scientists as being speculative, imaginative, and unsupportable in any rigorous way. While his critics see little empirically based science in Freud's formulations, his followers most enthusiastically seem to endorse and use them in efforts to treat sufferers of mental disorder.

Jung. Once a follower of Freud's, the Swiss psychiatrist Carl Gustav Jung had a substantial influence in the field of personality theory but did not develop the kind of massive following Freud did. Jung had considerable contact with Freud during the early period of their work, until, beginning in 1914, there were clear signs of a break between them, apparently prompted at least in part by Jung's rejection of Freud's pervasive concern with the sex drive as a basis for psychoanalytic interpretation. Jung agreed with Freud in attaching great importance to unconscious motivations but, unlike Freud, he placed even greater emphasis on interpreting behaviour as consciously purposive and as guided by one's awareness of goals and aspirations.

Recognition was given to the biological basis of personality; for Jung, much in the way of predispositions in human personality is referred to the biological history of man as an animal species. In this context, Jung theorized the inborn existence in each person of a "collective unconscious" that represents one's inherited accumulation of "memories" of ancestral experience dating from the earliest prehistoric man. The emphasis of this concept is on the common biological heritage that a species develops, all men being said to share a similar physical structure and correspondingly similar potential for behaviour.

In Jung's theory of personality, the ego constitutes the conscious "mind," reflected in the person's own feelings of self-identity. In addition (beyond the "collective unconscious"), Jung proposed for each person a "personal unconscious" in which one's unique, individual history of experience resides.

Jung agreed with Freud in holding that the effects of prior experiences that have been pushed back (repressed) out of consciousness may be drawn from the "personal unconscious." Jung's theory is full of historical, speculative, and imaginative references and analogues; his notion of personality as persona, for example, derives from the ancient (Etruscan?) name for the masks worn by actors. In this sense, personality is conceived as the mask the individual advances in maintaining his conventional place in society. As persona, personality is seen to be one's public presentation of himself in a manner likely to be socially acceptable. Jung also held persona to be consistent with the primordial images (archetypes) that he supposed were drawn from the "collective unconscious," which theoretically served as the repository of ancestral memories. The reappearance of deluge, or flood, myths (*e.g.*, that of Noah and the ark) in many different cultures worldwide was offered by Jung as evidence for a common set of archetypes shared by all people.

Jung's archetypes

Adler. While the writings of Freud and Jung may be contrasted with more experimentally grounded, so-called eclectic scientific theories of personality, many other authors were strongly influenced by Freudian and Jungian notions. More recent analytic theories have tended to stress social factors, emphasizing the malleability of the individual and the effects of social experience on the formation of his personality. Some theories, such as the "Individual Psychology" of Alfred Adler, an Austrian psychiatrist who once was a prominent associate of Freud and Jung, gave little or no weight to sexual and archetypal motives. Adler's theory of personality is known for its emphasis on a human "drive to power" (superiority) and on the importance of conscious (rather than unconscious) activities. Adler wrote extensively that all individuals must cope with feelings of inferiority, stressing compensatory processes (*e.g.*, as exhibited by the short person who compensates by being aggressive). In striving for social acceptance and influence, the person was viewed by Adler as constantly struggling toward the goal of what many theorists call self-realization.

Social analytic theories. Most of the socially oriented analytic theories of personality are of relatively recent vintage; what is common to these various approaches is that they are not basically experimental, having arisen largely in clinical settings and from intuitive thinking. To the extent that they have removed the conceptual rigidities and mysticism of the Freudian fountainhead from which they were largely drawn, such theories have tended to gain acceptance more readily among academic, empirically oriented psychologists such as Henry A. Murray, whose system for defining and measuring the strength of human needs has influenced research extensively. Concern with the study of what Murray called need-achievement, for example, is a major emphasis among social psychologists who seek to subject human behaviour to objective, statistical analysis.

The theories of Murray and Allport

Murray's list of human needs seems to show considerable correspondence to sets of personality characteristics that have been delineated through the painstaking statistical analysis of the responses given by large groups of people to items on personality tests. What is most characteristic of theories such as Murray's is evidence that they have been generated in much broader investigative contexts than was the case for their Freudian forebear. Thus, Murray sought to adapt his theory to a wide range of research data from psychological, anthropological, and sociological studies. Murray particularly worked toward the quantitative, empirical validation of his speculative concepts, constructing the Thematic Apperception Test (see PERSONALITY, MEASUREMENT OF) in an effort to translate his concerns with unconscious and other subtle motivational aspects of personality into measurable dimensions.

The view of personality presented by the U.S. psychologist Gordon W. Allport represents a focus on the individual as a unique, qualitatively different (idiographic) entity in each case. Allport, for example, initiated efforts to measure morality or values in personality and helped to develop a well-known personality inventory called the Allport-Vernon-Lindzey Study of Values. He viewed human behaviour as being motivated by relatively stable personality characteristics, or traits; his theorizing tended to be eclectic and to be compatible with clearly defined scientific tradition. Allport also is known particularly for the introduction of such concepts as that of functional autonomy, which describes the tendency of learned motives to be self-sustaining long after the historic reason for them has ceased to exist. Thus, an adult neurotic who suffers from an abnormal fear of heights (acrophobia) may have learned the trait as an infant by responding to the anxious behaviour of oversolicitous parents. Yet, long after the parental signs of alarm have ceased, often long after the parents have died, the acrophobia persists as a functionally autonomous personality characteristic.

RECENT ECLECTIC THEORIES

Contemporary personality study is characterized by a tendency of investigators to interpret their research findings on the basis of the eclectic use of many theories rather than in terms of a single, universally applicable, "grand" theory of personality. Indeed, this very practice may seem to support (possibly unfair) conclusions that the discipline of personality study suffers from comparative scientific immaturity, even though it has a rich history dating back to the earliest recorded human cultures (see PSYCHOLOGY, HISTORY OF). It also must be observed that lay people and workers in other fields who are not aware of the various developments in the growth of academic psychology and sociology may not recognize the evidence that many topics in personality theory were well developed early in the history of these disciplines. Contributions such as those of the U.S. psychologist William James (born 1842), concerning the analysis of human instincts and learned behaviour, are reasonably well known.

Role theories. On the other hand, monumental and incisive work on the development of human personality that was the contribution of an American, James Mark Baldwin, is less known, even among modern investigators and clinicians. Baldwin presented the skeleton and much of the fleshing of what can be described as a behaviourist theory of personality socialization and had an important influence on the U.S. sociologist Charles Horton Cooley. Cooley's volume on *Human Nature and the Social Order* at the turn of the century provided for a generation an eclectic and more or less scientifically respectable introduction to a social psychology focussed on the individual and his social behaviour. This tradition led subsequently to further elaboration and theoretical development in the description of learning processes through which the human infant is brought into society and develops the behavioral characteristics of a socially active person. This particular behaviouristic orientation has been associated with the concept that much social interaction has symbolic meaning to the participants, particularly in the social roles they play. So-called role theory has not yet produced a coherent body of empirically based generalizations; still, proponents of this approach (the study of personality as social role playing, as persona) have the advantage of being relatively free of the stifling constraints imposed by some rigidly dogmatic theories of personality.

Trait theories. An alternative approach to human personality among academic psychologists is found in the work of a British psychologist, William McDougall, and others who have outlined various classifications of personality traits. Though McDougall once was widely criticized for his use of the term instinct in specifying the driving forces of human personality, the notion of instinct (*q.v.*) has today recaptured considerable respectability in scientific circles. McDougall ordered personality characteristics as primary (inherited) propensities and as secondary (acquired or learned) tendencies. Such a stratification of personality characteristics is common to many theoretical attempts to organize personality traits.

McDougall's notion of instinct

In more modern dress personality theory drawn directly from the empirical study of human traits has attracted the attention of many psychologists. In part, this movement grew from developments in educational psychology, particularly from progress in the construction of psychological tests. The idea of using tests for measuring intellectual characteristics was well established by their successful use in the psychological screening of military personnel during World War I. By the 1930s the statistical technique called factor analysis (a procedure for the examination of the structure of a test) had stimulated this type of development. Contemporary psychologists published masses of empirical data showing the tests to measure sets of statistically distinct "factors" or behavioral dimensions that could be used to define personality traits. Efforts have been made to demonstrate parallel sets of factors derived from empirical biological data and from psychological tests designed to measure personality traits at the level of verbal behaviour. There has been some debate on the quality of evidence thus far offered to demonstrate psychobiological parallelism (or

Factor analysis

even unity) in human personality, but there is gross evidence that spectacular personality changes can be induced biologically (*e.g.*, by drugs).

Factor-analytic studies at times may appear to be exploratory and atheoretical, but intrinsically the theory that underlies them is that of all empirical science. In particular, the theory for these approaches rests on the axiom that there are generalizations about human personality that are empirically discoverable and verifiable. Indeed, agreement in repeated factor-analytic studies of personality measures by independent researchers has been considerable. One approach to the description of personality, for example, is through ratings of an individual by others. Factor-analytic studies of such ratings repeatedly have provided evidence for such generalized personality characteristics as assertiveness, sociability, emotionality, and conscientiousness. Traits based on ratings may be described as those ways in which individuals behave and differ from each other that are reliably discernible to others. This is sometimes taken as the focal definition of what one wants to know about a person; namely, how others see him. The structure of personality as defined at this level relates to the theoretically unimportant approach of symbolic social interaction. According to the symbolic-interaction theorists, the identity of a person is defined by his perception of the responses of others to him; that is, by how he perceives his value as a social stimulus.

In factor-analytic studies of formal measures of personality (*e.g.*, pencil-and-paper personality inventories or tests) there has been less agreement among different investigators in delineating specific human traits. In part this results from the great variety of tests available, the many different ways of asking questions of the test taker, and of scoring and interpreting test responses. Also, the theoretical ambition of having all the relevant questions about a person's preferences, attitudes, and values measured in any one test is not to be satisfied, if only for practical reasons. Thus, different tests do not necessarily sample the same kinds of behaviour. Additionally, personality-test construction has not been performed by equally sophisticated investigators in all circumstances.

One objective of factor analysis and of related techniques is to reduce the number of dimensions required to account for most of the variation in individual responses to all test items. Thus, a personality inventory with three hundred items may yield evidence for only ten or fifteen personality traits (factors or dimensions) that are relatively distinct (*i.e.*, statistically independent of each other). The traits identified by different investigators often seem to differ. Sometimes these discrepancies may be more apparent than fundamental, however, and may simply result from semantic differences in naming the factor. In other studies the differences seem to indicate basically different personality traits, as revealed when the same set of subjects is used to check two sets of earlier findings.

From the point of view of agreement, however, repeated factor-analytic studies with different personality tests do indicate some traits that emerge with at least a degree of consistency. These have commonly been identified as the dimensions of introversion–extroversion and of social adjustment. Particularly interesting in these findings are the suggested ties to earlier work; for example, Jung's early interest in and emphasis on the introversion–extroversion continuum in accounting for such abnormalities of personality as schizophrenia and the manic-depressive psychosis. Personality inventories based heavily on clinical hunches or intuitions (popular before the advent of more sophisticated test-construction techniques) emphasized concepts of the sort offered by Jung.

Statistical procedures used in the study of personality permit the characteristics of a single individual to be charted over periods of time, yielding evidence as to either the stability or transient nature of some traits. Although the technique is fraught with problems, what is called inverted factor analysis is used to examine the degree to which groups of individuals share the same traits. Thus, individual people take the place of tests or test items in the factor analysis. Investigators who use this technique hope to discover with its help more precise ways of grouping people for such purposes as psychiatric diagnosis.

At present, the scientific concerns with the study of personality tend to de-emphasize reliance on general theories of personality and to place more stress on the study of specific empirical aspects of human behaviour. Hopes for developing a universal theory of personality probably are no more realizable than is the dream of a global theory to account for the functioning of the human organism as a biological entity. Apparently, the theory can be no less complex than the organism itself. What constitutes personality theory at any point in time thus depends on a cumulative mass of information, segments of which may be understood in great detail while others may remain to be discovered.

BIBLIOGRAPHY. E.F. BORGATTA and W.W. LAMBERT (eds.), *Handbook of Personality Theory and Research* (1968), an encyclopaedic collection of reviews on many topics, with extensive bibliographies; I.G. SARASON, *Personality: An Objective Approach* (1966), a comprehensive textbook emphasizing personality theories; J.M. WEPMAN and R.W. HEINE (eds.), *Concepts of Personality* (1963), topical presentations, a number of which focus on theoretical approaches to personality; R.B. CATTELL, *The Scientific Analysis of Personality* (1965), a volume representative of personality theory as viewed from the perspective of a measurement-oriented scientist; J.P. GUILFORD, *Personality* (1959), a comprehensive survey of personality from a traditional psychological point of view, relatively eclectic; N. SANFORD, *Issues in Personality Theory* (1970), a recent consideration of important problems in the study of personality.

(E.F.B.)

Personnel Administration

Personnel administration is the management of the people in working organizations. It is also frequently called personnel management, industrial relations, employee relations, and manpower management. It represents a major subsystem in the general management system, in which it refers to the management of human resources, as distinguished from financial or material resources. The term may be used to refer to selected specific functions or activities assigned to specialized personnel officers of departments. It is also used to identify the entire scope of management policies and programs in the recruitment, allocation, leadership, and direction of manpower.

Personnel administration begins with the process of organizing—the process of defining the required quantities of particular personal capabilities. Thereafter, people must be found, recruited, selected, trained or retrained, negotiated with, counselled, led, directed, committed, rewarded, transferred, promoted, and perhaps released or retired. In many of these relations, managers deal with their associates as individuals (the field takes its name in part from this type of relationship). In some working organizations, however, employees are represented by unions, and managers bargain with these associations. Such collective bargaining relationships are generally described as labour relations.

HISTORY

Early practices. The history of personnel jobs is one of persistent job enlargement. Today's personnel departments and activities bear little resemblance to earlier personnel officers and assignments. Responsibilities have expanded and job specifications have changed from low-skill requirements to technical and professional levels of competence.

Early personnel positions

Many early personnel positions were created to implement the policies of paternalistic owners. Those who held these early assignments in England have been described as "welfare workers"; counterparts in the United States have been characterized as "amateur social workers." Expected to convey to rank and file employees the employer's interest in their welfare, early personnel people had to become acquainted with rank and file workers and their personal and family circumstances and to express the employer's sympathy when they faced difficult personal and family problems.

In what was perhaps the most common early development, assignments were expanded to include responsibilities for selection. As this function evolved, personnel workers interviewed applicants "at the gate," and later, in reception rooms. They investigated work histories and experience. Still later, some of them began simple testing programs. Their offices became depositories for personnel records. When unions became common, some personnel officers were expected to prevent them by blacklisting troublemakers. As collective bargaining expanded, some departments handled negotiations and contract administration.

Contemporary practices. All around the world, assignments to personnel departments clearly reflect expanding industrialization and accompanying social and political developments. Developing nations have faced the problem of adapting their agricultural labour forces to industrial employment. Specialized personnel officers have been expected to reorient workers and to enforce safety and security programs. In the economic development process, public policy has turned toward objectives of "full employment," shorter working days and weeks, higher real wages, equal employment opportunity, and increased economic security. Technological change has created new skill requirements and reduced employment for "raw," unskilled labour. Lead times between decisions to hire and actual employment have become longer. Geographic boundaries of labour markets have broadened as labour mobility has increased.

As skill shortages created new demands for manpower planning and training, rising levels of citizen education and sophistication complicated the employer's problem of gaining commitment. More sophisticated employees sought promises of new satisfactions in their work. Problems in employment behaviour attracted growing attention from the rejuvenated behavioral sciences. New theories of human behaviour suggested new approaches in manipulation and direction. Managers recognized the need for new patterns and styles of leadership and new structures of working organizations.

Utilization of human resources

Today's personnel assignments have emerged in managerial reactions to industrialization. In modern industrial economies, the manager's responsibilities for human resources require a complex mixture of programs. In less industrialized nations, on the other hand, programs emphasize recruitment, reorientation, and employment security, as political leaders seek to gain commitment to industrial jobs. Labour movements may emerge in these nations as protests against the regimen and control of mass employment and are likely to be considerably more explosive than are the older unions of more "mature" economies. New unions may be oriented more strongly toward political action than toward collective bargaining.

Government regulation. Governments in the developing nations, sensitive to worker fears of mass unemployment and of health hazards, usually try to provide extensive benefit programs. Personnel managers may enforce limitations on working hours, ensure approved working environments, and administer a variety of supplementary benefits. In India, for example, employment of an employee welfare officer with such responsibilities is required by law in firms with more than 500 employees. In contrast, in the more mature, industrialized nations, assignments to personnel departments are much broader and more varied. Many current discussions refer to the personnel "system," implying a scope as broad as the total of interacting variables that influence employment behaviour.

Personnel associations. Personnel workers have created numerous local, national, and international associations. They sponsor periodic meetings and conferences and publish journals that report on research, experience, and the opinions of practitioners. They include the European Association for Personnel Management, the Institute of Personnel Management in England, the American Society for Personnel Administration, the Society for Personnel Administration, and the Industrial Relations Research Association in the United States and Canada. Training and safety officers and wage and salary administrators have created more specialized associations (for example, the American Society for Training and Development and the American Compensation Association).

PERSONNEL DEPARTMENTS

Responsibilities. Current practice shows wide variation in the range of responsibilities assigned to personnel or industrial relations departments. Organizations with 100 or more employees generally provide some such specialized services. The average "personnel ratio" is about 0.50, meaning one full-time professional or technical personnel worker for each 200 employees. In larger organizations, personnel departments usually include a variety of personnel specialists.

Major areas of responsibility

Major areas of personnel department responsibilities include: (1) organizing—devising and revising organizational structures of authority and functional responsibility and facilitating two-way, reciprocal, vertical, and horizontal communication; (2) planning—forecasting personnel requirements in terms of numbers and special qualifications, scheduling inputs, and anticipating the need for appropriate managerial policies and programs; (3) staffing or manning—analyzing jobs, developing job descriptions and specifications, appraising and maintaining an inventory of available capabilities, recruiting, selecting, placing, transferring, demoting, promoting, and thus assuring qualified manpower when and where it is needed; (4) training and development—assisting team members in their continuing personal growth, from pre-employment, preparatory job training to executive development programs; (5) collective bargaining—negotiating agreements and following through in day-to-day administration; (6) rewarding—providing financial and nonfinancial incentives for individual commitment and contribution; (7) general administration—developing appropriate styles and patterns of leadership throughout the organization; (8) audit, review, and research—evaluating current performance and procedures in order to facilitate control and improve future practice.

Individual personnel departments may be assigned varying degrees of responsibility in a few, many, or all of these areas. In areas assigned to them, personnel departments exercise various levels of authority. Some officers and departments create policy and make major decisions and determinations. Some study and recommend policy and create and direct appropriate programs. The personnel vice president may be a member of the executive cabinet; he may be expected to assume leadership in and responsibility for all manpower management policy and programs. Other personnel departments are essentially "staff" or advisory; their activities are restricted to recommending, consulting, and providing such specified technical and professional services as are requested by operating managers. Such a department, for example, may provide the "legwork" of recruiting or of interviewing or testing. All important decisions may be reserved to line managers. In this latter type of assignment, personnel departments provide "staff services" for management. Among the most common examples, the personnel staff (when, as, and if requested and authorized by operating managers) analyzes jobs and prepares job specifications, visits sources of potential employees, conducts preliminary screening interviews, administers tests, supervises rating or appraisal programs, counsels employees, conducts training and development programs, negotiates agreements, monitors grievance settlements, maintains safety and accident control programs, administers employee benefits and services, prepares forecasts of future manpower requirements, recommends changes in organizational structures, supervises formal in-house communication, conducts employee attitude and morale surveys, and supplies other related services, sometimes including public relations.

Staffing policies and programs. Among the earliest and most persistent of assignments to personnel officers is that of finding and choosing people to fill present or future jobs. These staffing programs are built around several steps or stages, including: (1) manpower planning—forecasting manpower requirements and creating necessary

policies and programs for filling openings as they appear; (2) job analysis—providing information for job specifications that identify required skills, competence, and potential capabilities for each job; (3) recruitment—identifying likely sources of potential candidates who are informed of opportunities and encouraged to apply; (4) selection—using interviews, tests, references, and other sources of relevant information to identify candidates to be offered employment; (5) placement and induction—introducing newly selected employees to working colleagues and conditions and explaining task assignments.

Current policy objectives

Manpower planning. The simplest form of staffing policy proposes to find and hire whatever manpower is needed when it is needed. Current policy in private firms and public agencies frequently emphasizes these objectives, but more often more complex staffing is required. Public policy may insist on equal employment opportunities for all. Managers may prefer to hire for a long working career instead of for a single job. They may propose utilization of the highest skills and aptitudes and maximization of employee satisfaction in the job, relating this policy to participant commitment and contribution. Policy may propose preference for employees who are physically handicapped or are available only on a part-time basis. Policy may stress the care and advancement of present employees.

Job analysis. Staffing is designed to fill manpower needs identified in the organizing process, which includes division of labour, specification of tasks to be accomplished, and specification of combinations of tasks in jobs. Thereafter, job analysis spells out these operations, noting the knowledge, skills, and working conditions and jobholder qualifications thus prescribed. It provides the facts for "job descriptions," which systematically summarize this information. "Job specifications" detail the personal qualifications required in such terms as education, minimal scores on tests of knowledge, skill, and aptitudes, experience in the same or related occupations, and other such personal characteristics. "Staffing tables" or "manning tables" summarize quantitative requirements—that is, numbers of jobholders to be employed in each job at specified levels of plant operation. "Time study" may be used to develop job or performance standards for each job.

Recruitment. Sources of acceptable candidates for vacancies may be internal or external. Policy frequently proposes to give priority to internal sources—that is, present employees. That policy, described as "promotion from within," may be implemented by maintaining a "capability inventory" that summarizes the skills, experience, education, and talents of present employees. To fill vacancies, present employees may be promoted—that is, reassigned to jobs with greater responsibilities—or transferred to other jobs at similar levels. Recruitment from within may be guided by a "seniority system" expressing the intention to regard length of service in the organization as a major consideration in promotion or retention. Use of internal sources often places heavy weight on "personnel appraisals" or "ratings," which are periodic evaluations of individual performance or potential by supervisors or peers.

Among the most widely used external sources are "waiting lists" (lists of applicants who have sought employment but have not been hired), unions, schools, public and private employment agencies, executive-search and temporary-help agencies, and professional societies and associations. Search may require extensive lead time for some highly skilled, technical, and professional workers. Recruitment may extend beyond national boundaries. Recruiting programs may encourage nominations by present employees. Employers may advertise in local or distant publications. Recruiters may use brochures describing career and job opportunities. Recruiting teams may be sent to preferred sources. Experienced scouts may visit vocational or secondary schools, universities, or foreign sources.

Selection. Selection separates those offered employment from all others. The objective is to maximize the probability that selected applicants will meet employer specifications and expectations—to hire the best and reject the rest. The selection process has often been described as a series of successive hurdles. Each hurdle presumably pits the qualifications of candidates against one or more requirements called out in job or career specifications. Some programs identify crucial requirements—for example, citizenship, union membership, a credential, or a license—which may be given priority, so that unsuccessful candidates need go no further in the selection process. Among the most widely used screening techniques and devices are the following:

Identifying and measuring capabilities

1. Application blanks, in which applicants provide requested information on such characteristics as age, sex, education, special training, and experience. "Weighted" application blanks include items that can be scored to predict probabilities of success. Practice may require completion in handwriting by the applicant, sometimes designed to permit handwriting analysis.
2. References, the names and addresses of individuals who can provide personal evaluations. Some are essentially character references; others have knowledge of earlier working performance.
3. Physical examinations may be a critical hurdle. In current practice, because of the cost of medical services, these checks may be used only for the most likely candidates.
4. Interviews are the most widely used selective device. Preliminary screening interviews may determine whether or not candidates are admitted to the remainder of the process. Applicants may face a panel or board interview, meeting with a group of supervisors or managers, or they may be interviewed by an individual job supervisor or manager. Selection interviews may be structured or unstructured—that is, carefully preplanned or merely a conversation that follows its own course. Frequently the interviewer will ask certain specific questions. Weaknesses of the interview as a selective device are now widely recognized, but despite these deficiencies, interviews remain a major selective technique, in part because no entirely satisfactory substitute has been found.
5. Observational methods seek to read personality and character from physical traits and appearance. Despite evidence that assumed correlations are dubious and not demonstrable, some employers continue to use phrenology—the examination of head shape; others lean on a more inclusive assessment generally described as character analysis, which infers a wide range of personal traits from individual physique.
6. Tests are essentially small samples of the behaviour and capability of the candidate. Some of them are widely described as paper-and-pencil tests; they require written marks or replies to the questions they raise. Others are essentially performance tests. Major types used in selection include "achievement tests," measuring capability to perform specified tasks (for example, to type with various degrees of speed and accuracy); "aptitude tests," which identify and compare undeveloped as well as developed personal talent or capability (as for example, mechanical, musical, or literary aptitude); "interest tests," which use checklists and required completions to discover areas of major interest to the applicant; and "personality tests," which seek to reveal the whole pattern of basic personal characteristics, including maturity and temperament. Some personality tests are projective—the applicant's interpretations of pictures or inkblots are regarded as evidencing projections of and clues to his personality. Individual firms may develop their own tailor-made test batteries, but many of the most widely used tests are standardized; they have been administered to large numbers and analyzed statistically, thus providing average ranges and norms of scores.

Placement and induction. The process of introducing a new employee to his working colleagues and conditions and explaining his task assignments contains a number of elements: (1) he must have a simple, precise definition of the work to be done and of the conditions surrounding it; (2) he must be assigned specific work to do and be vested with the authority to do it; (3) he must be informed about his relations to others engaged in the en-

terprise; (4) he must be shown how to perform the tasks assigned to him; and (5) he must know the standards and means by which his work is evaluated.

Separation. In any dynamic economy or organization there is bound to be a labour turnover, and there have developed a few business methods of handling the problem (in addition to such governmental methods as unemployment insurance and employment services). The oldest and most common practice is the dismissal notice —an advance notice to the employee, allowing him to remain on the job for a certain length of time while seeking other work.

Company plans that provide dismissal notices usually also offer what is variously called "severance pay," "dismissal compensation," or "separation grant." This payment (which is in addition to any back wages or salary or any vacation pay) is given a worker when his employment is terminated after having served a minimum length of time (usually at least a year); the amount of the payment is geared to the length of service. Most plans grant severance pay for unavoidable—though permanent—layoffs resulting from the elimination of a job category, a plant shutdown, employee disability, or the like. Many plans pay the allowance even to employees discharged for incompetency or similar cause. Few make payments to anyone who voluntarily quits.

Especially in large corporations there has also developed the alternative to dismissal known as the transfer policy. With a knowledge of the situation in each division or subsidiary of the corporation, the personnel department is in a position to anticipate changes in personnel requirements and to transfer surplus workers to other sections requiring additional help.

Changing staffing policy and programs. Questions and concerns about several traditional staffing programs have been widely publicized. Some of the traditional hurdles may be inconsistent with a public policy that proposes equal employment opportunity. Job specifications may be unrealistic, so that artificial barriers interfere with hiring. For example, requirements of secondary school or university graduation or of lengthy experience, and limitations based on sex, colour, and age are frequently questioned, as are abilities to provide a surety bond or a record of no arrests or convictions. Cultural bias in tests and interviews is recognized; new tests and revised procedures seek to ensure "culture-fair" applications. Selection programs may also involve unacceptable invasions of personal privacy, as in use of lie detectors and the investigation of financial histories and credit ratings.

Training and development programs. Helping both managers and workers learn whatever knowledge and skills are required is a widely accepted responsibility in modern manpower management. In-house programs assist individuals in becoming adjusted to and prepared for work. Their behavioral objectives are more satisfactory performance and more satisfying on-the-job experience. Programs concerned with technical skills are described as "training," whereas those that deal with analytical and leadership skills and understanding for supervisory and management jobs are labelled "development."

Improving technical and managerial skills

Several levels of instruction may be distinguished, including: (1) pre-employment training, designed to aid inexperienced workers to make the transition from schools or unemployment into the employed labour force; (2) job training, for specific job requirements of knowledge and skill; (3) supervisor training; (4) manager or executive development; and (5) a variety of special courses tailored to prepare participants for new problems, rapidly changing assignments, transfers to foreign operations, new materials, new processes, and new organizational arrangements. Examples include special training for supervision of minority groups; preparation for retirement; basic reading, writing, and speaking; so-called sensitivity training (in which individuals come together in a kind of group therapy—studying their own adjustments and reactions to group situations); team development courses; and preparation for citizenship.

The range of training and development programs is a function of general manpower policy, organizational resources, and environmental constraints. For some job assignments, management may be forced to offer training courses because no source can provide qualified recruits. Most training and development programs, however, express a typical "buy-or-make" decision: some employers may offer very little training, preferring to purchase required capabilities, whereas other employers may conclude that such programs are less costly than recruiting, assure superior quality, are more dependable, or tend to improve employee commitment.

A policy stressing "investments in people" encourages in-house training and development programs on the assumption that courses pay off by increasing contributions throughout the careers of employees. Training may be related to the learning-curve or unit-cost function, which shows that continued employee experience and improvement in routine operations can result in reductions in unit labour costs.

Training techniques and practices range from lectures and demonstrations to programmed and computer-assisted instruction. Methods include classroom recitation, discussions, role playing, conferences, coaching, understudy arrangements, and apprenticeship. Training may take place on the job, in the shop, or in "vestibule schools," where trainees are under less pressure and do not interfere with ongoing production. Some programs provide impressive special facilities, permitting a multimedia approach with a variety of audiovisual facilities that mobilize all of the individual senses in learning. Some courses may be provided by local schools and colleges, with scholarships and tuition-refund plans to encourage employee participation. Executive development programs may include assignment to one or more of the several hundred university programs.

Commitment and incentive. Many experienced managers regard the task of gaining and holding the interest, commitment, and enthusiasm of their working crews as their most persistent and difficult task. Personnel departments are assigned numerous program responsibilities in this area. Worldwide transition from agricultural to industrialized economies has made historic incentives inappropriate; ancient attitudes of being "born to work" and of accepting orders from owners and masters have lost most of their earlier influence. Industrialized economies are tied to exchange systems; wages become the major rewards for effort and commitment, but managers must also recognize the importance of nonfinancial considerations. Personnel officers manage programs designed to maximize the payoff from both financial and nonfinancial inducements.

Managers seek high levels of individual commitment to the mission and to the goals of the organization. They hope to increase the interest, effort, and contribution of each participant. Commitment policy seeks to attain this objective through suitable reward or incentive systems. Major dependence is placed on exchangeable financial rewards. Recent, more sophisticated policy, however, recognizes employee sensitivity to a wide range of nonfinancial rewards that promise opportunities to satisfy a variety of personal needs. Rewards or commitment theory has thus moved from rather simple to much more complex models. Effort and contribution are assumed to be functions of opportunities for comprehensive personal self-actualization. Commitment levels are a response, not simply to financial rewards, but also to on-the-job promises of such satisfactions as economic security, group membership, status, self-esteem, and self-fulfillment. Some theory further suggests that opportunities for satisfactions in the work itself may be more influential than those offered by the work setting or environment.

Reward and incentive systems

Wages and salaries. Many personnel departments include a subdivision assigned special responsibility for "wage and salary administration" and involving two major types of compensation programs. One type provides direct money payments for participation and contribution. The other type offers deferred, indirect rewards—employee "benefits" and "services." Some immediate rewards are time based. Others are scaled to match measured performance and output; such "incentive wage"

plans vary from simple piece rates to more complicated arrangements that pay for time saved and provide bonuses for superior performance. "Wages" are paid to hourly rated employees; "salaries" are calculated by the month or year. Wage structures may provide "flat rates" (*i.e.*, a single amount) for each job level or classification, or "rate ranges" for "labour grades," with a series of step rates within each range. Adjustments in individual pay may be "merit" increases, or they may be "automatic"—granted at regular intervals, usually each half-year or year. Rates or amounts of pay are subject to a variety of adjustments—for example, compensation for changes in costs of living. Other adjustments match productivity increases, family allowances provide supplements based on numbers of dependents, and sliding scales adjust to match prices of products. Some programs assure guaranteed annual wages. Legal requirements may specify minimum hourly rates.

"Job evaluation" is widely used to establish comparative values of jobs. For professionals, "maturity curves," which relate pay to years of postgraduate experience, may be substituted for job evaluation. Managers use wage and salary surveys to discover current practice in other firms, agencies, and localities. Compensation frequently combines salaries and a variety of deferred payments (including options to buy stock at discounted prices).

Benefits and services

Numerous indirect, deferred payment programs are also widely used. Some benefits may be required by law—for example, those for unemployment, retirement, and work-related accidents and illness. Noncompulsory benefits are created on the initiative of employers or of unions. Benefits may be contributory; potential receivers may or may not make direct contributions. In terms of purposes or objectives, four types of such benefits and services are common: (1) economic security; (2) protection against health hazards; (3) old age and retirement; and (4) special personal assistance, such as legal aid, credit, counselling, recreational and social assistance.

Nonfinancial rewards. Nonfinancial rewards promise satisfactions not available in the market through purchase or exchange. They assume that work itself and the working environment can offer opportunities for satisfying a variety of human needs, and jobs and organizations are structured to provide such satisfactions. All such programs lean heavily on assumptions about the unsatisfied needs and expectations of rewardees. Managers may undertake periodic morale or attitude surveys to appraise individual and group job satisfaction and to identify unsatisfied needs of participants. Special programs encourage listening by supervisors and improved communications from workers to supervisors. Managers may publish house magazines, develop employee handbooks and manuals of standard operating practice, and assign specific responsibility for answering rumours and gossip. Some nonfinancial reward programs approach commitment through job redesign, job restructuring, and job enrichment, the basic concept being one of adapting jobs to the special and changing interests of jobholders. In all such efforts, the persistent problem for managers is that of limited understanding about what variables influence and condition on-the-job interest and effort.

CURRENT TRENDS

Growing manager recognition of the central importance of human resources has, over the present century, broadened the responsibilities of personnel departments to encompass every aspect of manager–worker relationships, from recruitment to retirement. In some current working organizations, this change is symbolized by describing the personnel operations simply as "administration" and giving its officers heavy responsibility for the total administrative style or stance of management. These overall administrative styles may be compared according to the degree of participation afforded workers in solving problems, influencing plans, decisions, and organizational structure, and thus exposing and reducing tensions and conflict. Patterns in total administration may be described in such terms as "authoritarian" or "participative."

Many personnel departments charged with such broad responsibilities are actively promoting programs of organizational or team development, people-centred supervision, group problem solving, supportive management, consultive management, conflict resolution, and organizational renewal. The long-term trend appears to be toward such broader assignments, coupled with "reverse delegation," in which line supervisors and managers accept greater responsibility for such specific programs as personnel appraisals, training crew members, day-to-day communications, compensation, and counselling.

BIBLIOGRAPHY. For more detail on the scope of current programs, see D.S. BEACH, *Personnel: The Management of People at Work* (1965); C.H. NORTHCUTT, *Personnel Management Principles and Practice*, 3rd ed. (1956); G.M. OXLEY, "The Personnel Manager for International Operations," in R.E. FINLEY (ed.), *The Personnel Man and His Job* (1962); G. RITZER and H.M. TRICE, *An Occupation in Conflict: A Study of the Personnel Manager* (1969); M.N. RUDRABASAVARAI, *Dynamic Personnel Administration* (1967); and D. YODER, *Personnel Management and Industrial Relations*, 6th ed. (1970). For the most thorough history of changing assignments, see C.C. LING, *The Management of Personnel Relations* (1965). For an excellent comparative description of current national systems, see F.H. HARBISON and C.A. MYERS, *Management in the Industrial World: An International Analysis* (1959).

(D.Y.)

Persuasion

Persuasion, in the sense to be used in this article, refers to the process by which a person's attitudes or behaviour are, without duress, influenced by verbal communications from other people. One's attitudes (*q.v.*) and behaviour are affected by other factors (for example, verbal threats, physical coercion, one's physiological states). Not all verbal communication is directly persuasive; some of it has such effects as informing, entertaining, or deceiving. But persuasion, in which communication and belief intersect, is a particularly pervasive and poignant manifestation of humanity, shown perhaps not at its best but at its most human. Persuasion involves manipulating other people, and for this reason many find the exercise distasteful. Others might argue that without some degree of social control and mutual accommodation, the human community becomes disordered and that persuasion indeed gains moral acceptability when the alternatives are considered. To paraphrase Winston Churchill's evaluation of democracy as a form of government, persuasion is the worst method of social control except for all the others.

Ancient writings on persuasion

So important a process inevitably attracts study from many quarters. Persuasion has been treated as an art, a craft, and a science by the great cultures that have left a substantial historical record. Thinkers from the time of and of the stature of Aristotle and Cicero devoted whole treatises to the topic. In European universities of the Middle Ages it was one of the basic liberal arts to be mastered by any educated man; and from the days of imperial Rome through the Reformation it was raised to a fine art by preachers who used the spoken word to move men to virtue and the Holy Land. In the modern era, persuasion, in the form of advertising, supports a major industry; only education and the military command a greater share of the gross national product.

Analysis of persuasive communication. The process of persuasion can be conveniently analyzed in a preliminary way by distinguishing verbal communication (as cause, or stimulus) from the associated changes in attitudes (as effect, or response).

Components of verbal communication. Alternative schemes for analyzing verbal communications into manageable sets of components often have been presented as questions to be answered. Just as medieval schoolmen suggested such questions as *quis, quid,* and *quando* ("who, what, and when"), so contemporary news reporters are trained to communicate their stories by describing who, what, when, where, why, and how. An American political scientist of the 20th century, Harold Dwight Lasswell, has said that communication is a matter of who says what to whom, how, and with what effect. In the language of the communication engineer, these factors are source, message, channel, receiver, and destination.

Source factors include all of the characteristics of the perceived sender of the communication. The "message" refers to what he says and includes style, content, and organization; "channel" designates the medium (*e.g.*, press, radio) through which the message is communicated; "receiver factors" are the characteristics (*e.g.*, age, sex) of the person to whom the communication is directed. "Destination" indicates the behaviour (*e.g.*, voting) the communication is designed to influence.

Steps in persuasion. Analysis has led to the delineation of a series of successive steps through which a person goes in being persuaded. The communication first is presented; the person pays attention (*q.v.*) to it, and he comprehends its contents (including the basic conclusion being urged and perhaps also the evidence offered in its support). For persuasion to be effected the individual must yield to, or agree with, the point being urged; and, unless only the most immediate impact is of interest, he must retain this new position until opportunity arises to act on it. Finally, if the fruits of the persuasive process are to be reaped, he must carry out the behaviour implied by his new attitudinal position; *e.g.*, he enlists in the army or becomes a Buddhist monk or begins to eat cornflakes for breakfast.

Theories of persuasion. Persuasion research has been undertaken with a great variety of orientation and purpose, and the consequent theoretical diversity, though it gives an untidy appearance, provides a rich set of alternative viewpoints. The theories arose more in ignorance of than in opposition to one another and should be viewed as supplementary rather than contradictory. When there is theoretical conflict, resulting empirical research may provide clarification.

Learning. Some, but by no means all, theorists emphasize a similarity between education and persuasion. They hold that persuasion closely resembles the learning of new information through informative communication. Thus, since repetition in communication will modify learning, they infer that it will have persuasive impact as well; that, for example, principles of verbal learning and conditioning are widely and profitably applied by persuaders (as in the judicious repetition of television advertisements). The learning approach (see LEARNING THEORIES) tends to emphasize attention, comprehension, and retention of the message.

Perception. One's reaction to persuasive communication depends in part on the message and to a considerable extent on the way in which the receiver perceives or interprets it. Words in a newspaper advertisement may take on different persuasive qualities if they are printed in red instead of in black. Perceptual theorists regard persuasion as altering the person's perception of any object of his attitudes (see LEARNING, PERCEPTUAL). Perceptual approaches also rest on evidence that the receiver's preconceptions are at least as important as the message content in determining what will be understood. The approach stresses attention and comprehension.

Functional approaches. While learning and perceptual theorists may stress objective intellectual steps involved in the process of being persuaded, another, functional, approach emphasizes more subjective motivational aspects. It views man as essentially ego-defensive; that is, as an organism with beliefs that function to satisfy his conscious and unconscious personal needs and that may have little to do with the objects toward which his attitudes and actions are directed. Functional theorists say, for example, that one's ethnic prejudice and other forms of social hostility are often manifestations of his underlying personality structure, reflecting his difficulties in dealing with authority. It is held that his childhood hostility toward his father was severely punished and that he represses such tendencies toward authority in later life, tending to keep such feelings in check by idealizing the powerful groups within society and by despising deviant and low-prestige people. His hostility to such outgroups is said to derive more from his own personality structure than from information about the nature of these groups.

The role of prejudice

According to functional theorists, such attitudes are to be changed by helping the person handle his underlying personality problems rather than by informing him about the people he dislikes. Particularly directed toward emotion-laden attitudes, such theories are less relevant to a broad range of persuasive efforts that deal with more objective, less emotional issues and in which unconscious motivation may play a lesser role.

Conflict resolution. Other theories view the person confronted with persuasive communication as being in the harassed role of trying to find some reasonable compromise among many conflicting forces acting on him; *e.g.*, his own desires, his information, and the social pressures put upon him, including this new communication. Those (frequently called congruity, balance, consistency, or dissonance theorists) who stress this conflict-resolution model of man focus on how people weigh these forces in adjusting their attitudes. Some theorists who take this point of departure stress the intellectual aspects of persuasion, while others emphasize emotional considerations.

Information processing. Each of the approaches considered above tends to neglect one or more steps in the process of being persuaded and thus serves to supplement rather than supplant the others. Another more eclectic and inclusive approach, growing out of information-processing theory, is oriented to a consideration of all of the options implied by the communication aspects of source, message, channel, receiver, and destination; and each option is appraised for its persuasive efficacy in terms of presentation, attention, comprehension, yielding, retention, and overt behaviour.

Methods of research. A typical persuasion experiment of the 1970s may begin with a hypothesis about how some communication factor will affect persuasive impact. In connection with a public-health campaign to persuade people to get an annual physical checkup to detect cancer, for example, the working hypothesis may be that a person's intense anxiety about the prevalence and seriousness of cancer will motivate him to seek a diagnostic examination. The researcher may develop this hypothesis further by theorizing that messages designed to arouse high anxiety might be so frightening that they will discourage people from facing the potential "horror" that a medical examination might reveal.

A typical experiment in persuasion

To evaluate such hypotheses, the investigator might design a set of messages in each of which the informational part is the same but each of which could be expected to evoke different levels of anxiety by varying the stress and vividness with which the incidence and seriousness of cancer are described. He would then devise measures (*e.g.*, questionnaires) of how well the communication is understood (comprehension) and of the extent to which people agree with the recommendations (yielding) and intend to get the medical checkup; he may also prepare a measure of retention, hopefully arranging to check if medical examinations are actually obtained (overt behaviour). The researcher then assembles and studies as many matched samples of people as he has versions of the message, one version being presented and evaluated for each group of receivers.

Receivers' responses on the attitude questionnaires and on other instruments are statistically analyzed to discover any relationships between each communication and its persuasive impact. It may be found, for example, that intermediate levels of anxiety are optimal, that communications designed to evoke either very high or very low levels of anxiety are least persuasive. Other studies may focus on how persuasion is affected by such aspects as message complexity, receivers' previous tendency to show chronic worry about health, and so on. Aside from their theoretical value, such experiments may provide specific information that can guide the conduct of practical, large-scale campaigns.

Source factors. *Credibility.* There is evidence to suggest that persons who attempt to persuade others are most likely to be believed if they appear to know the truth and to be motivated to reveal it; that is, their credibility is a function of their perceived expertise and trustworthiness. Perceived expertise on the part of a person who is a message source is found to exert major persuasive impact even

when it is not specific to the issue at hand. It can derive from his general educational attainment or occupational status; for example, those members of juries with relatively high levels of education are likely to be elected foremen and to be especially influential in determining verdicts even if they exhibit only the subtlest kinds of cues to their status. At the same time, it appears that people also tend to be particularly influenced by sources who seem most like themselves, perhaps through some perceived identity of interest. Thus, while the source may gain credibility from his apparent superior knowledgeability, he may lose it if he seems too different from his receivers. In such a situation it can be predicted that people are most influenced by those whom they judge to be somewhat more expert than they are but not a great deal more. Indeed, it is found that children are more influenced by those a few years older than themselves than they are by those who are either their age or very much older.

Aside from having to appear to know enough, one's credibility appears as well to depend on his being perceived as highly motivated to tell the truth as he sees it. People will apparently tend to judge him as less trustworthy if he seems intent on persuading them, or likely to gain from their agreement. In general, however, such suspect motivation appears to entail surprisingly little loss of persuasive impact; in some studies the source actually is found to gain in impact from being explicitly depicted as trying to manipulate the recipient. It would seem that while the tactic of clearly revealing the source's persuasive goal may have a detrimental effect on yielding, it apparently clarifies the purpose of the communication thus facilitating attention, comprehension, and retention. Also, when someone is motivated to ingratiate himself with the source (*e.g.*, a student trying to please a teacher), the clearer the persuasive intent is made the more likely he is to go along.

Attractiveness. The attractiveness of the source to the audience apparently adds to his persuasive impact. This may reflect the extent to which the recipient seeks some satisfying relationship with him; for example, if the source is a greatly admired person, people tend to pursue gratification from identifying with him or from gaining social esteem by appearing to be like him. The source's familiarity and similarity to the receiver also enhance his attractiveness and persuasive impact. Ideological similarity seems particularly potent; thus, people who believe in racial segregation tend to prefer those of other races who share this belief over members of their own ethnic group who endorse racial integration.

Power. There are indications that the recipient is influenced by the power he perceives the source to have over him. In such case the receiver may believe that the source actually can control his rewards and punishments. Some theorists have conjectured that power to reward should add more to persuasive impact than does power to punish, but both have proved to be equally efficacious. There are grounds for believing that the source must be perceived as interested in exerting power in order to obtain the recipient's compliance. Seemingly, one is influenced by the members of his own group to the extent that the fate of all depends on each person's compliance to achieve unanimity. Finally, if power is to be effective, the source seemingly must be able to observe whether or not the individual is complying; thus, sources who derive impact from power to reward and punish show more control over the individual's public action than over his private beliefs.

Reward and punishment

Message factors. *Appeals used.* The cancer campaign noted earlier is an example of the use of fear-appeals in persuasive messages. It was hypothesized that maximum persuasion will occur with an intermediate level of fear arousal and that this optimal level is lower when receivers are chronically anxious. Research has given some corroboration to such hypotheses, though the operation of appeals to fear in persuasive communications is proving to be unexpectedly complex.

Students have been interested in the relative persuasiveness of ethical, emotional, and logical appeals as far back as the ancient Greeks and Romans. Empirical research appears to corroborate the classic notion that all three can be persuasive under given circumstances; but the three kinds of appeals are so interrelated that it is unwise to assume that one can produce purely moral or rational or emotional messages for comparative study.

Stylistic factors have been of perennial interest to such practitioners of persuasion as politicians and salesmen. The question of a dynamic versus a subdued presentation, or of the "hard" versus the "soft" approach in advertising (*q.v.*), for example, has been investigated repeatedly in the laboratory and in the field. It is found that the dynamic delivery tends to reduce the degree of yielding but gains in attention and comprehension. The use of humour in the message can enhance yielding; apparently it puts the recipient in a more pleasant, agreeable state. Even the manifestly extraneous pleasure of eating a snack while listening to a message has been found to increase persuasive impact.

Information included. The most studied aspect of what is best included or left out of a persuasive message has to do with the opposition's arguments. A number of studies have been made to determine whether persuasive impact is most enhanced by ignoring counterarguments or by mentioning and refuting them. With an initially hostile or highly intelligent audience, it is found efficacious to mention and refute competing views rather than to ignore them; presumably such an audience is already aware of them. By contrast, when recipients are more sympathetic or less intelligent, it appears to be better to ignore one's opponents; it is likely that members of this audience have not yet thought of other views, and their mention (even in refutation) suggests new difficulties to such people. In general, however, refutation is particularly appropriate when the aim is not only to change opinions but also to make them resistant to subsequent efforts by the opposition.

Refuting the opposition

Another well investigated problem relates to whether it is best explicitly to mention the conclusion desired or whether it should be left to be drawn by the receiver himself. Theoretically, persuasive impact might be enhanced if the receiver participates more fully in the communication process and feels that the conclusion is his own. Research data indicate the opposite, however; messages tend to be much more persuasive when specific conclusions are incorporated. It appears to be the case that receivers in most persuasion situations may lack the ability or motivation to make the intended conclusions for themselves.

Organization or sequence. There are indications that the order in which points in a message are presented can change its persuasiveness. To the extent that a communication is likely to be confusing or the issue unclear, for example, it has been found best to offer the conclusion at the outset; this technique seems to provide receiver orientation that enhances attention and comprehension. When the issue is clear but the audience is resistant, it is demonstrably better to put the conclusion at the end of the message. People seem prone to "tune out" and miss a message that indicates opposition to their present views too early.

In two-sided presentations, as in a debate or a jury trial, participants may argue about which side is to make its presentation first. Experimental data indicate the tentative generalization that whether the first or second side is favoured depends on timing. The greater the intervening time between the messages of the opponents, the more the advantage tends to accrue to the one heard last. As the interval between the last message and the time at which the attitude is measured increases, results tend to benefit the first side.

Channel factors. Channel factors refer to the sensory modality or the medium through which any communication is transmitted.

Sensory modality. The relative efficacy of the written versus the spoken word is a problem that arises in connection with persuasion as well as education. Persuasive messages have in general been found to have about the same impact whether they are presented in written or spoken form, but such difference as there may be is in the direc-

tion of a slightly greater impact for the spoken word. Since attention and comprehension of the message contents are usually greater with written presentations, these results indicate that the comprehension advantage of the written word is overridden by a sizably greater yielding to spoken presentations. It has been suggested that the individual may feel in a closer relationship to the source or under more implicit pressure from him when the source is present to deliver his message orally.

Mass media. Various activities in modern society, such as advertising, public relations, and political campaigning, usually involve the investment of vast resources in the use of the mass media to persuade the public to buy a certain product, to love a certain institution, or to vote for a certain candidate. Studies assessing these campaigns have not been successful in demonstrating their persuasive impact, even as regards attitude change, much less changes in actual purchasing or voting behaviour. Advocates of the mass media have suggested various excuses for this lack of demonstrated impact. Some have argued that they do have an impact but that it results from a "two-step" process in which the mass media affect opinion leaders who in turn affect the masses through face-to-face discussion. Others have argued that the mass media may not be effective in changing opinion but do affect the perceived importance of products, candidates, or issues. Still others have said the media are effective but that alternative products and candidates are advertised about equally so that their effects mutually cancel.

Paradoxically, while the media have difficulty in demonstrating that their persuasive communications have the intended effect, there is some evidence that entertainment programs may have considerable unintended effects. It has, for example, been argued that the high amount of violence depicted in entertainment programs has been responsible for an increased level of violence in society. The arguments here are based on laboratory research showing that nonverbal communication in the form of violent behaviour of models evokes subsequent aggressive behaviour of children. The importance of this phenomenon in mass media presentations in the natural environment has yet to be demonstrated but is currently receiving considerable research attention.

Direct contact. The contact between the recipient and the attitude object in a persuasion situation is only indirect, being mediated through the communication. It thus invites comparison, as regards attitudinal effect, with direct contact between the believer and the object. This topic has been studied most fully in connection with attitudes of racial hostility and with how they are affected by racial integration. An optimistic "to know them is to love them" hypothesis leads to the prediction that when hostile groups of people are brought into direct contact so that they can interact and learn about one another, they should develop more positive mutual attitudes. But the limited results of research suggest that the situation is rather more complex. It may be that there is some small tendency for integration to produce more positive attitudes, but it would also appear that the stronger effect of integration is to multiply pre-existing attitudes so that those with initially favourable attitudes become more favourable while the initially hostile become increasingly hostile. It would seem that man can derive information from interpersonal contacts to bolster whatever initial preconceptions he might have had.

Available research findings appear to specify those conditions under which contact between hostile groups is most likely to result in the development of favourable attitudes. There is, for example, evidence to indicate that this is likely to happen to the extent that the interaction involves two groups on equal status levels, requires the participants to work for a mutually desirable goal, allows them to experience success and discover shared attitudes, and is continued over an extended period of time.

In general, work on channel factors in persuasion has been somewhat neglected over the past several decades. For one thing, the factors involved, while often of practical interest, have little theoretical relevance and so offer little intellectual stimulation to scholars. Also, investigation of these factors often requires research in society at large, while most recent work has been done in the laboratory. It may, however, be the case that the pendulum is swinging back to correct this neglect of channel factors.

Receiver factors. Considerable effort has been extended to discover what kinds of people (receivers) are most readily persuadable.

Demographic characteristics. Among adults, it appears that females generally are more persuadable than males, though the difference is less clear in early childhood years, suggesting that the differences are learned rather than innate. There also seems to be a tendency for the individual's personality characteristics to predict persuadability to a greater extent in males than in females. Perhaps this means that in the social role assigned to him, the man's own proclivities are left to determine his degree of susceptibility while outside factors tend to impose the level of susceptibility in the female regardless of her personal predilections.

Chronological age is another demographic variable that has received considerable attention. Studies indicate that one's persuadability tends to increase during the first five or six years of life, presumably through a growing ability to attend to and to understand the persuasive message, and then perhaps falls slightly before it levels off during the adolescent period for the remainder of the active life span. Persuadability manifests less clear-cut relationships to other demographic variables such as ethnic origin, social class, and regional residence.

Individual differences in ability and personality. When theoretical analysis focusses on the phenomenon of yielding to persuasion, it frequently leads to predictions that the more intelligent person will be more resistant to persuasion (in view of his high critical ability, self-esteem, and willingness to hold deviant attitudes). But when it is considered that he first must understand and remember the persuasive message if he is to be influenced, the more intelligent person may be expected to be more vulnerable to persuasion. Indeed, empirical findings are as complex as these two logically tenable approaches would suggest. In situations that require relatively simple understandings, people with undistinguished intelligence tend to be most responsive to persuasion. When the situation is more demanding of attention and comprehension, high intelligence tends to be associated with persuadability. In general, people at intermediate levels of intelligence appear most persuadable, with those who are very low in intelligence being resistant as a result of poor reception and those of very high intelligence being less likely to yield.

Persuadability of disturbed persons

The relationship of psychiatric disturbance to persuadability is similarly complicated. Those who exhibit minor neurotic signs seem more susceptible to influence in many situations, plausibly as a result of low self-esteem and a greater proneness to yield. Seriously disturbed people (*e.g.*, psychotics) seem to be more difficult to persuade, presumably as a result of their low attention and comprehension; indeed, even with the simple repetitive messages used to induce hypnosis (*q.v.*), schizophrenics are notoriously resistant as compared with more normal people.

In studies of personality characteristics, such as anxiety and self-esteem, people who exhibit moderate degrees of a given trait tend to be more susceptible to influence than are those who are very high or low. The evidence suggests that there may be two opposed underlying factors pushing the relationship in opposite directions. The person who is chronically anxious, for example, seems to have a low self-image that prompts him to yield, yet it appears that his preoccupation with his anxiety tends to interfere with attention, comprehension, and retention. At any rate, the typical research finding is that those who suffer intermediate levels of anxiety are the most vulnerable.

Destination factors. Destination factors refer to the specific, observable activity at which the persuasive communication is aimed—for example, verbalized changes of attitude toward a politician, or overt behaviour in purchasing a particular brand of breakfast food.

Verbalized attitude versus overt behaviour. Research findings with regard to verbalized changes of attitude are

often quite different from those that emerge from studies of gross behavioral response. To some extent this can be understood by assuming greater difficulty in changing behaviour than in changing attitude. Thus, a public-health campaign designed to persuade individuals to give up smoking leads to more who *say* they intend to comply than who actually do. But the disparity is more complicated than that. Often there seems to be more change in gross behaviour than in attitude; among American prisoners of war in North Korea, for example, indoctrination led to substantial alterations in gross behaviour (*e.g.*, participation in propaganda broadcasts), while underlying attitudes apparently were little affected.

It is generally supposed that one's attitudes should be altered if his behaviour is to change. Yet, in other studies it has been found that behaviour often changes first, and attitudes only later. Religious missionaries, for example, may encourage potential converts first to "bend the knee" in the expectation that true faith will follow.

Immediate and delayed impact. It is more convenient to measure immediate changes of attitude than it is to assess long-term effects, although the latter may be of primary interest.

Even as the message tends to be forgotten with the passage of time, there is a general tendency in the long run for a change of attitude produced by persuasive communication to dissipate. The rate of such decay varies widely; depending on conditions, different studies have shown that about half of an originally produced attitude change may decay in anywhere from one week to six months after original exposure to the communication. The rate can be shown to depend on the characteristics of the original message and those of the recipient, as well as other factors.

Delayed-action effects

In many studies of persuasion, a delayed-action effect has seemed to develop over a period of a few days or weeks after receipt of the communication. Delayed-action effects are likely to be observed when the communication initially appears untrustworthy (*e.g.*, is attributed to a suspect source) or when the message is complex or subtle or when its arguments are perceived to hedge to some degree. A major difference between simply forgetting the contents of a message and decay of induced attitude change is the prevalence of delayed-action effects in persuasion.

Inducing resistance to persuasion. Techniques have also been sought by which to make people more resistant to persuasion. Such efforts rest on a most diverse set of approaches, some of the more productive of which are presented below.

Prior commitment. One effective means of making a person more resistant to subsequent persuasive attack appears to be to commit him in advance to his initial belief by having him publicly announce it or simply privately consider it in some formal or effortful way. Indeed, the latter seems to be as effective as the former in making the person resist subsequent attacks on the belief. Commitment becomes even firmer when the person is led to make irreversible decisions and to take action on the basis of his initial belief. With such commitment he tends not only to resist changing his attitude but actively to recruit others in support of his own viewpoint.

Anchoring. A person's belief can apparently be made resistant by anchoring it to other aspects of his life. It may be anchored, for example, to other people or groups who are significant to him, by providing him with firmer information than he already has that they endorse the belief. A belief can be anchored to the individual's values or moral principles by showing him that his adherence to it leads to the attainment of goals he seeks and the avoidance of outcomes that he holds undesirable. Or the belief may be anchored to other beliefs since, as the person becomes more aware of such interrelations, he tends to resist change in any given belief that he perceives would require major readjustments throughout his system of attitudes.

Personality reconstruction. Fundamental alterations in personality might be produced to make one more resistant on all issues. One such approach that has been explored is to increase the general level of hostility, but such efforts have been found to create new problems and do not seem generally effective. Instead, the more hostile person tends to be more resistant to messages that argue for benign and tolerant positions and more receptive to those that argue for punitive stands.

Enhancing the subject's self-esteem

A more attractive approach, which also promises to be more effective, is to raise the person's general level of self-esteem. It has been found that the higher the individual's self-esteem the less prone he will be to pressure. On the other hand, to the extent that enhanced self-esteem also increases attention and comprehension of persuasive messages, this approach could actually add to persuadability. If the person is manipulated in the direction of heightened anxiety, yielding should increase but message reception should decrease. With complex relations such as these, the use of personality restructuring to enhance resistance to persuasion could be hazardous. Under these circumstances it can be difficult to predict whether manipulating a given personality characteristic will increase or decrease persuadability.

Evasion training. People protect themselves against being persuaded by avoiding exposure to positions with which they do not agree or by distorting disagreeable contents or by repressing such dissonant material as they do learn. While these tendencies may seem to be unhealthy, they nevertheless can be used to develop resistance to persuasion. A person can be trained to detect persuasive communications aimed against his belief and quickly to bring into play selective avoidance mechanisms.

Inoculation. Metaphorically akin to smallpox immunization, inoculation against being persuaded has been a heavily studied mode of producing resistance. In this approach a person may be pre-exposed to a weakened dose of attacking arguments sufficient to stimulate his defenses without overcoming them. The method assumes that people typically live in an overprotected ideological environment that leaves belief intensity at a level analogous to the health of a laboratory animal reared in a germ-free environment. Such animals prove most vulnerable when suddenly exposed to germs. By analogy, to protect the individual he is first exposed to weakened arguments that give him motivation and practice in summoning up arguments to bolster his belief in the face of later, massive attack.

Empirical work has shown the approach to be successful. Considerable resistance had indeed been developed through exposure to weakened counterarguments. As in the case of biological inoculation, resistance to subsequent persuasive attack requires time to develop, building up to a maximum over a few days after exposure, after which it decays slowly. The analogy to smallpox holds in other aspects of persuasion as well, including the strength of the pre-exposure and its similarity to the attacking material.

Conclusion. Persuasion has been a major research topic in the social and behavioral sciences during recent years. This research has been directed toward testing old insights accumulated over many centuries of theorizing about the persuasion process and also has revealed new relationships. The number of investigations shows no sign of diminishing, and it is likely that coming decades will show similar progress. Since manipulation of attitudes and behaviour obviously threatens the autonomy of the individual, it is fortunate that study continues to be devoted to ways of producing resistance as well as compliance to persuasive communication.

BIBLIOGRAPHY. A succinct yet comprehensive survey of empirical findings and theories regarding persuasion may be found in W.J. MCGUIRE, "The Nature of Attitudes and Attitude Change," in G. LINDZEY and E. ARONSON (eds.), *Handbook of Social Psychology*, 2nd ed., vol. 3 (1969). Major experimental studies of persuasion are reprinted in R. ROSNOW and E.J. ROBINSON (eds.), *Experiments in Persuasion* (1967); in M. JAHODA and N. WARREN (eds.), *Attitudes* (1966); and in A.G. GREENWALD, T.C. BROCK, and T.M. OSTROM (eds.), *Psychological Foundations of Attitudes* (1968). Theoretical analyses are found in C.A. INSKO, *Theories of Attitude Change* (1967); and in C.A. KIESLER, B.E. COLLINS, and N. MILLER,

Attitude Change (1969). The role of personality factors in persuadability is described in W.J. MCGUIRE, "Personality and Susceptibility to Social Influence," in E.F. BORGATTA and W.W. LAMBERT (eds.), *Handbook of Personality Theory and Research* (1968). Practical applications are discussed in D.J. BEM, *Beliefs, Attitudes, and Human Affairs* (1970); and in P.G. ZIMBARDO and E. EBBESEN, *Influencing Attitudes and Changing Behavior* (1969).

(W.J.McG.)

Peru

A republic bordering the Pacific Ocean, straddling the Andes mountains, and containing the headwaters of the Amazon River, Peru is the third largest nation in South America after Brazil and Argentina. Its name is derived from a Quechua Indian word implying land of abundance, a reference to the economic plenitude produced by the highly organized agricultural technology of the Inca civilization that ruled the region for centuries. The nation's vast natural resources, many of them virtually untapped, long have been exploited by foreign companies and a handful of wealthy Peruvian families. Since 1968, however, a revolutionary government has pressed toward nationalization of industry and of resource development in an attempt to gain greater economic freedom and to improve the living standards and income of the nation's approximately 14,000,000 people.

An overview of the nation

With Ecuador to the north, Bolivia to the southeast, and Chile to the south, all or most of Peru's borders run across the high Andes, whereas the borders with Colombia to the north and Brazil to the east run through lower ranges or tropical forests. In the past, lengthy disputes over certain of these boundaries, some resulting from inexact geographical knowledge, have produced open warfare with Ecuador and Chile, but since a protocol signed in Rio de Janeiro in 1942, only a few landmarks along the Ecuadorian border have remained unsettled. The 496,224 square miles (1,285,216 square kilometres) of Peru, equal to the combined area of Great Britain, France, and Spain, is supplemented by the 238,417 square miles (617,500 square kilometres) of territorial waters, reaching 200 miles (320 kilometres) into the Pacific, that it claims.

Peru is a tropical country, with its northern tip nearly touching the Equator. A great diversity of climate, of way of life, and of economic activity is brought about, however, by the high altitude of much of the country and by the prevailing southwest winds that sweep in from the cold Humboldt, or Peru, Current in the Pacific. The immense difficulties of travel have long impeded national unity. Iquitos on the upper Amazon lies only about 600 miles from Lima, the capital, but before the airplane, travellers between the cities were forced to choose between arduous mountaineering or a 7,000-mile trip that included the Amazon, the Atlantic, the Isthmus of Panama, and the Pacific. The latter was the more frequent route. (For information on related topics, see the articles PERU, HISTORY OF; LIMA; ANDES MOUNTAIN RANGES; and TITICACA, LAKE.)

THE NATURAL AND HUMAN LANDSCAPE

Geographical regions. According to Spanish tradition, Peru comprises three major areas: a narrow belt of arid desert along most of the Pacific; the steep and complex Andes system, or highlands; and the wet, densely forested Montaña, or selva, on the lower eastern slopes of the Andes. The divisions made by the pre-Hispanic peoples, however, are both more complex and in greater agreement with modern geographical principles.

The Chala region, or coastal plain. Bordering the sea and rising to about 1,500 feet (500 metres) above sea level, the Chala is a region of arid plains formed by marine abrasion or fluvial action on old mountain systems and fractured or eroded sedimentary rock. Although it receives less rain than the Sahara, it is the population and economic centre of Peru. Every few decades, the Chala's delicate balance of wind, water, and plant and animal life is disastrously upset by the inundation of the desert coast in great quantities of rain. This sudden visitation, called El Niño (The Child), generally occurs in December. The resulting floods devastate both the natural and human landscapes; the fish leave the coast for months, damaging the fishing industry, and the birds, lacking their customary food, die, littering the beaches by the thousands. The plains are irrigated at intervals by irregularly flowing rivers that produce cultivable oases and woods. In northern and central areas, sandy deserts predominate, whereas rough hills may reach to the ocean in the south.

Coastal climate

The Yunga regions: the Inter-Andean Valley. A maritime Yunga ("jungle valley") lies on the western slope of the Andes, rising from about 1,000 to 7,500 feet above sea level, whereas the fluvial Yunga on the eastern slopes lies at altitudes of about 3,200 to 7,500 feet. These are regions of narrow, enclosed valleys at the bottom of rocky canyons. A hot climate produces a high incidence of yaws, malaria, and goitre among the scanty population. Despite frequent downpours and accompanying mud avalanches, water for agriculture is scarce.

The Quechua, or temperate, region. The principal cities of the Inca Empire—on or near the sites of such present-day cities as Cuzco, Cajamarca, Tarma, and Jauja—were built in this rolling landscape, which is from 7,500 to 11,500 feet in altitude. The region is heavily populated because of the moderate climate, and the rich volcanic soils are cultivable for vegetable and tree-fruit crops.

The Suni, or high-barren, region. The high, rough terrain of the Suni rises from about 11,500 to 13,000 feet and has numerous cliffs, escarpments, and other natural obstacles. The climate is cold, frosts frequent, and soils thin, limiting crops to a few leafy and tuberous varieties. Species of eucalyptus are among the several trees adapted to life in this environment.

The Puna region: the Altiplano, or High Andes. Rising from about 13,000 to 15,700 feet, the Puna is characterized by vast tablelands open to blazing sun and constant winds. The cold climate has wide daily fluctuations in temperature, but varieties of potato and fruit can be grown to about 14,000 feet, the upper limit of human habitation in Peru. Lacking in natural forestation, it is a world of forage grasses for the llama, alpaca, vicuña, and guanaco.

The Janca, or high cordilleran, region. Above 15,700 feet, at the nucleus of the Andean system, the Janca comprises a succession of craggy and snow-covered volcanic cones or shafts. Nighttime temperatures are below freezing throughout the year, and no spontaneous vegetation occurs. Human settlement is ephemeral.

The Rupa-Rupa, or high-forest, region. Descending from the eastern fluvial Yunga to about 1,300 feet, the Rupa-Rupa comprises a succession of valleys, ridges, and plains related to Andean and eastern Peruvian earth-fault systems. Its climate is warm, with cool evenings, and its rough and rolling relief is covered with trees and pastures that are kept green throughout the year by heavy annual precipitation of more than 157 inches. Livestock raising and plantation agriculture are widespread, and large forest reserves contain a great variety of woods.

The eastern forests

The Omagua, Amazonian, or low-forest region. The Amazon River and its affluents wander through the huge plain of the Omagua, creating meanders, lakes, marshes, and other formations. Soils are rich in alluvial and organic deposits. The climate is warm, the precipitation more than 118 inches, and flooding is frequent. In its woods and waters live thousands of tropical, terrestrial, and aquatic species, including colourful ornamental fish. The heavy forestation is commercially important for its rubber, chicle, and rosewood, and a rudimentary agriculture is pursued by primitive tribes on the shifting stream banks and higher locations.

Drainage. Fifty-four rivers empty into the Pacific from the Andes. Of these, only the Río Tumbes is navigable. Only 20 percent of the flow is utilized by agriculture. The Amazon, with the largest volume flow of any river in the world, rises in Peru as the Río Ucayali. It is navigable, as are its seven principal affluents. In the south, four rivers cross the Collao Plateau to empty into Lake Titica-

ca, which is shared with Bolivia and is—at 12,500 feet (3,810 metres) in altitude and 3,205 square miles (8,300 square kilometres) in size—the world's highest navigable body of water.

Patterns of human settlement. The nature of Peruvian life, whether urban or rural, varies by physiographic region. The greater degree of commonality exists, however, in relation to population densities.

Rural Peru. In the coastal desert, populated areas along wagon roads and bridle paths usually have a single, unpaved street. Their one-story houses are built of such light materials as *quincha*, a sticks-and-clay compound, dictated by the danger of earthquakes. The settlements lack services; in some places, the houses have wells and small groves of tropical and subtropical fruit trees.

Going upward, settlements are located at distances of a day's foot or horseback journey—about nine miles—taking advantage of crossroads, river mouths, or high ground. In the warm Yungas, houses generally have one story, the front of which faces the interior of the property. They are separated from one another by groves of fruit trees. Kitchens and bread ovens are set apart. Roofless, circular enclosures sometimes guard the livestock. In the temperate Quechua and the cold Suni regions houses have walls of solid adobe, of adobe bricks, or of stones mortared with clay. Of the usual two stories, the first is for living, the second for storing crops. Corrals hold the livestock and sheds the kindling, whereas kitchens have special devices for fattening guinea pigs. In the very cold Punas, houses are either made similarly or are of straw or sod; kitchens are attached to the house to provide heat. The small crop fields are protected by enclosures, or by stone walls, as are the pens for llamas, alpacas, and other animals.

In both the high and low eastern forests, houses are placed among the cultivated fields close to communications routes and to sources of drinking water—although water may be captured from zinc roofs. The houses are built with bases of termite- and fungi-resistant woods and with palm-leaf roofs that are replaced frequently. In flood areas, houses are raised on supports. Kitchens are always at a distance to avoid fires. Rafts and canoes are moored on the banks of navigable lakes and rivers.

In all of the regions, the old landed estates, or haciendas, are characterized by their large buildings, solid and well protected. Construction varies with the particular activity of the hacienda. There are no logging haciendas, but establishments for fishing have been organized recently.

Urban Peru. Lima (*q.v.*), with more than 3,000,000 inhabitants, is Peru's administrative, commercial, and cultural centre.

The Peruvian cities

Since World War II, glass and steel have come to form striking contrasts with the sprawling slums, or *barriadas*, of the Indians who have left their mountains to seek a better life.

Arequipa, about 60 miles east of the southern coast, lies at the foot of El Misti volcano, presumably inactive. Originally an Inca city, Arequipa is sometimes called the White City because it is constructed almost entirely of white volcanic rock. The principal population centre of the northern Andes is Cajamarca, a sleepy town where an Inca king gave the Spanish conquistador Francisco Pizarro piles of gold and silver for his freedom but was then executed. To the south, Cerro de Pasco, the nation's mining centre, is, at 14,300 feet, one of the world's highest cities. Huancayo, a farming centre, is famous for its colourful Sunday fairs, during which Indians from the countryside sell a variety of products, including figurines, llama-wool blankets, and the high-domed hats worn by Indian women everywhere in the Andes.

The principal Andean centre, however, is the ancient city of Cuzco, once the capital of the Inca Empire. Tourists from around the world visit Inca remains in Cuzco and environs as well as the many Spanish colonial churches, often regarded as the finest in the Western Hemisphere. The Inca past is apparent in many places. Inca walls topped by Spanish-style houses stand along several streets. The base of the Temple of the Sun, portions of which were plated in gold, was used as the foundation for the Church of Santo Domingo, which stands

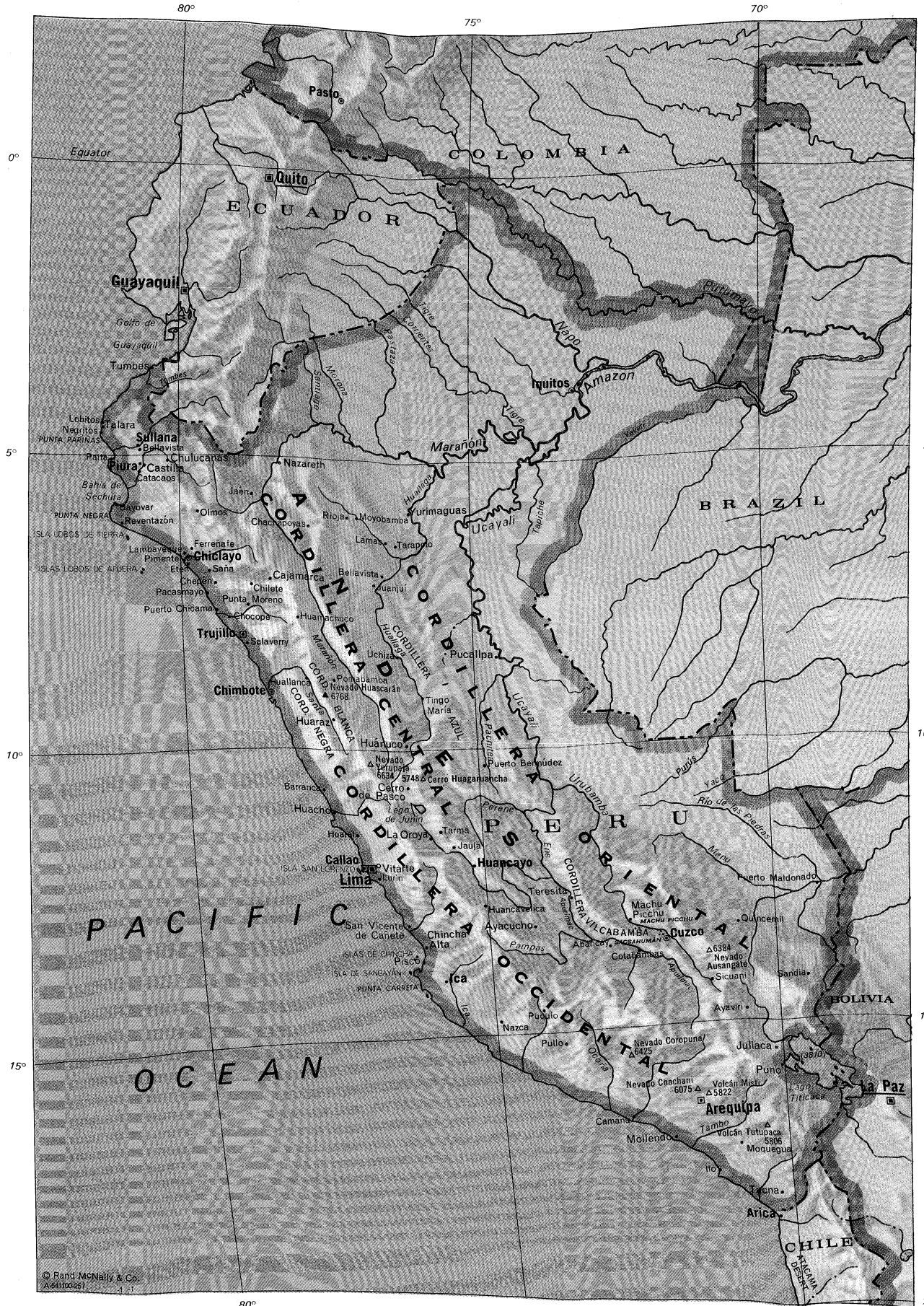
MAP INDEX

Cities and towns

Abancay 13·35s 72·55w
Arequipa 16·24s 71·33w
Ayacucho 13·07s 74·13w
Ayaviri 14·52s 70·35w
Barranca 10·42s 77·45w
Bayovar 5·50s 81·03w
Bellavista 4·54s 80·42w
Bellavista 6·58s 76·32w
Cajamarca 7·10s 78·31w
Callao 12·02s 77·05w
Camaná 16·36s 72·40w
Cañete, see San Vicente de Cañete
Castilla 5·12s 80·38w
Catacaos 5·15s 80·42w
Cerro de Pasco 10·41s 76·16w
Chachapoyas 6·10s 77·50w
Chepén 7·15s 79·25w
Chiclayo 6·46s 79·50w
Chilete 7·14s 78·51w
Chimbote 9·05s 78·36w
Chincha Alta 13·27s 76·08w
Chocope 7·47s 79·12w
Chulucanas 5·06s 80·10w
Cotabambas 13·45s 72·20w
Cuzco 13·31s 71·59w
Eten 6·55s 79·50w
Ferreñafe 6·35s 79·45w
Huacho 11·07s 77·37w
Huallanca 8·50s 77·50w
Huamachuco 7·50s 78·05w
Huancavelica 12·46s 75·02w
Huancayo 12·04s 75·14w
Huánuco 9·55s 76·14w
Huaral 11·32s 77·13w
Huaraz 9·32s 77·32w
Ica 14·04s 75·42w
Ilo 17·38s 71·20w
Iquitos 3·50s 73·15w
Jaén 5·21s 78·28w
Jauja 11·46s 75·28w
Juanjuí 7·10s 76·45w
Juliaca 15·30s 70·08w
Lamas 6·25s 76·30w
Lambayeque 6·42s 79·55w
La Oroya 11·32s 75·54w
Lima 12·03s 77·03w
Lobitos 4·27s 81·15w
Lurín 12·17s 76·52w
Machu Picchu 13·07s 72·34w
Mollendo 17·02s 72·01w
Moquegua 17·20s 70·55w
Moyobamba 6·02s 76·58w
Nazareth 5·08s 78·20w
Nazca 14·50s 74·55w
Negritos 4·36s 81·15w
Olmos 5·59s 79·46w
Pacasmayo 7·20s 79·35w
Paita 5·05s 81·10w
Pimentel 6·45s 79·55w
Pisco 13·42s 76·13w
Piura 5·12s 80·38w
Pomabamba 8·50s 77·25w
Pucallpa 8·20s 74·30w
Puerto Bermúdez 10·17s 74·57w
Puerto Chicama 7·42s 79·27w
Puerto Maldonado 12·36s 69·11w
Pullo 15·10s 73·48w
Puno 15·50s 70·02w
Punta Moreno 7·36s 78·54w
Puquio 14·43s 74·17w
Quincemil 13·15s 70·40w
Reventazón 6·10s 80·58w
Rioja 6·03s 77·05w
Salaverry 8·14s 78·58w
Saña 6·54s 79·36w
Sandia 14·14s 69·25w
San Vicente de Cañete 13·08s 76·28w
Sicuani 14·15s 71·15w
Sullana 4·53s 80·42w
Tacna 18·01s 70·15w
Talara 4·35s 81·25w
Tarapoto 6·30s 76·20w
Tarma 11·25s 75·43w
Teresita 12·33s 73·50w
Tingo María 9·09s 75·56w
Trujillo 8·10s 79·02w
Uchiza 8·25s 76·20w
Vitarte 12·02s 76·54w
Yurimaguas 5·54s 76·05w

Physical features and points of interest

Amazon, *river* 4·21s 70·02w
Andes, *mountains* 8·00s 77·00w
Apurímac, *river* 12·17s 73·56w
Ausangate, Nevado, *mountain* 13·47s 71·13w
Azul, Cordillera, *mountains* 9·00s 76·00w
Blanca, Cordillera, *mountains* 9·00s 77·30w
Carreta, Punta, *point* 14·12s 76·17w
Central, Cordillera, *mountains* 0·09s 77·00w
Chachani, Nevado, *mountain* 16·12s 71·32w
Chincha, Islas de, *islands* 13·39s 76·24w
Coropuna, Nevado, *mountain* 15·30s 72·41w
Corrientes, *river* 3·43s 74·35w
Ene, *river* 11·10s 74·18w
Huagaruancha, Cerro, *mountain* 10·32s 75·56w
Huallaga, *river* 5·07s 75·30w
Huascarán, Nevado, *mountain* 9·07s 77·37w
Ica, *river* 14·53s 75·33w
Junín, Lago de, *lake* 11·00s 76·09w
Las Piedras, Río de, *river* 12·37s 69·10w
Lobos de Afuera, Islas, *islands* 6·57s 80·42w
Lobos de Tierra, Isla, *island* 6·27s 80·52w
Machu Picchu, *ruins* 13·08s 72·30w
Manu, *river* 12·16s 70·55w
Marañón, *river* 4·30s 73·35w
Misti, Volcán, *volcano* 16·17s 71·24w
Morona, *river* 4·40s 77·10w
Negra, Cordillera, *mountains* 9·30s 77·30w
Negra, Punta, *point* 6·06s 81·10w
Napo, *river* 3·20s 72·40w
Occidental, Cordillera, *mountains* 13·00s 75·00w
Ocoña, *river* 16·28s 73·08w
Oriental, Cordillera, *mountains* 13·00s 72·00w
Pachitea, *river* 8·46s 74·33w
Pacific Ocean 15·00s 80·00w
Pampas, *river* 13·24s 73·12w
Pariñas, Punta, *point* 4·40s 81·20w
Pastaza, *river* 4·50s 76·25w
Perene, *river* 11·09s 74·14w
Purús, *river* 9·25s 70·35w
Putumayo, *river* 2·40s 70·05w
Sacsahumán, *ruins* 13·30s 71·59w
Sangayán, Isla de, *island* 13·51s 76·28w
San Lorenzo, Isla, *island* 12·06s 77·13w
Santa, *river* 8·58s 78·39w
Santiago, *river* 4·27s 77·38w
Sechura, Bahía de, *bay* 5·35s 81·00w
Tambo, *river* 17·10s 71·52w
Tapiche, *river* 5·03s 73·51w
Tigre, *river* 4·30s 74·10w
Titicaca, Lago, *lake* 15·50s 69·20w
Tumbes, *river* 3·30s 80·27w
Tutupaca, Volcan, *volcano* 17·02s 70·22w
Ucayali, *river* 4·30s 73·30w
Urubamba, *river* 10·43s 73·48w
Vilcabamba, Cordillera, *mountains* 13·00s 73·00w
Yaco, *river* 10·45s 10·40w
Yavarí, *river* 4·21s 70·02w
Yerupajá, Nevado, *mountain* 10·16s 76·54w


80°
75°
70°
0°
5°
10°
15°
20°
Equator
COLOMBIA
ECUADOR
PERU
BRAZIL
BOLIVIA
CHILE
PACIFIC OCEAN
Pasto
Quito
Guayaquil
Golfo de Guayaquil
Tumbes
Lobitos
Negritos
PUNTA PARIÑAS
Talara
Sullana
Bellavista
Paita
Piura
Castilla
Catacaos
Chulucanas
Bahía de Sechura
Sechura
Nazareth
Jaén
Bayovar
PUNTA NEGRA
Reventazón
ISLA LOBOS DE TIERRA
Olmos
Chachapoyas
Rioja
Moyobamba
Lamas
Tarapoto
Lambayeque
Ferreñafe
Pimentel
Chiclayo
ISLAS LOBOS DE AFUERA
Eten
Saña
Chepén
Cajamarca
Bellavista
Juanjui
Pacasmayo
Chilete
Punta Moreno
Puerto Chicama
Chocope
Huamachuco
Trujillo
Salaverry
Uchiza
Pucallpa
Chimbote
Huallanca
Pomabamba
Nevado Huascarán 6768
Tingo María
Huaraz
Huánuco
Nevado Yerupajá 6634
Cerro Huagaruancha 5748
Puerto Bermúdez
Barranca
Cerro de Pasco
Huacho
Lago de Junín
Huaral
La Oroya
Tarma
Jauja
Callao
ISLA SAN LORENZO
Lima
Vitarte
Lurín
Huancayo
Teresita
Huancavelica
Ayacucho
Chincha Alta
San Vicente de Cañete
ISLAS DE CHINCHA
Pisco
ISLA DE SANGAYAN
Ica
PUNTA CARRETA
Pampas
Abancay
Machu Picchu
MACHU PICCHU
MACHUPICCHU
Cuzco
Cotabambas
Quincemil
Nevado Ausangate 6384
Sicuani
Sandia
Ayaviri
Puquio
Nazca
Pullo
Nevado Coropuna 6425
Juliaca
Puno
Nevado Chachani 6075
Volcán Misti 5822
Arequipa
La Paz
Lago Titicaca
Camana
Mollendo
Tambo
Volcán Tutupaca 5806
Moquegua
Ilo
Tacna
Arica
ATACAMA DESERT
Río Marañón
Río Huallaga
Río Ucayali
Río Pachitea
Río Perené
Río Apurímac
Río Urubamba
Río Pachitea
Río Pisqui
Río Yavarí
Río Napo
Río Tigre
Río Corrientes
Río Pastaza
Río Huallaga
Río Tapiche
Río Ucayali
Río Purús
Río Yacu
Río de las Piedras
Río Manu
Río Acre
Río Amazon
CORDILLERA OCCIDENTAL
CORDILLERA BLANCA
CORDILLERA NEGRA
CORDILLERA CENTRAL
CORDILLERA ORIENTAL
CORDILLERA VILCABAMBA
Iquitos
Yurimaguas
Urubamba
Ucayali
Marañón
Huallaga
Tingo María
Huánuco
Huancayo
Cuzco
Arequipa
Puno
La Paz
Iquitos
Pucallpa
Juliaca
Ayacucho
Huancavelica
Ica
Nazca
Pullo
Puquio
Abancay
Machu Picchu
Quincemil
Puerto Maldonado
Sandia
Sicuani
Cotabambas
Nevado Ausangate 6384
Nevado Coropuna 6425
Nevado Chachani 6075
Volcán Misti 5822
Volcán Tutupaca 5806
Moquegua
Tacna
Arica
Mollendo
Camana
Tambo
Ayaviri
Juliaca
Puno
Lago Titicaca
Teresita
Pisco
Chincha Alta
San Vicente de Cañete
Callao
Lima
Vitarte
Lurín
Barranca
Huacho
Huaral
La Oroya
Tarma
Jauja
Cerro de Pasco
Puerto Bermúdez
Pucallpa
Yurimaguas
Iquitos
Nazareth
Chulucanas
Sullana
Talara
Tumbes
Guayaquil
Quito
Pasto
Chiclayo
Trujillo
Chimbote
Huaraz
Cajamarca
Chachapoyas
Moyobamba
Rioja
Juanjui
Tarapoto
Lamas
Olmos
Jaén
Piura
Castilla
Catacaos
Paita
Bellavista
Salaverry
Chocope
Huamachuco
Punta Moreno
Chilete
Pacasmayo
Chepén
Saña
Eten
Pimentel
Ferreñafe
Lambayeque
Reventazón
Bayovar
Sechura
Bahía de Sechura
Golfo de Guayaquil
Tumbes
Lobitos
Negritos
PUNTA PARIÑAS
PUNTA NEGRA
ISLA LOBOS DE TIERRA
ISLAS LOBOS DE AFUERA
ISLA SAN LORENZO
ISLAS DE CHINCHA
ISLA DE SANGAYAN
PUNTA CARRETA
Río Marañón
Río Huallaga
Río Ucayali
Río Pachitea
Río Perené
Río Apurímac
Río Urubamba
Río Pisqui
Río Yavarí
Río Napo
Río Tigre
Río Corrientes
Río Pastaza
Río Tapiche
Río Purús
Río Yacu
Río de las Piedras
Río Manu
Río Acre
Río Amazon
Lago de Junín
Lago Titicaca
CORDILLERA OCCIDENTAL
CORDILLERA BLANCA
CORDILLERA NEGRA
CORDILLERA CENTRAL
CORDILLERA ORIENTAL
CORDILLERA VILCABAMBA
Pampas
Abancay
Machu Picchu
MACHU PICCHU
MACHUPICCHU
Cuzco
Cotabambas
Quincemil
Nevado Ausangate 6384
Sicuani
Sandia
Ayaviri
Puquio
Nazca
Pullo
Nevado Coropuna 6425
Juliaca
Puno
Nevado Chachani 6075
Volcán Misti 5822
Arequipa
La Paz
Camana
Mollendo
Tambo
Volcán Tutupaca 5806
Moquegua
Ilo
Tacna
Arica
ATACAMA DESERT
© Rand McNally & Co.
A-641100-251
PERU
Size of symbol indicates relative size of town
Elevations in metres
0 100 200 400 600 km
0 100 200 400 mi

on the main plaza. The most monumental Inca ruins are those of the fortress of Sacsahuamán, consisting of a gigantic three-tiered stone wall built of monoliths as large as 18 feet in length and weighing 300 tons.

The bygone world of Spanish colonial power is evident in the tile-roofed buildings, museums, and churches of Cuzco. Among the most impressive is the cathedral (*c.* 1550), the only building that survived the earthquake of 1650. Other churches are celebrated for their 17th-century Baroque facades, in the mestizo style, which is an exuberant mixture of Indian and European motifs.

The major city of eastern Peru is Iquitos, which was a small jungle outpost on the upper Amazon until the rubber boom of the 1880s. Since the boom ended, lumber has been the major product. Among other major centres are Talara, a city amid the northern desert serving the petroleum industry; and Ica, another desert enclave and a viticulture centre. Huánuco is a university city at the gateway to the forests, whereas Ayacucho is famous for its temples and colonial mansions.

THE PEOPLES OF PERU

Pre-Hispanic groups. Diverse groups of indigenous Indians occupied the various natural regions during the pre-Hispanic millennium. The inhabitants of the coast were and are still called Chalacos. Farmers of the oases and of the deserts and skilled fishermen, they spoke languages, such as Sec, Quignam, and Pescadora, that today have disappeared. The Yunkachus, or Mapayungas, who inhabited the Yungas, were fruit and coca growers and spoke a variety of languages. Some were identical to the languages of the coast, for which reason all acquired the generic designation Yunga languages. The Quechua of the temperate highlands were expert agriculturalists who bestowed their name on a race, a culture, and a language. The Quechua spoke, and still speak, the Runa-Shimi language. In the Suni lived the Shucuy, a guinea-pig raiser, forester, and dehydrator of food products, who spoke Runa-Shimi and Cauqui. The Jatire, or Jamilli, of the Puna was a herder and grower who still speaks Runa-Shimi and Aymara. The Chuncho, inhabiting the high forests, was a hunter and gatherer; his numerous languages and dialects evidence a diverse linguistic heritage. In the low-forest region lived the Omagua, navigator of the big rivers, fisherman, gatherer, and eventual farmer. He spoke and still speaks languages related to those of the Chunchos.

The Peruvian Indians

Development of ethnic diversity. The Spanish conquerors settled in the coastal deserts and oases and in the cities that they established in the cold and temperate zones on the ruins of the cities that formerly had belonged to the indigenous peoples—*e.g.*, Cuzco, Cajamarca, Arequipa, Huaraz, Huánuco, Tarma, Huancavelica, and Ayacucho. They were a prototypical dominating group, masters in every area: politics, religion, and economics. They brought their European culture and transmitted their white skin, Spanish language, and Catholic religion to their descendants, who have become greatly mixed, however. Although they have received infusions from new groups of Europeans, there is currently no hereditarily pure white group in the Spanish tradition.

With the Spaniards, blacks came as slaves. Following the revolution and prohibition of slavery, Chinese were imported as farm labourers, and new groups of Spaniards, Nordic whites, and Japanese arrived. A complete absence of racial prejudice has permitted these diverse racial and ethnic groups to intermarry and produce various new racial types: mestizo, the mixture of white with Indian; mulatto, white with black; injerto, white with Chinese or Japanese; zambo, black with Indian; and chinocholo, black with Chinese or Japanese. More varied mixtures in succeeding generations have led to numerous additional distinctions. Joined together in the Peru of today, they comprise a complex mosaic of racial intermingling.

Differences in life-styles and attitudes are pronounced. Peruvians of Spanish descent and the mestizos live mainly along the coast and own the lion's share of the nation's wealth. The Indians of the mountains make up half of the population and live in extreme poverty in one of the world's harshest environments. Despite frequent attempts to improve their lot, they remain both indifferent to, and outside of, the large affairs of the country.

Contrasts of life-styles

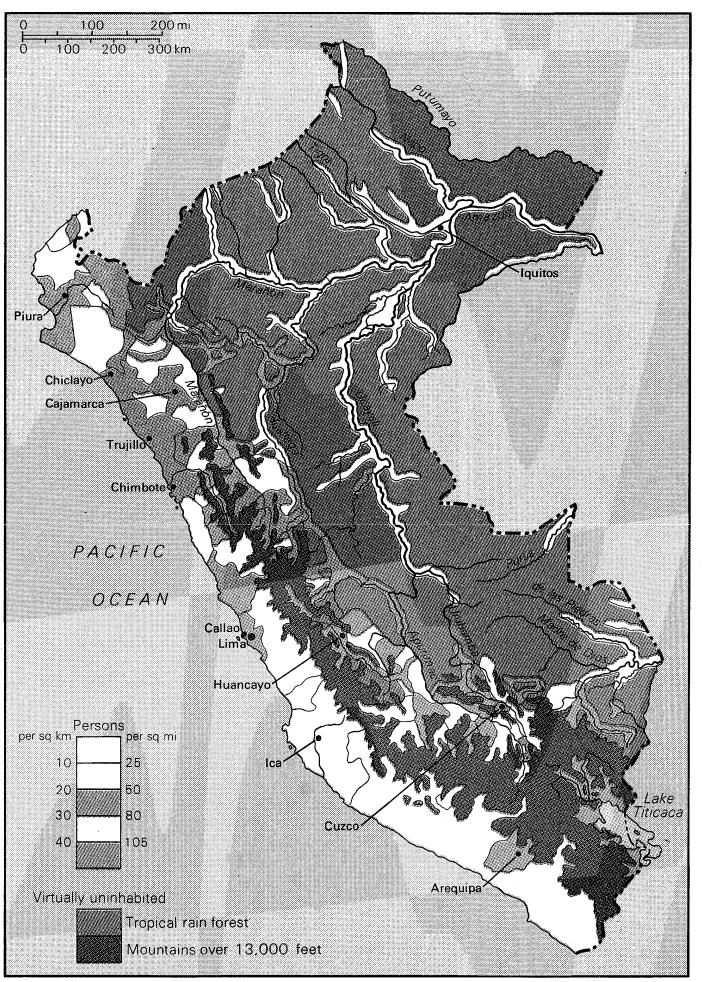
Typically, the small group of white people who are part of the ruling class live in urban homes or rural haciendas, and share a widespread belief in the Indians as shiftless and lazy. Mestizo culture is a blend of Indian and European ways known as *criollo*.

The Indian population, about 70 percent of the mountain people, has been a submerged element since the Spanish conquest. Living apart from the mestizos and Europeans, they shepherd their llama herds and work tiny plots of land to eke out a subsistence living. On the coast, they lived under near-feudal conditions, as tenant farmers on the large estates or as day labourers and sharecroppers. Like their Inca ancestors, the overwhelming number of Indians read neither their own nor another language.

Religion. Ancient Peru had various fetishistic, polytheistic, and pantheistic religions. The most important god was Viracocha—lord, creator, and father of men. The sun, the lightning, and the moon were worshipped. Each traditional culture raised great temples to honour its local divinity.

The Spanish indoctrinated the Indians and spread the Catholic religion, built hundreds of sumptuous churches, and held fiestas for patron saints in every village. The people were not strict in their practices, however, and Protestant sects have proliferated. The Indians have adapted many Catholic events in the light of the older practices to produce a rich diversity in religious practice.

Demography. Carbon-dated remains found in Ayacucho place the length of man's stay in Peru at 18,000 years. Under the Incas, the population reached 12,000,000, but in the first 50 years of Spanish domination, nearly



Population density of Peru.

10,000,000 Indians perished. The 1793 census showed 1,076,000 inhabitants; the 1940 census, more than 7,000,000; that for 1961, more than 9,900,000. During the five-year period 1965–70, the number of live births per 1,000 inhabitants was 43.5, the death rate, 12.3, giving a rate of natural increase of 31.2 and an estimated population in the early 1970s of about 14,000,000. These figures include an estimated 100,000 persons of the Amazon jungles living in Stone Age conditions.

The population structure is pyramidal. In 1966, about 55 percent of Peruvians were 19 years of age and under, whereas those over 60 years of age represented only 5 percent. Only about 31 percent were in an economically active age bracket in 1970. Between 1940 and 1971 there was a great increase in the urban population to more than 50 percent.

The population of Peru is doubling by natural increase every 25 years. Peru is considered large and empty, though new agricultural lands are being opened by irrigation works and by the building of roads to the eastern forests. The population is moving down from the highlands to the lowlands adjacent to the sea and to the Amazonian rivers, and the urban areas are growing while the cold countryside is gradually being depopulated.

Peru, Area and Population

	area		population	
	sq mi	sq km	1961 census	1972 census‡
Constitutional Province (*provincia constitucional*)				
Callao	29	74	214,000	316,000
Departments (*departamentos*)				
Amazonas	15,945	41,297	118,000	196,000
Ancash	14,019	36,308	583,000	727,000
Apurímac	7,975	20,655	288,000	308,000
Arequipa	24,528	63,528	389,000	531,000
Ayacucho	17,058	44,181	411,000	460,000
Cajamarca	13,675	35,418	747,000	916,000
Cuzco	29,431	76,225	612,000	709,000
Huancavelica	8,139	21,079	303,000	331,000
Huánuco	13,635	35,315	329,000	421,000
Ica	8,205	21,251	256,000	358,000
Junín	16,751	43,384	521,000	691,000
La Libertad	8,973	23,241	582,000	806,000
Lambayeque	6,404	16,586	342,000	515,000
Lima	13,087	33,895	2,031,000	3,485,000
Loreto	184,686	478,336	337,000	495,000
Madre de Dios	30,271	78,403	15,000	22,000
Moquegua	6,245	16,175	52,000	75,000
Pasco	8,438	21,854	138,000	177,000
Piura	12,767	33,067	669,000	855,000
Puno	27,947	72,382	686,000	780,000
San Martín	20,488	53,064	162,000	224,000
Tacna	5,701	14,767	66,000	96,000
Tumbes	1,827	4,732	56,000	75,000
Total Peru	496,224	1,285,216*	9,907,000†	13,568,000*

*Figures do not add to total given because of rounding. †Excludes an estimated 101,000 Indian jungle population. ‡Preliminary.
Source: Official government figures.

THE NATIONAL ECONOMY

Economic overview

Peru is the world's leading fishing country, the prime exporter of fish meal, the foremost producer of bismuth, the third largest producer of silver, the fifth largest copper producer, and the sole producer of alpaca wool and of the fertilizer known as guano. It contributes the world's finest wool, vicuña, and the natural food most complete in nutrients, quinoa, a leafy vegetable eaten like spinach.

Serious economic problems persist, however, in several areas. A shortage of investment capital is paralleled by a loss of university-trained technicians to offers of higher salaries abroad. Extensive destruction of highways and other physical properties occurs each year from earthquakes, slides, and other geological disasters. The limited agricultural areas are failing to meet the needs of a rapidly expanding population, resulting in continually rising imports and in attempts to alter both the nation's planting and eating habits. The nationalization of the petroleum and other industries since 1968 has been directed toward remedying these and other economic deficiencies.

Economic resources and utilization. *Mining.* Copper, iron, phosphorus, and manganese exist in huge quantities of high-yield ores. Gold and silver are found extensively, as are other rare metals. In all, more than 40 metallic elements and compounds are commercially exploitable. Petroleum and natural-gas fields of the northern desert and continental shelf supply fuel along with the huge carboniferous belts producing peats and various coals and the great stores of sugar-cane waste, or bagasse.

In spite of this potential wealth, exploitation has varied greatly with foreign investment, price fluctuations on the world markets, lack of roads, a scarcity of adequate installations, and, in some cases, the exhaustion of deposits. Modern mining legislation and the intensification of mining operations are based on the expectation of greater importance in the world market. An investment of more than $1,225,000,000 has been calculated for the period 1968–81. Most of this is designated to achieve a greater copper and phosphate production.

Biological resources. Between 1968 and 1970, farmlands increased nearly 20 percent, to about 7,000,000 acres (3,000,000 hectares). The principal crops are cotton, sugarcane, rice, cereals, coffee, tea, and fruits. The decline in the production of guano, an organic fertilizer of unique quality, has meant a serious reverse for agriculture, for it has not yet been adequately replaced by the chemical fertilizers.

Agriculture and fishing

An insufficient area under cultivation has caused increasing importation of foodstuffs, which now account for nearly 20 percent of total imports. Peru is in the process of changing its national diet away from the former staples—corn, potatoes, cassava, sweet potatoes, quinoa, and llama—but as yet it must import much of the wheat, beef and other livestock, and dairy products. Meeting these new requirements has become Peru's most urgent problem.

The fishing industry, developed within recent decades, is based on the enormous schools of anchovy. Prudence limits the catch, however, to about 13,000,000 tons per year. Current technology does not permit total recovery of all the oil extracted during fish-meal production, nor is it possible to take due advantage of full nutritional potentiality in the catch. Fishing for human consumption is still insignificant, and the country is in the anomalous position of exporting proteins via the fish meal while lacking them in the popular diet.

Many of Peru's hundreds of land and water animals have disappeared, including the chinchilla, the puna ostrich, and the mountain turkey; others are in danger, such as the guanay, whose droppings are the valuable guano; the fur-bearing vicuña and guanaco; and the pirarucu, an important food fish. Recent conservationist measures prohibit the hunting of about 50 species and fix guidelines for the export of such wild animals as ornamental fish, monkeys, crocodiles, iguanas, otters (nutria), jaguars, and the manatee or sea cow. Strict, closed seasons are observed in the fishing of anchovy, river prawn, and trout. The hunting of vicuña and chinchilla, whose repopulation is being attempted, is prohibited. Artificial hatcheries for river prawn and *paiche* have been built.

Flora and fauna

More than half of Peru is covered with natural hardwood and softwood forests. Species most worked commercially include cedar, mahogany, and walnut. Paper production from softwoods is relatively recent, and newly introduced techniques in pressed wood and veneering have given rise to an important industry.

Electrical power. The hydroelectric potentiality in Peru is great, especially on the Amazon and its tributaries; but until 1968, only 2 percent of this had been harnessed. Several plants were projected for the early 1970s, and it was anticipated that interconnections could be affected between the several operating plants, both thermal and hydroelectric. Nearly half of the nation's production and demand are in Lima Department, though mining activities are the biggest single user of power. In the early 1970s, public and private ownership remained about equal.

Manufacturing. Government plans for the period 1970–75 call for industry to increase its share of gross

national product from 20.7 to 26.5 percent. To utilize the nation's natural resources to achieve self-sustained growth, a particular push was given to such basic industries as chemical and steel refining and production. Companies remaining in private hands were to pass gradually into the public sector, as were the housing, educational, clothing, and food industries. The goal of various new laws was to make industrial profits benefit the masses.

Foreign commerce. During 1968, Peru had a favourable balance of trade amounting to $240,000,000. A decline in imports was due to a monetary devaluation in 1967, to a customs surcharge, and to import restrictions. Exports increased due to greater sales of fish meal, copper, silver, sugar, and coffee. More recently, lower sales of fishmeal and a drop in mineral prices, aggravated by strikes among the mine workers, have injured foreign commerce. The availability of dollars has decreased, but large loans as well as purchases and investments have been obtained from socialist countries.

Economic administration. *Current policies.* The financing of investment in works and studies for 1970 amounted to nearly 10,000,000,000 soles (38.70 soles = $1 U.S.; 93.27 soles = £1 U.K. on May 1, 1971). Seventy percent of these sums was covered by the treasury and its own resources, the remainder by loans from foreign and international monetary agencies. The government's goal has been to raise the gross national product by a higher figure than was achieved in 1970 and to lower drastically the unemployment and underemployment figures. Heavy investments were made in fishing research and in state enterprises in iron and steel, electricity, and mining. Construction was begun on the copper and zinc refineries and reconstruction of the areas destroyed by earthquake in 1970 was undertaken. In housing, the government planned to strengthen housing cooperatives and to control the current process of urbanization.

Planning and investment

The public and private sectors. Private enterprise continued its heavy investment in the Peruvian economy in the early 1970s, with manufacturing and construction receiving about 10,000,000,000 and 13,000,000,000 soles, respectively. Mining received about 4,000,000,000, whereas agriculture, fishing, and power investment was about one-half that figure. The heaviest governmental expenditures—more than 20,000,000,000 soles—were for operations, but billions were poured into the same sectors of the economy as well as into transportation and communications, housing and education. At present, numerous firms and governments are interested in investing in Peruvian petroleum, copper, phosphates, irrigation, and metallurgy.

Taxation. Taking 1960 as the base year, the index of direct taxes rose more than 50 percent by 1966, and the indirect tax index rose by nearly 250 percent. First place went to the stamp tax, which in that year accounted for more than one-third of tax revenue. Personal income taxes are at a maximum of more than 50 percent, and luxury items are heavily taxed.

Labour. The Confederación de Trabajadores del Perú unites blue- and white-collar workers, most notably in the bus transportation, sugar, mining, banking, and teaching forces. There is also a dissident labour federation of Communists. Legislation since 1968 has created agricultural cooperatives and some industrial organizations that may in the long run decrease the power of the established unions. Strikes are subject to mediation among labour, management, and government. Recent legislation requires biennial collective work agreements.

Transportation. Peru's transportation system faces the challenge of the Andes and of the complex Amazon River system. The only really integrated systems are the roads and the airlines, for the railroads have not succeeded in forming a connecting network, and maritime traffic is primarily dependent on calls from world shipping lines. The 31,103 miles of roadway—asphalted, macadamized and metalled, or unpaved but drivable—cross the country from north to south, form the penetration roads that run west and east, or link the small villages to close up the road network. Among the longitudinal roads are the Pan American Highway, which crosses the country from Ecuador to Chile, and the macadamized Carretera Longitudinal de la Sierra, which will unite Ecuador with Bolivia. The principal penetration roads uniting the seacoast with the ports on the Amazonian rivers follow Chiclayo–Jaén–Nazareth, Lima–Pucallpa, Pisco–Teresita, and Mollendo–Puerto Maldonado routes. The Piura–Moyobamba–Yurimaguas will soon enter into service.

The major Peruvian railroad, the Ferrocarril Central, is the world's highest rail line, rising from the coast at Callao to cross the continental divide at 15,806 feet. It connects with the line from Cerro de Pasco, making it of great importance to the mining industry. Begun in the 1860s, it passes through 66 tunnels and over 59 bridges during its 215-mile course. A longer southern line, the Ferrocarriles del Sur, serves Cuzco, Arequipa, and other cities; much of its traffic originates in Bolivia.

Callao is the largest of Peru's numerous maritime, river, and lake ports. Chimbote is the world's largest fishing port. Iquitos, 2,300 miles up the Amazon, is a characteristic river port, with inadequate dock facilities, and unreachable by larger boats in the dry season.

The rough terrain compels the use of the airplane, yet makes it more difficult. Commercial aviation was founded in 1928, and four domestic lines currently operate in addition to APSA, an international carrier. There are also military aid and assistance services. An international airport in Callao is the most important of the more than 150 airports in the country.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. Since October 1968, Peru has been ruled by a revolutionary government of the armed forces, which has suspended the constitution, dissolved the Congress, reorganized the judiciary, and placed Gen. Juan Velasco Alvarado in the presidency. The present Peruvian regime has been concerned with increasing the nation's economic self-sufficiency. Principal measures of the new government include nationalization of petroleum (without excluding the possibility of granting new concessions); an agricultural reform law that terminated the old landed estates; a water law that converted all water into state property; a mining law to stimulate foreign investment; an industries law to strengthen national industry; and the nationalization of the bank.

The revolutionary government

Prior to the coup d'etat in 1968, Peru was nominally a democratic republic but one in which the great majority of the economic and political power in fact was wielded by a minute percentage of the population. The president and the members of the bicameral Congress were elected for six-year terms, and the Supreme Court of Justice was elected by the Congress. Local government was divided among municipalities by 23 departments, 1 constitutional province, 149 provinces, and 1,672 districts. Voting was compulsory for all adult citizens except the military, who also could not hold office. The majority political party was the Alianza Popular Revolucionaria Americana (APRA), founded in 1924 by Víctor Raúl Haya de la Torre.

Armed forces. Theoretically, all Peruvian males between 20 and 25 years of age are subject to two years of military duty, but the law has been poorly observed. The regular army, navy, and air force, though small, have modern equipment, most of it from the United States and France. The military also does road-construction work.

Social conditions. The oligarchy that long ruled Peru was severely disrupted by the 1968 coup. The haciendas have been given over to cooperatives of former employees, whereas workers were granted access to ownership of factories through the Industrial Community (Comunidad Industrial). The majority of Peruvians continue to earn less than $1 a day, however, and the country suffers from inflation. Sanitation is a major problem; most cities lack adequate sewerage, lighting, and paving.

Education. The revolutionary government has undertaken major reforms in education to strengthen the roles of the people in the restructuring of society and the increase of national self-sufficiency. Among the means is an ongoing drive against illiteracy, which in 1940 had held about 60 percent of the nation in its grip.

Education is free and compulsory for all children between six and 15, though outside urban centres—and often inside—this is difficult to enforce. In the mid-1960s, primary and secondary curricula were under review with the help of United Nations Educational, Scientific, and Cultural Organization (UNESCO), and teachers were being prepared for bilingual work in the mountains and the Montaña. Twenty-one national universities are supported by the state, and 12 other institutions are private. The University of San Marcos, in Lima, is the oldest university in South America, having been founded in 1551.

Health and housing. Infant and young-child mortality long was a major scourge of Peru, and it is one of the major concerns of government. Numerous public agencies are involved in national health and insurance work, including the military. With only about 6,500 doctors and 4,000 nurses, however, the nation faced a difficult road to adequate service for all age levels.

Peru is about 1,000,000 housing units short, especially in the *barriadas* but also in the countryside. An earthquake in May 1970 further aggravated the problem by destroying 10 percent of the housing units in the country and 200 towns and villages. Private construction has been aided by a system of mutual aid, and the state is building large developments.

CULTURAL LIFE AND INSTITUTIONS

The complex ethnic and cultural miscegenation of Peru presents an entwining of aboriginal pantheism, Spanish mysticism, and African religious practices that erupts into song, literature, textile art, engraved gourds, gold- and silverworking, and a bounteous cuisine.

Archaeological remains

The ancient culture of Peru perhaps is best known not only in its Spanish–Inca cities but also in such remains as Machu Picchu, a "lost city" of the Incas discovered in 1911 on an 8,000-foot-high mountain 50 miles from Cuzco. The most extraordinary ruin in the Americas, Machu Picchu comprises five square miles of terraced construction, 3,000 steps connecting its many levels, and a multitude of small houses built of unmortared stone.

The dryness of the central and southern coasts has preserved the remains of a long succession of pre-Inca peoples, about whom little is known except for their exquisitely decorated jugs and bowls and embroidered mantles. The Mochica people detailed every aspect of their lives, including erotic practices, on their jugs.

Folk culture. Peruvian folk culture is deeply tinged with inheritances from the ancestral races. In both towns and countryside notable examples of pre-Hispanic and *mestizo* lore abound in myths, song, superstitions, and dances representing a historical synthesis. An impressive system of learning based on the course of the seasons and of agricultural cycles includes mastery of the medicinal qualities of plants. The leaves of the coca plant are chewed by nearly all Indians to produce a narcotic effect. Intricately carved figures and elaborate horned, toothed, and tufted masks are popular with tourists and collectors. Ceramics often represent a blend of Catholic and Indian symbolism.

Literature, music, and the visual arts

Westernized arts. The arts long have occupied positions of esteem among Peru's educated minority. Since the late 19th century, most writers have felt a ceaseless duty to analyze their society. Ricardo Palma (died 1919) was among the first to utilize Peruvian themes. The poet José Santos Chocano (died 1934) was one of the first to write sympathetically about the Incas. The *pensador* (intellectual) Manuel González Prada (died 1919) founded a literary circle to encourage writers alienated by the nation's cultural dichotomies. *Aves sin nido* (1889; *Birds Without a Nest*, Eng. trans., 1904) by Clorinda Matto de Turner, a member of Prada's group, was the first of many books whose authors exposed the conditions of Indian life. César Vallejo is often regarded as Peru's finest poet; he died (1938) in poverty in Paris.

Valle-Riestra's *Ollanta* and Vicente Stea's *Sinfonía autóctona* (*Aboriginal Symphony*) were the basic musical works of 19th-century Peru. Later, Luís Duncker Lavalle incorporated Peruvian motifs into Western forms. Other users of folk materials include Daniel Alomía Robles, notably in his "Hymn to the Sun" and *The Condor Passes.*

Painting reached its zenith with the famous Cuzco school during the 17th and 18th centuries. Most of the thousands of paintings and sculptures are anonymous; the works show resemblances both to Byzantine and to Asian forms. Modern Peruvian art has followed an abstract course, notably in the work of the painter Fernando de Szyszlo and the sculptor Joaquín Roca Rey.

Architecture. The ancient Peruvians were great builders of houses, temples, palaces, fortresses, roads, and canals, adapting their architecture to the landscape. The oldest colonial work is the Cathedral of Lima, and the most important architectural jewel is the Convent and Church of San Francisco of Lima. Modern architecture has been characterized by the so-called neo-Peruvian, or Peruvian Baroque, and, recently, by the introduction of concrete and steel structures.

Communications. Some of Peru's daily newspapers, such as *El Peruano* and *El Comercio,* are among the oldest in America. Others of the more than 500 publications include popular magazines and the organs of specialized institutions in the arts and sciences. More than 200 radio stations blanket the entire country. Lima has seven television channels, whereas Arequipa, Huancayo, Trujillo, and Piura have one each.

Prospects. Peru's 1,000 years of cultural tradition are in the process of radical change. In a land of vast mineral wealth, grave shortages continue in many aspects of human life. Problems in internal communications, housing, health, and illiteracy are being tackled vigorously, and attempts are being made to draw the mass of workers and peasants more closely into the mainstream of national life. Few Peruvians pretend that the processes of change will be easy or quick.

BIBLIOGRAPHY

General works: A complete picture of Peru is found in the *Atlas histórico, geográfico y de paisajes peruanos* (1970), published by the President's National Planning Institute; other comprehensive studies include D.A. ROBINSON, *Peru in Four Dimensions* (1964); GEORGE PENDLE, *The Land and People of Peru* (1966); R.J. OWENS, *Peru* (1963); PRESTON E. JAMES, *Latin America,* 3rd ed., pp. 160–203 (1959).

Physical geography: Boundaries, dimensions, relief, and other features are treated in the analytical work of the MILITARY GEOGRAPHIC INSTITUTE, *La carta nacional* (1945–71), as well as in ISAIAH BOWMAN, *The Andes of Southern Peru* (1916); CARLOS PENAHERRERA DEL AGUILA, *Geografía general del Perú,* vol. 1, *Aspectos físicos* (1969); G.R. JOHNSON and R.R. PLATT, *Peru from the Air* (1930), aerial photographs; oceanography and marine biology are dealt with by ERWIN SCHWEIGGER, *El litoral peruano,* 2nd ed. (1964); and by ZACARIAS POPOVICI, *El Pacífico peruano* (1970); other works on the subject are OLIVIER DOLLFUS, *Les Andes centrales du Pérou et leurs piémonts* (1965); and CLAUDE COLLIN DELAVAUD, *Les Régions côtières du Pérou septentrional* (1968); geological works include GUSTAV STEINMANN, *Geologie von Perú* (1929); A. ALBERCA CEVALLOS, *Los minerales en el Perú* (1970); LEONIDAS CASTRO BASTOS, *Anotaciones geológicas y petrológicas sobre las ocho regiones naturales* (1967); JUAN DE LA CRUZ LOPEZ, *Algunas observaciones sobre las unidades morfológicas en las ochos regiones naturales del Perú* (1967); the lands of Peru are treated in CARLOS J. ZAMORA, *Suelos del Perú y capacidad de uso de los suelos* (1970); a detailed analysis of the natural regions of Peru and the principal types of its vegetation are found in JAVIER PULGAR-VIDAL, *Geografía del Perú: las ocho regiones naturales* (1967) and in his *Análisis geográfico sobre las ocho regiones naturales del Perú* (1960); also valuable are JOSEPH A. TOSI, JR., *Zonas de vida natural en el Perú* (1960); and AUGUST WEBERBAUER, *El mundo vegetal de los Andes peruanos,* new ed. (1945).

People: The life of the Peruvian population is dealt with in JULIO C. TELLO, *Origen y desarrollo de las civilizaciones prehistóricas andinas* (1942); EFRAIN ORBEGOSO RODRIGUEZ, *Geografía humana del Perú* (1968) and *Migraciones en el Perú* (1970); AVENCIO VILLAREJO, *Así es la selva,* 2nd ed. (1953), and *La selva y el hombre* (1959); JOHN P. GILLIN, *Moche: A Peruvian Coastal Community* (1947); HARRY TSCHOPIK, *Highland Communities of Central Peru: A Regional Survey* (1947); BERNARD MISHKIN, "The Contemporary Quechua," in JULIAN H. STEWARD (ed.), *Handbook of South American Indians,* vol. 2 (1946); JOSE MATOS MAR, "The 'Barriadas' of Lima: An Example of Integration into Urban

Life," in PHILIP M. HAUSER (ed.), *Urbanization in Latin America* (1961); birth and death rates, ethnic composition, and demographic trends are treated in numerous works published between 1964 and 1971 by the NATIONAL PLANNING INSTITUTE OF PERU.

Economics: A realistic picture of the economic situation in Peru is provided by the ACCION PARA EL DESARROLLO (Developmental Action Committee), *Memorandum sobre la situación económica nacional* (1970); economic information in English is found in D.R. KILTY, *Planning for Development in Peru* (1967); and FERNANDO BELAUNDE-TERRY, *Peru's Own Conquest* (1965); a systematic study of the economic geography of Peru is provided in EMILIO ROMERO, *Geografía económica del Perú*, 3rd ed. (1953), and economic history in his *Historia económica del Perú* (1971).

Social and political life: Excellent analyses are found in VICTOR RAUL HAYA DE LA TORRE, *El antimperialismo y el Apra* (1936); HARRY KANTOR, *The Ideology and Program of the Peruvian Aprista Movement* (1953); FRANCOIS BOURRICAUD, *Pouvoir et société dans le Pérou contemporain* (1967; Eng. trans., *Power and Society in Contemporary Peru*, 1970); and J.C. CAREY, *Peru and the United States, 1900–1962* (1964).

Culture: LUIS C. LUMBRERAS, *Panorama de la arqueología peruana* (1970); JORGE GUILLERMO LLOSA, *La cultura peruana en el siglo XX* (1963); LUIS ALBERTO SANCHEZ, *Panorama cultural del Perú* (1968); AUGUSTO TAMAYO VARGAS, *La literatura en el Perú* (1968); JUAN MANUEL UGARTE ELESPURU, *Problema y posibilidad de la pintura actual* (1968); HECTOR VERLARDE, *La arquitectura peruana del siglo XX* (1963).

(J.P.-V.)

Peru, History of

Before its conquest from the Incas by the Spaniards under Francisco Pizarro in 1531–33, Peru had been the home of advanced Indian civilizations for more than a thousand years (see ANDEAN CIVILIZATION, HISTORY OF).

Spain consolidated its control over Peru, and the region remained under Spanish rule for nearly 300 years (see LATIN AMERICA AND THE CARIBBEAN, COLONIAL). In the early 19th century, despite moves toward independence by other Spanish colonies, Peru remained loyal to Spain because of the conservatism of its aristocracy, the presence of many Spaniards in the region, the concentration of Spanish military power in Lima, and the effectiveness of Spanish officials in Peru in suppressing uprisings. Independence was imposed on it primarily by outsiders.

Peru from 1824 to 1930. The Venezuelan general Antonio José de Sucre's victory over Spanish forces at Ayacucho in the south central Peruvian Andes in 1824 brought independence to Peru. The end of Spanish rule did not, however, provide a solution to the many political, social, and economic problems facing the country. The transition from a colonial dependency to a modern state proved difficult and was still incomplete in the 1970s.

The caudillos

Struggle for power. At the outset of Peru's national existence, military leaders (*caudillos*) who had gained prominence in the struggle for independence sought to seize power. The departure of Simón Bolívar, the South American liberator whose army had been essential in freeing the country, removed a stabilizing influence. The aims of the caudillos were furthered by the absence of a tradition of self-government, by the prevalence of a feudal society of creoles and Indians, and by the reluctance of civilians to assume political responsibility. Despite military influences, a liberal constitution was adopted in 1828. This did not prevent Gen. Agustín Gamarra a year later from becoming head of the government by illegal means. His position was successfully challenged in 1835 by another self-seeking caudillo, Gen. Felipe Salaverry.

Peruvian-Bolivian confederation. The ambitions of Gamarra and Salaverry were thwarted by Andrés Santa Cruz, an intelligent military commander of Spanish-Indian descent, who proposed a confederation of Peru and Bolivia. For three years Santa Cruz, though born in La Paz, obtained the backing of influential groups in Peru and maintained the political union. But his hopes were shattered at the Battle of Yungay in 1839 by a joint force of nationalist-minded Peruvians and of Chileans fearing a threat to the balance of power in the Pacific.

Establishment of order. During the initial period of statehood in Peru, liberal and conservative parties with ill-defined programs emerged. Their rivalry enhanced the political instability of the country.

Growth under Ramón Castilla

Gen. Ramón Castilla assumed the presidency in 1845. Castilla dominated the political scene from 1845 to 1851 and from 1855 to 1862, in spite of his mestizo (racially mixed) background. His greatest accomplishment for the promotion of national wealth was the exploitation of the guano deposits along the coast and offshore islands. Taxes on this industry, which was controlled by foreign corporations, furnished the principal source of government revenue for several decades. Castilla appealed to liberals by abolishing the payment of tribute by Indians and by emancipating the Negro slaves. Landowners on the coastal plantations, however, were permitted to import thousands of Chinese coolies in order to have a sufficient labour supply. As an additional concession to the liberals, Castilla established a system of state education at the primary and secondary levels. Owing to his influence, an assembly in 1860 adopted a constitution that lasted into the 1900s.

National setbacks

The Spanish invasion (1864–69). In the second half of the 19th century, Peru's history was characterized by many setbacks. In 1864 Spain dispatched a naval force to the Pacific, ostensibly to protect the rights of Basque immigrants, but in reality to re-establish domination over its former colony. In 1869, after meeting with the determined opposition of Peru and Chile, Spain withdrew and recognized Peru's independence for the first time, but the conflict was a heavy drain on Peru's treasury.

The Meiggs railway project. Dissatisfaction with military rule resulted in 1871 in the formation of the Civilian Party, representing an oligarchy of landowners and merchants. This party, headed by Manuel Pardo (president, 1872–76), approved a costly program of internal development, which included the construction of railroads across the Andes by Henry Meiggs, an unscrupulous U.S. entrepreneur. Corruption and fraud on the part of government officials and contractors characterized the work. Although it decreased the isolation of the Peruvian interior, it increased enormously the national debt.

The War of the Pacific (1879–84). Another untoward event was the War of the Pacific with Chile, 1879–84, caused mainly by rivalry over the exploitation of the rich nitrate beds in the Atacama desert then in Peru, now between Chile and Peru. Chile's superior resources and military discipline brought overwhelming defeat to Peru and its ally Bolivia.

At the Battle of Iquique (then in Peru, now in Chile), on May 21, 1879, the Peruvians suffered the loss of their best warship, the "Independencia." The eventual surrender of control of the sea permitted a Chilean army to land on the Peruvian coast. On January 17, 1881, Chilean forces captured the capital, Lima. Looting and pillage followed, with the destruction of the National Library, which had over 58,000 volumes. According to the terms of the Treaty of Ancón (October 20, 1883), Peru turned over to Chile full possession of the province of Tarapacá (now in northern Chile) and the administration for ten years of the provinces of Tacna and Arica, after which a plebiscite was to determine their future sovereignty.

The Peruvian Corporation. Expenditures for the war, and the consequent loss of revenue from the nitrate fields, created the possibility of imminent bankruptcy. To avert this disaster, the Civilian regime accepted in 1889 a plan proposed by the bondholders for handling the debt. The Peruvian Corporation, representing the creditors, with headquarters in London, was to control the railroads for 66 years, to mine up to 3,000,000 tons of guano, and to receive 33 annual payments of £80,000 each. The plan worked satisfactorily to the benefit of both parties.

Peaceful civilian rule

Social reforms and economic development. The decline in national prestige created an atmosphere conducive to political change. The Democratic Party was formed, and in 1895, under the leadership of Nicolás de Piérola, it won the presidential election. Having a broad, popular base, it championed direct suffrage and the restoration of municipal elections. Public education was fostered, but schools for the children of the poor were lacking in both urban and rural areas.

An orderly political scene, marked by rivalry between the Democratic and Civilian groups, accelerated economic development. There was an increase in the production of minerals, notably copper, and of such agricultural commodities as cotton, sugar, and wool. In the mining of copper, U.S. capital acquired important interests, founding the Cerro de Pasco Corporation in the central Peruvian highlands. Small factories designed for the domestic market appeared in the cities. Railroad construction, halted in the 1880s, was resumed, stimulating trade.

The presidencies of Augusto Leguía (1908–12; 1919–30). In 1908 Augusto Leguía, chief spokesman for the Civilians, assumed the presidency; a former minister of the treasury, he was closely identified with business groups. By selling guano to owners of the large coastal estates the government expanded indirectly the production of sugar and cotton—thenceforth important props for the economy.

Leguía's second period in office was generally less successful. By advocating social reform and the modernization of the country, he appealed strongly to many sectors, including the newly formed Socialist Party, but he permanently undermined the Civilian and Democratic parties. Eventually, his leadership rested on a popular basis. In line with his theme of progressiveness, Peru received a new constitution in 1920 to replace the one of 1860. The new charter authorized the state to limit property rights; to build hospitals, clinics, and asylums; and to provide free primary education; Article 58 protected the communal lands of the Indians from sale and seizure.

The administration's economic policies achieved a temporary prosperity. An expensive program of public works aided a number of cities and communities, especially Lima, the capital. Irrigation projects opened up 100,000 acres along the coast. To provide technicians, the government founded a National Agricultural School in Lima. Foreign investments were encouraged. In 1924 the International Petroleum Company, a subsidiary of the U.S.-owned Standard Oil Company of New Jersey, acquired the oil fields of La Brea-Pariñas, near Talara on the northern coast, from the London and Pacific Petroleum Company, and built a refinery to supply the nation with oil and gasoline; the nature of the British company's exploitation of the concessions raised the question of the legality of the purchase. To finance its policies, the Peruvian government resorted to large loans from U.S. banks.

Indianism

The period of the 1920s witnessed a significant development of Indianism. The government's failure to implement Article 58 of the constitution made the issue a critical one. Peruvian intellectuals created a vogue of Inca history, tradition, and archaeology; the Indian, they argued, should enjoy equality in society hitherto denied him. While a majority of thinkers urged gradual reforms, some favoured radical means to achieve this goal. This latter group found support in the ideology of the Mexican Revolution of 1910; one of these was José Carlos Mariátegui, the founder of the Peruvian Communist Party.

Formation of the Aprista movement. The Alianza Popular Revolucionaria Americana (APRA), known as the "Aprista movement," was formed in 1924 in Mexico City by Víctor Raúl Haya de la Torre, an intellectual then in exile. Internationally, it expressed the ideals of the unity of American Indians and the elimination of U.S. imperialism. Nationally, it proclaimed the need to end the exploitation of the Peruvian masses through the institution of a planned economy and the nationalization of foreign-owned enterprises. Its anticapitalist and anti-imperialist stand appealed to intellectuals, to the Indians, and to the lower middle class. Many teachers and students participated in the movement, and the University of San Marcos became a stronghold of radicalism.

Peru since 1930. By 1930 Leguía had experienced a definite loss in popularity. Final settlement of the long-standing Tacna-Arica dispute with Chile, by which Peru ceded the province of Arica, angered the extreme nationalists. The world depression, resulting in a serious decline in exports and a curtailment of loans from U.S. bankers, cost Leguía the support of the commercial and financial groups. Finally, the establishment of a repressive dictatorship alienated the liberals. In 1930 a military junta headed by Gen. Luis Sánchez Cerro, a mestizo, engineered Leguía's overthrow.

The presidencies of Luis Sánchez Cerro and Oscar Benavides (1931–39). Politically, the era of military domination in the 1930s was one of the most strife-torn in Peru's history. In 1931 Sánchez Cerro obtained the presidency over the Aprista candidate Haya de la Torre. In 1933 a new constitution set up a presidential parliamentary system with a strong executive to be chosen for a five year term.

Aprista opposition

Sánchez Cerro had to deal both with strong political opposition and with the effects of the depression. APRA claimed that the 1931 election was fraudulent and instigated a campaign to discredit the regime. The threat from the left led to the emergence of a Fascist group, whose chief exponent was the conservative historian José de la Riva Agüero. Sánchez Cerro's efforts to win over the Indians by a promise of agrarian reform were futile. In July 1932 Apristas organized an uprising in Trujillo, along the northern coast. During their withdrawal the revolutionaries shot hostages, including officers and enlisted men; the army retaliated with wholesale executions. Sánchez Cerro was assassinated by an Aprista in 1933. These incidents created an enduring enmity between the military establishment and APRA.

Sánchez Cerro's successor was Gen. Oscar Benavides (1933–39), who restored confidence in the economy. Benavides settled a dangerous boundary controversy with Colombia that had erupted into open fighting in 1932. The scene of the dispute was the port of Leticia on the upper Amazon and a finger of land giving access to the river, both of which had been ceded to Colombia in a treaty of 1922. To avoid full-scale war, Benavides disclaimed the acts of invading Peruvians and returned the territory to Colombia.

Benavides reduced the strength of the APRA, now claiming a majority of the middle and lower sector of the population, by outlawing the party, by a relentless persecution of its leaders, and by the adoption of social assistance projects. In the presidential election of 1939 the Apristas backed Benavides' choice—Manuel Prado, a banker and member of an aristocratic family of Lima.

The first presidency of Manuel Prado (1939–45). During World War II Peru cooperated with the United States. Prado authorized Allied use of airfields and ports and arranged to sell them petroleum, cotton, and minerals. In 1942 Peru severed diplomatic relations with the Axis powers, and in 1945 it declared war against them. During the war Peru succeeded, with U.S. backing, in getting a favourable settlement of a boundary dispute with Ecuador, which recognized Peru's sovereignty over a stretch of territory along the Marañon River (an Amazon headwater) and over the town of Zarumilla.

Alignment with the United States

Gains by the Apristas. World War II not only brought economic prosperity but also hope for real democracy. Prado, swayed by public opinion, approved the presidential candidacy in 1945 of José Bustamante, a lawyer from Arequipa with liberal leanings, who represented a coalition of middle- and upper-class elements. APRA, again a legal party, obtained a majority of seats in the lower house and half the seats in the senate. Because of Haya de la Torre's secret acquiescence in Bustamante's selection, APRA believed that many of its policies would be accepted. But Bustamante generally followed an independent course, and the Apristas withdrew their collaboration. After Apristas staged an abortive insurrection in Callao, near Lima, the president outlawed the party.

The dictatorship of Manuel Odría (1948–56). In October 1948 Gen. Manuel Odría seized power in the capital, protesting the president's lack of firmness in dealing with the radicals; Odría served as president until 1956. Extreme measures were taken to suppress the Apristas. Haya de la Torre found refuge in the Colombian embassy, where he stayed five years before leaving Peru.

From 1948 to 1956 Peru had an authoritarian regime. Political stability allowed the revival of prosperity, which had been hurt by a postwar recession. The Korean conflict of the early 1950s augmented foreign trade because

of heavy U.S. demand for Peruvian minerals. A friendly policy toward foreign capital prompted investments on a large scale. To appeal to the lower classes, Odría advocated medical and social assistance; but these proved to be illusory. Some progress was made, however, in the establishment of primary and secondary schools. Lavish expenditures in the capital, marked by official graft, and the ending of the Korean War led to the adoption of inflationary policies by the treasury.

Prado's second term (1956–62). In the election of 1956, Manuel Prado, who was supported by Odría, won a second term. Prado defeated Fernando Belaúnde-Terry, a U.S.-educated architect and a descendant of a prominent Peruvian family, who captivated intellectuals, students, and professional groups by advocating thorough social and economic reforms. A surprising feature of the election was the decline of APRA, some of whose members joined Belaúnde's National Front Party.

Decline of APRA

By currying favour with the Apristas, again a legal body, Prado avoided disorders. He countered the financial crisis inherited from Odría by appointing as minister of the treasury Pedro Beltrán, director of the conservative Lima newspaper *La Prensa,* whose orthodox policies contributed to a 4½ percent annual increase in the gross national product. Aiding the economy was the expansion of the fishing industry based on the Peruvian Current; by 1966 Peru was foremost in the world in anchovy and fishmeal products. Beltrán's measures did not, however, lessen the pressure from the landless Indians and the underpaid urban proletariat.

The presidency of Fernando Belaúnde-Terry (1963–68). With political tension at a high level in 1962, none of the presidential candidates of the three leading parties received the necessary one-third vote for election. Haya de la Torre had a slight plurality, followed closely by Belaúnde-Terry and Odría. The election went to the congress for a decision, but the military, fearing an alliance between the conservatives and the Apristas that might have given the latter a strong influence, stepped in and seized the government. A new election called in 1963 by the junta permitted Belaúnde's party, now called "Popular Action," to be victorious.

Reforms of Belaúnde

Belaúnde promised practical solutions to the country's economic and social problems, especially the high rate of population increase (in 1966 Peru had an estimated population of 12,000,000, three-fourths of whom had a subnormal standard of living). An agrarian act of 1964 provided for more effective use of the land, with expropriation of unused or misused agricultural properties; by 1966 more than 500,000 acres had been distributed to the peons. To enhance living conditions and the dignity of the Indians, the government organized community development projects; Indians were encouraged to colonize land in the foothills east of the Andes. Irrigation schemes were instituted for arid areas on the coast, and a network of roads was planned for the interior. Education was promoted with the establishment of new universities and with attacks on illiteracy—by 1961 down to 40 percent. Excessive outlays for these projects provoked a financial crisis.

Military rule. On October 3, 1968, the military reasserted its power and forced the resignation of Belaúnde on the pretext of his unduly moderate position toward the International Petroleum Company. Behind this, however, was a fear of APRA acquiring power in the forthcoming elections. The junta, headed by Juan Velasco Alvarado, jailed opposing politicians and suspended constitutional liberties. On October 9 the government expropriated the holdings of the International Petroleum Company, declaring that the company owed $690,500,000 for exploitation of illegal concessions. This further strained relations with the U.S., already affected by seizures of U.S. fishing craft in a 200-mile limit claimed by Peru as national waters and by the purchase of supersonic jet fighters from France. The U.S. State Department protested but refrained from cutting off foreign aid. As a gesture of independence Peru opened diplomatic relations with the Communist Bloc nations.

In 1969 the junta embarked on a vigorous program of economic nationalism. Velasco, while denouncing Marxism, proclaimed his intent "to combine public and private ownership of the economy." This policy would affect U.S. capital investments totalling $600,000,000, which had contributed to the economic growth of the country. In 13 months, three basic reform measures were enacted; the Agrarian Law (June 24, 1969); the Mining Law (April 14, 1970); and the Industrial Law (July 30, 1970). Accordingly, the government seized on August 22, 1969, the Paramonga chemical and paper complex of W.R. Grace & Company, the largest U.S. interest in Peru next to that of Standard Oil, and many foreign and domestic owned plantations on the coast. It sought also to control essential industries and public services through outright ownership and by "Peruvianization"—insistence that a majority of the stock of a foreign company be held by nationals. The success of the government's program depended on the continued support of the Indian masses and of the growing middle sectors and on the exercise of managerial ability. The occurrence on May 31, 1970, in the northeastern Andean region and in the adjoining coastal strip, of the worst Latin American earthquake in the 20th century partly jeopardized the financial stability of the regime. Loss of life was estimated at 50,000, with property damage in the hundreds of millions of dollars.

Economic nationalism

BIBLIOGRAPHY

General works: JORGE BASADRE, *Historia de la república del Perú,* 5th ed., 7 vol. (1963–64), a monumental account covering the years 1824 to 1930 by one of Peru's most highly regarded historians; SIR ROBERT MARETT, *Peru* (1969), a commendable synthesis from earliest times to the military dictatorship of Velasco; SIR CLEMENT MARKHAM, *A History of Peru* (1892), an old but reliable account; F.B. PIKE, *The Modern History of Peru* (1967), by far the most accurate and complete history in English.

Specialized works: C.A. ASTIZ, *Pressure Groups and Power Elites in Peruvian Politics* (1969), an analysis of political forces in the contemporary era; FERNANDO BELAUNDE-TERRY, *Peru's Own Conquest* (1965), a popular account of the program of reform embracing all sections of the country; J.C. CAREY, *Peru and the United States, 1900–1962* (1964), a brief but satisfactory treatment of diplomatic relations between the two countries; PAN AMERICAN UNION, *Constitution of the Republic of Peru, 1933* (1962); W.J. DENNIS, *Tacna and Arica* (1931), an account of the territorial dispute between Peru and Chile dating from the War of the Pacific; T.R. FORD, *Man and Land in Peru* (1962), a sociological study; V.R. HAYA DE LA TORRE, *El Plan del aprismo, programa de gobierno del Partido aprista peruano* (1933), an exposition of the ideals of the Aprista Party by its founder; J.W. HOPKINS, *The Government Executive of Modern Peru* (1967), a comparative study of the policy-making section of the bureaucracy; HARRY KANTOR, *The Ideology and Program of the Peruvian Aprista Movement* (1953), a basic treatment of the ideas and organization of the Aprista movement at the height of its influence; MIGUEL MARTINEZ, *La vida heroica del Gran Mariscal don Ramón Castilla* (1952), a sound biography of a national hero; GUSTAVO PONS MUZZO, *Historia del conflicto entre el Perú y España, 1864–1866: el 2 de Mayo de 1866* (1966), a narrative of the Spanish-Peruvian war; WATT STEWART, *Chinese Bondage in Peru: A History of the Chinese Coolie in Peru, 1849–1874* (1951), a study of the introduction of Chinese workers, with mention of the Chinese-Peruvian treaty, the first signed by China with a Latin American state, and *Henry Meiggs: Yankee Pizarro* (1946), an intriguing story of an American entrepreneur in railroad construction of the 1870s; ALBERTO ULLOA Y SOTOMAYOR, *Don Nicolás de Piérola: una época de la historia del Perú* (1949), an objective biography of the founder of the Democratic Party; A.P. WHITAKER, *The United States and South America: The Northern Republics* (1948), includes diplomatic relations with Peru in the post-World War II period; A.C. WILGUS (ed.), *South American Dictators During the First Century of Independence* (1937).

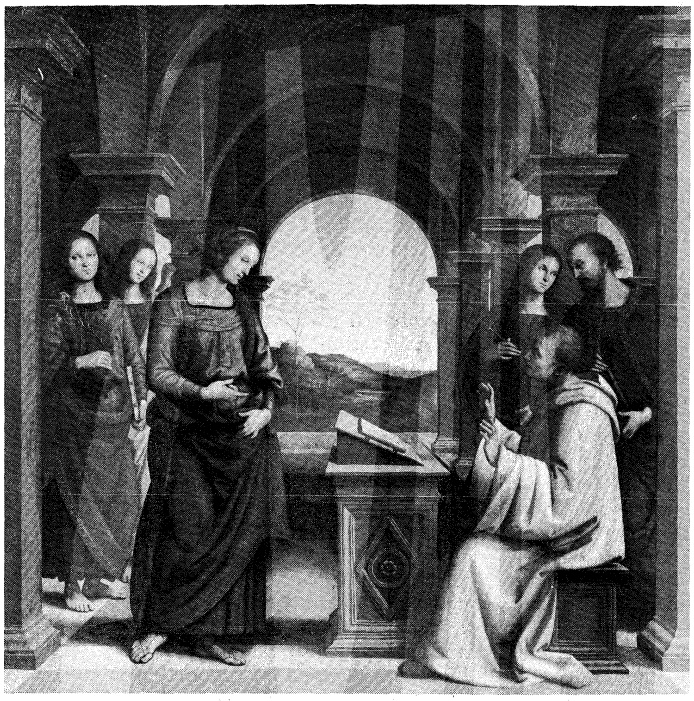
(J.P.M.)

Perugino

Of all the Italian Renaissance painters of the late 15th century, Perugino, an artist from the province of Umbria, was one of the most significant. The compositional clarity, sense of spaciousness, and economy of formal elements in his works were to become some of the distinguishing aesthetic principles of the High Renaissance

style of the early 16th century. Perugino is also important for his influential role as the teacher of the famous painter Raphael, whose paintings are regarded as the quintessence of High Renaissance art.

By courtesy of the Alte Pinakothek, Munich



"Vision of St. Bernard," wood panel by Perugino, *c.* 1491–94. In the Alte Pinakothek, Munich. 1.73 × 1.70 m.

Born about 1450 at Città della Pieve, Italy, Pietro di Cristoforo Vannucci became known as Perugino, derived from the Umbrian city of Perugia located near his birthplace. Nothing is known for certain of his early training, but he may have been a pupil of Fiorenzo di Lorenzo (*c.* 1440-1525), a minor painter in Perugia, and of the renowned Umbrian Piero della Francesca (*c.* 1420–92) in Arezzo, in which case he would have been a fellow pupil of one of his most famous contemporaries, Luca Signorelli. The two men were acquainted, and an occasional influence from Signorelli is visible in Perugino's work, notably in the direction of an increased hardness of drawing (*e.g.*, "Crucifixion and Saints," *c.* 1480–1500; Uffizi, Florence). In Florence, where he is first recorded in 1472, he almost certainly worked in the shop of the important painter and sculptor Andrea del Verrocchio, where the young Leonardo da Vinci was apprenticed.

Early work

The first certain work by Perugino is a "Saint Sebastian," at Cerqueto, near Perugia. This fresco, or mural painted on plaster with water-dissolved pigments, dates from 1478 and is typical of Perugino's style. He must have attained a considerable reputation by this time, since he probably worked for Pope Sixtus IV in Rome, 1478–79, on frescoes now lost. Sixtus IV also employed him to paint a number of the frescoes in the Sistine Chapel in the Vatican Palace. Completed between 1481 and 1482, three narrative scenes behind the altar were destroyed by Michelangelo in 1535–36 in order to use the space for his fresco of the "Last Judgment." Of the scenes completely by Perugino's own hand, only the fresco "Giving of the Keys to St. Peter" has survived. The simple and lucid arrangement of the composition reveals the centre of narrative action unlike the frescoes in the same series by the Florentine painter Sandro Botticelli, which, in comparison, appear overcrowded and confused in their narrative focus. After completing his work in the Sistine Chapel, Perugino returned to Florence, where he was commissioned to work in the Palazzo della Signoria. In 1491 he was invited to sit on the committee concerned with finishing the Florence Cathedral.

Summit of artistic career

From approximately 1490 to 1500 Perugino was at his most productive and at the artistic summit of his career. Among the finest of his works executed during this time are the "Vision of St. Bernard," the "Madonna and Saints," the "Pietà," and the fresco of the "Crucifixion" for the Florentine convent of Sta. Maria Maddalena dei Pazzi. These works are characterized by ample sculpturesque figures gracefully posed in simple Renaissance architectural settings, which act as a frame to the images and the narrative. The harmonious space is tightly controlled in the foreground and middle ground, while the background effect is conversely one of infinite space. During this period he painted his best known portrait, a likeness of "Francesco delle Opere." Perugino must have been well acquainted with the late 15th-century portraiture of Flanders, since the influence of the Flemish painter Hans Memling is unmistakable.

Commissioned by the guild of bankers of Perugia, Perugino painted a fresco cycle in their Sala dell'Udienza that is believed to have been completed during or shortly after 1500, the date that appears opposite Perugino's self-portrait in one of the scenes. The importance of these frescoes lies less in their artistic merit than in the fact that the young Raphael, Perugino's pupil around 1500, probably was an assistant learning the technique of fresco painting. An allegorical figure of Fortitude from this series is often attributed to Raphael.

Creative decline

After 1500, Perugino's art began to decline, and he frequently repeated his earlier compositions in a routine manner. Giorgio Vasari, a 16th-century biographer and artist, wrote that the critical Florentines began to lampoon him, and Perugino replied that they had once praised his work, and, if he now gave the same designs, they had no right to blame him. It is certainly true that the "Combat of Love and Chastity" (Louvre, Paris) was commissioned in 1503 by Isabella d'Este and was delivered only in 1505, after a great many letters had passed between all concerned, at which time Isabella expressed herself as satisfied but only moderately so. In fact, Perugino left Florence about 1505 and began to work principally for the less critical public of Umbria. In 1508 he made a temporary comeback by painting roundels on the ceiling of the Stanza dell'Incendio in the Vatican. The commission for the frescoes on the walls of the room went to his pupil Raphael, who, in the few years after leaving Perugino's studio, proved himself the greater artist. One of Perugino's last commissions was the completion in 1521 of some frescoes in S. Severo, Perugia, which had been begun by Raphael. He was still painting in February or March 1523, when he died of the plague at Fontignano, near Perugia. The fresco of the "Nativity," in the National Gallery, London, comes from Fontignano and is generally supposed to be Perugino's last work.

MAJOR WORKS

"St. Sebastian" (1478; Cerqueto, Italy); "Giving of the Keys to St. Peter" (1481–82; Vatican, Sistine Chapel, Rome); "Crucifixion" (*c.* 1485; National Gallery, Washington D.C.); "Vision of St. Bernard" (*c.* 1491–94; Alte Pinakothek, Munich); "Adoration of the Child, Crucifixion, and Saints" (1491; Villa Torlonia, formerly Albani, Rome); "Madonna and Saints" (1493; Uffizi, Florence); "Madonna and Saints" (1493; Kunsthistorisches Museum, Vienna); "Crucifixion" (1493–96; Convent of Sta. Maria Maddalena dei Pazzi); "Francesco delle Opere" (1494; Uffizi, Florence); "Pietà" (*c.* 1494–95; Uffizi, Florence); "Virgin in Glory with Saints" (*c.* 1495–96; Pinacoteca Nazionale, Bologna); "Agony in the Garden" (1495; Galleria Palatina, Florence); Altarpiece (1497; Church of Sta. Maria Nuova, Fano, Italy); "Ascension" (1496–98; Musée Municipal des Beaux-Arts, Lyons), panels belonging to this painting in museums at Nantes and Rouen; "Annunciation" (1498; Church of Sta. Maria Nuova, Fano, Italy); "Archangel Michael, The Virgin Adoring the Child, Archangel Raphael with Tobias" (*c.* 1499; National Gallery, London); "Biagio Milanesi" (1500; Uffizi, Florence); "The Monk Baldassare" (1500; Uffizi, Florence); "Prophets, Heroes, Sibyls, and Sages, the Transfiguration and the Adoration of the Shepherds" (*c.* 1497–1500; Sala dell'Udienza, Collegio del Cambio, Perugia); "St. Sebastian" (1500–08; Hermitage, Leningrad); "Resurrection" (*c.* 1501; Vatican Gallery, Rome); "Adoration of the Shepherds" (1501; Galleria

Nazionale dell'Umbria, Perugia); "Crucifixion and Saints" (1502–06; Church of S. Agostino, Siena); "Marriage of the Virgin" (1503–04; Musée de l'Hôtel de Ville, Caen, France); "Combat of Love and Chastity" (1505; Louvre, Paris); "Deposition" (1505–07; Uffizi, Florence); "Assumption" (1506; Uffizi, Florence); "Ascension" (*c.* 1509; Cathedral, San Sepolcro); Altarpiece (before 1517; Galleria Nazionale dell'Umbria, Perugia); "Adoration of the Magi" (1521; Church of Sta. Maria delle Lacrime, Trevi, Italy); "Nativity" (1522; National Gallery, London).

BIBLIOGRAPHY. The most complete documentation on Perugino is in FIORENZO CANUTI, *Il Perugino*, 2 vol. (1931); but there is a useful modern monograph by E. CAMESASCA, *Tutta la pittura del Perugino* (1959). See also the lengthy article by FRANCESCO SANTI in the *Encyclopedia of World Art*, vol. 11, col. 265–271 (1966); and FREDERICK HARTT, *A History of Italian Renaissance Art* (1969).

(P.J.Mu.)

Pestalozzi, Johann Heinrich

Johann Heinrich Pestalozzi, Swiss educational reformer, was among the first to recognize the social problems of the industrial age and to stress the need for better popular education. Believing that every man, even the lowest, has inherent powers capable of development and has equal rights to be educated, he devised a teaching method aimed at strengthening the child's own faculties to make him proficient and secure, able to help himself and others. His "method" became widely accepted and gave impetus to teacher training. Most of his principles have been absorbed into modern elementary education.

By courtesy of the Bibliotheque Nationale Suisse



Pestalozzi, lithograph after a drawing by G.A. Hippius, 1818. In the Pestalozzianum, Zürich.

Pestalozzi was born in Zürich on January 12, 1746. He was influenced by the political conditions of his country and by the educational ideas of Jean-Jacques Rousseau; he abandoned the study of theology to go "back to Nature." In 1769 he took up agriculture on neglected land near the River Aare—the Neuhof. When this enterprise collapsed in 1774, he took poor children into his house, making them work by spinning and weaving and learn simultaneously to become self-supporting. This project also failed materially, although Pestalozzi had gained valuable experience. He continued to take an active interest in Swiss politics.

As practical realization of his ideas was denied him, he turned to writing. *The Evening Hour of a Hermit* (1780) outlines his fundamental theory that education must be "according to nature" and that security in the home is the foundation of man's happiness. His novel *Leonard and Gertrude* (1781–87), written for "the people," was a literary success as the first realistic representation of rural life in German. It describes how an ideal woman exposes corrupt practices and, by her well-ordered homelife, sets a model for the village school and the larger community. The important role of the mother in early education is a recurrent theme in Pestalozzi's writings.

Life at Neuhof

For 30 years Pestalozzi lived in isolation on his Neuhof estate, writing profusely on educational, political, and economic topics, indicating ways of improving the lot of the poor. His proposals were ignored by his own countrymen, and he became increasingly despondent. He would have accepted the post of educational adviser anywhere in Europe had it been forthcoming. His main philosophical treatise, *Meine Nachforschungen über den Gang der Natur* (1797; "My Inquiries into the Course of Nature in the Development of Mankind"), reflects his personal disappointment but expresses his unshakable belief in the resources of human nature and his conviction that man himself is responsible for his moral and intellectual state. Thus, Pestalozzi was convinced, education should develop man's innate faculties to think for himself.

Pestalozzi's chance to act came after the French Revolution, when he was over 50 years old. The French-imposed Helvetian Republic in Switzerland invited him to organize higher education, but he preferred to begin at the beginning. He collected scores of destitute war orphans and cared for them almost single-handedly, attempting to create a family atmosphere and to restore their moral qualities. These few exhausting months in Stans (1799) were, according to Pestalozzi's own account, the happiest days of his life.

From 1800 to 1804 he directed an educational establishment in Burgdorf and from 1805 until 1825 a boarding school at Yverdon, near Neuchâtel. Both schools relied for funds on fee-paying pupils, though some poor children were taken in; and these institutes served as experimental bases for proving his method in its three branches —intellectual, moral, and physical, the latter including vocational and civic training. They also were to finance his life's "dream," an industrial (*i.e.*, poor) school. The Yverdon Institute became world famous, drawing pupils from all over Europe as well as many foreign visitors. Some visiting educators—*e.g.*, Friedrich Froebel, J.F. Herbart, and Carl Ritter—were so impressed that they stayed on to study the method and later introduced it into their own teaching.

The Yverdon Institute

While dedicated assistants carried on the teaching, Pestalozzi remained the institute's heart and soul and continued to work out his method. *Wie Gertrud ihre Kinder lehrt* (1801; *How Gertrude Teaches Her Children*, 1894) contains the main principles of intellectual education: that the child's innate faculties should be evolved, that he should learn how to think, proceeding gradually from observation to comprehension to form clear ideas. Although the teaching method is treated in greater detail, Pestalozzi considered moral education pre-eminent.

The family spirit prevailing in Yverdon was shattered in later years by a progressively severe dispute among the teachers for first place by Pestalozzi's side. The longed-for poor school, established by means of the proceeds from publication of his collected works, existed for only two years. The Yverdon Institute lost its fame and its pupils to Pestalozzi's great distress. His efforts at reconciliation were in vain. With a few pupils he retreated to Neuhof in 1825, sad but convinced that his ideas would prevail in the end. His *Schwanengesang* (1826; "Swan Song") culminated in the maxim "Life itself educates." He died on February 17, 1827.

Pestalozzi was an odd-looking man but an impressive personality, highly esteemed by his contemporaries. His concept of education embraced politics, economics, and philosophy, and the influence of his "method" was immense.

BIBLIOGRAPHY. JOHANN HEINRICH PESTALOZZI, *Sämtliche Werke und Briefe* (1927– and 1946–), is the scholarly critical edition in German of Pestalozzi's complete works (26 vol.) and letters (12 vol.), the majority of which was hitherto unknown. KATE SILBER, *Pestalozzi: The Man and His Work*, 2nd ed. (1965), containing a select bibliography of texts and literature, is the only comprehensive account of Pestalozzi's life and personality and interpretation of his writings in English. FRIEDRICH DELEKAT, *Johann Heinrich Pestalozzi*, 3rd ed. (1968), is a most recent excellent study on Pestalozzi in German.

(K.Si.)

Pest Control

When man hunted animals and foraged for food he shared the natural resources with other organisms in the community. As man's culture developed and his population rose, he made ever-increasing demands on these resources. In order to satisfy these increasing demands, man has modified his environment with little regard for the biological consequences of his actions.

One result of changing the environment has been a great increase in the number of species that are now recognized as competitors of man. These competitors are usually referred to as pests. The definition of pest is, of course, subjective. An ecologist would not necessarily consider several leaf-eating catepillars on a plant as pests, whereas a gardener who cultivated the plant might very well do so. And only one bat, rat, or mouse is enough to qualify as a household pest. As the term will be used in this article, however, a pest is an animal or plant that occurs in such abundance as to present a distinct threat, economically or medically, to man or his interests. This article is concerned mainly with animal pests; for plant pests see WEED CONTROL.

A pest defined

Natural communities have always contained organisms that were economically significant; locusts, for example, have plagued man throughout history, and grain from Stone Age locations has been found infected with bunt and ergot diseases. Most species that became pests, however, did so because of environmental modification, occasionally from natural causes but usually from the activities of man.

In order to appreciate some of the methods devised to combat pests, it is instructive to consider how advancing technology has increased the number of harmful insect species. The change from natural vegetation to large areas of single-crop (monoculture) agriculture has three consequences. First, given a more uniform plant cover, some plant-eating species increase to large populations. Second, uniform vegetation is easily invaded by attacking pests. Third, the introduction of new crops over large areas results in the transfer of previously harmless insects from scattered native plants to the new and abundant sources of food. Cultural practices such as fertilization, irrigation, and the use of selected varieties as crops enhance still further the ability of pest species to increase rapidly in abundance. Additional reasons for the increase in the number of pest problems include the increasingly higher standards of crop production, and the occurrence of unintended effects in some pest control programs.

Pests are found throughout the plant and animal kingdoms. Weeds are often strong competitors of crop plants. Micro-organisms such as fungi, bacteria, and viruses are here considered with the pests even though they are usually thought of as agents of disease. Most animal pests are invertebrates, among them protozoa, flatworms, nematodes, snails, slugs, insects, and mites. Among the vertebrates, rabbits, elk, deer, and many kinds of rodents are sometimes injurious to crops. Because weeds, plant diseases, and arthropods (insects and mites) constitute the chief threats to the welfare of man, the development of pest control has been directed mostly toward restricting the abundance of these forms.

In spite of the immense efforts expended in controlling pests, their significance for man appears to be increasing. Pesticides have contributed toward the evolution of pests that are increasingly difficult to control and have often caused the replacement of one species of pest by another. Toxic materials have polluted biotic communities throughout the world. As a result the effects of these materials on living organisms has become a matter of major concern.

This article discusses various means, both artificial and natural, that can be used to combat man's competitors and draws attention to their merits and deficiencies.

BACKGROUND

The extent of pest damage. Although the number of species that inhabit the earth is very large, the number that are serious pests is remarkably small, less than 1%, most of which are insects. In spite of the small number of pest species, the losses attributable to insects and mites, plant diseases, and weeds are substantial. One recent estimate of annual worldwide crop loss, expressed as a percentage of the market value of the crop, puts the loss from insect damage at 13.8%, from diseases at 11.6%, and from weeds at 9.5%. The annual regional losses, in percentage of crop market value, are estimated as follows: North and Central America, 28.7%; South America, 33%; Europe, 25%; Africa, 41.6%; Asia, 43.3%; Oceania, 27.9%; U.S.S.R. and China, 29.7%. Annual losses, in millions of tons, for various crops are estimated as follows: cereals, 506; potatoes, 129; sugar beets and sugar cane, 636; vegetables, 78; fruit crops, 56; oil crops, 42; and fibre crops and natural rubber, 8. Added to these losses are the costs of pest control programs.

Economic impact of pests

Insects are also serious pests because some of them play an essential role in the transmission of disease. Each year millions of lives, particularly in the tropics, are threatened by insect-borne diseases. Malaria and yellow fever are transmitted by mosquitoes, plague by fleas, typhus by the human louse, sleeping sickness by tsetse flies, Chagas' disease by blood-sucking bugs, and leishmaniasis by sandflies. Other diseases may be spread by insects accidentally as a result of their habits.

Pest control programs. Pest control began to receive attention in the 18th century, and it has become increasingly important since then. As control programs were developed, they tended to fall primarily into two somewhat mutually exclusive categories: chemical and biological. Physical or mechanical methods, including sticky barriers, heat killing (for storage pests), and flooding (for ground pests), were also developed. Because of their limited utility and short-term effectiveness, they have been superseded largely by chemical and biological methods.

The chemical, or pesticide, approach to pest control probably began with the use of poisonous plant compounds—ground tobacco was used in France to kill aphids about 1763. Other natural products such as nicotine, rotenone, petroleum, kerosene, creosote, and turpentine were used in the 19th century. Inorganic compounds such as Paris green, lime sulfur, Bordeaux mixture, hydrogen cyanide, and lead arsenate were also introduced in the 1800s.

With the appearance of the synthetic organic compounds during World War II, a dramatic change occurred in pest control. Some organic compounds, such as the dinitrophenols, had been used earlier, but the discovery of the insecticidal properties of DDT (dichloro-diphenyl-trichloroethane) and of BHC (benzene hexachloride) made the concept of pest-free crops possible. Research in the 1930s on plant hormones and related compounds led to the development of the selective herbicide 2,4-D (2,4-dichlorophenoxyacetic acid) and this became commercially available about the same time as DDT. Following the appearance of these new synthetic organic compounds, a whole new series of pesticides—insecticides, fungicides, herbicides, and plant growth regulators—was introduced.

Although the ancient Chinese used predacious ants to control foliage-feeding insects, the control of pests by biological means appears to have been initiated in the Western world with the importation of the Indian Mynah bird into Mauritius in 1762 to control the red locust. Control of bedbugs by predatory agents was recommended in 1776. A number of other projects in biological control were carried out in the 1800s, but the modern era in this phase of pest control began in 1888 with the importation into California of the vedalia beetle (*Rodolia cardinalis*) to control the cottony-cushion scale (*Icerya purchasi*). This beetle, imported from Australia, saved the citrus-fruit industry.

Beginnings of biological control

The use of plants resistant to insects was suggested about 1788 as a means of controlling the Hessian fly, a pest of wheat. The classic example of this approach was the control of phylloxera, aphid-like insects that attacked the root stocks of the European wine grape and almost ruined the European wine industry. The solution lay in grafting the European wine grape onto resistant American stocks. The effort, about 1900, to control the lantana shrub in Hawaii

Table 1: Categories of Chemical Pesticides

class	type	example
Insecticides	stomach poisons	
	arsenicals	lead arsenate, Paris green, calcium arsenate
	fluorine compounds	sodium fluoride, cryolite
	contact poisons (natural)	
	plant extracts	pyrethrum, nicotine, rotenone
	hydrocarbons	oil
	(synthetic)	
	chlorinated hydrocarbons	DDT, BHC, lindane, Chlorobenzilate, methoxychlor, and cyclodienes including aldrin, dieldrin, chlordane, heptachlor, endrin, toxaphene
	organophosphates	parathion, malathion, Diazinon, naled, methyl parathion, dichlorvos
	carbamates	carbamyl, methomyl, carbofuran
	hormone mimics	farnesol, sesquiterpenes
	fumigants	hydrogen cyanide, naphthalene, nicotine
	for stored products	methyl bromide
	for soil	chloropicrin, dichloropropane
	systemic poisons	
	plant	demeton, oxydemetonmethyl, dimethoate, phorate, some carbamates
	animal	ronnel
Miticides		Kelthane, Aramite, dicofol
Repellents		dimethyl phthalate, deet
Chemosterilants	alkylating agents	apholate
	antimetabolites	aminopterin
	miscellaneous	triazine herbicides
Antibiotics		streptomycin
Fungicides	inorganic	lime sulfur, Bordeaux mixture, copper sulfate, copper oxide, sulfur, mercuric chloride, cadmium chloride
	organic mercury compounds	phenylmercuric acetate
	dithiocarbamates	ferbam, nabam, maneb, thiram
	chlorine based compounds	folpet, chloranil, dichloran, dichlone
Herbicides	inorganic compounds	sodium arsenite
	metal-organic compounds	cacodylic acid
	carboxylic aromatic	2,4-D; 2,4,5-T; dicamba; endothall; naptalam
	aliphatic acid compounds	dalapon
	heterocyclic nitrogen derivatives	atrazine
	aliphatic organic nitrogen compounds	diuron, barban
Rodenticides	inorganic	thallium sulfate, arsenic trioxide, arsenic sulfide
	botanicals	strychnine
	anticoagulants	coumafuryl, coumafene
	fluorides	sodium fluoroacetate

by introducing an insect appears to have been the first attempt to control weeds biologically. The use of microorganisms to destroy insect pests began in the late 1800s and early 1900s.

With the appearance of the synthetic organic insecticides, however, emphasis was placed on chemical control, and biological means fell into neglect. The extensive use of powerful chemical agents soon resulted in a number of serious ecological problems. Consequently, current pest control practice minimizes the use of pesticides and combines them with biological methods in an approach called integrated control.

CHEMICAL PEST CONTROL

All pesticides act alike in blocking some metabolic process. They differ, however, in composition, potency, mode of action, speed of effect, dosage required, and stage of pest against which they may be used. They are often classified according to the type of organism they are principally intended to control; hence the terms rodenticide, insecticide, acaricide (against mites), nematicide (against roundworms), herbicide (against plants), fungicide, etc. See Table 1.

The chief categories of chemical pesticides. *Insecticides.* Insecticides can be classified on the basis of their chemistry, toxicological action, or principal method of penetration. The method of penetration scheme allows insecticides to be listed as stomach poisons, contact poisons, fumigants, or systemics. The distinction between these latter categories is somewhat arbitrary and a given compound may fall into two or more of them.

Stomach poisons are used against insects with biting mouthparts, such as caterpillars, and are toxic only if ingested. The inorganics, such as the arsenic compounds, are thought to be general protoplasmic poisons, and the fluorine compounds are believed to interfere with enzymes associated with cellular oxidation. This group of compounds has gradually been replaced by the synthetic organic insecticides, which are generally less dangerous to mammals.

Contact poisons penetrate the skin of the pest and are used against those arthropods, such as aphids, that pierce the surface of a plant and suck out juices. The contact insecticides are composed of two main groups: naturally occurring and synthetic organic.

Naturally occurring contact insecticides are nicotine, developed from tobacco; pyrethrum, from flowers of *Chrysanthemum cinerariaefolium*; rotenone, from roots of *Derris* species and related plants; and oils, from petroleum. Although these compounds are originally derived mainly from plant extracts, the toxic agents of some of them (*e.g.*, pyrethrins) have been synthesized. Natural in-

secticides are usually short-lived on plants and cannot provide protection against prolonged invasions. They are effective, however, provided they hit the pests directly. Except for pyrethrum, they have been largely replaced by the newer synthetic organic insecticides.

Synthetic contact insecticides

The main synthetic groups are chlorinated hydrocarbons, organic phosphates (organophosphates), carbamates, and the insect hormones. Other synthetics include thiocyanates, dinitrophenols, sulfonates, sulfides, sulfones, and sulfites.

Chlorinated hydrocarbons were developed following the discovery of the insecticidal properties of DDT. Other examples of this series are BHC, lindane, the cyclodienes, Chlorobenzilate, and methoxychlor. All penetrate the organism readily and are toxic to a wide range of species. Some of these compounds are quite stable and have long residual action; they are, therefore, particularly valuable where protection is required for prolonged periods. In addition to plant protection, they have been of outstanding value in human and animal health fields. Their toxic action is not fully understood but they are known to disrupt the nervous system.

Organophosphates are currently the largest and most versatile class of insecticides. Two widely used compounds in this class are parathion and malathion; others are Diazinon, naled, methyl parathion, and dichlorvos. Usually, they have little residual action and are important, therefore, where residue tolerances limit the choice of insecticides. These compounds are variably hazardous but are generally much more toxic than the chlorinated hydrocarbons. Organophosphates inhibit the enzyme cholinesterase, which is essential in the functioning of the nervous system.

Carbamates form a relatively new group of insecticides that includes such compounds as carbamyl, methomyl, and carbofuran. They are rapidly detoxified and eliminated from animal tissues. Because many of these compounds have a low toxicity to mammals and are selective, they fit well into integrated control programs (see below). The toxicity of the carbamates is thought to arise from a mechanism somewhat similar to that for the organophosphates.

Insect growth hormones have recently been viewed as potential insecticides. Some chemicals, such as the sesquiterpenes and farnesol, have properties similar to such hormones. When they are applied to the target insect at a stage when the natural hormone is absent, they disrupt normal development. These compounds appear to have many advantages, including specificity to particular insect species.

Fumigants such as hydrogen cyanide, naphthalene, nicotine, and methyl bromide are toxic compounds that enter the respiratory system of the pest; they are used mainly for killing insect pests of stored products or nursery stock. Methyl bromide, dichloropropane, and chloropicrin are also used for soil fumigation, principally to control plant parasitic nematodes.

Systemic poisons are absorbed into a plant and later kill sucking insects and mites feeding on it. Absorption into the plant is achieved by spraying the leaves and stems or by applying solutions or granules impregnated with the chemical to the soil, so that intake occurs through the roots. Oxydemetonmethyl, demeton, dimethoate, and phorate are examples of systemic organophosphates. Some carbamates show systemic action. Animal systemics, such as ronnel, are used to control parasites of livestock.

Miticides or acaricides. Because many insecticides also kill mites, the separation of this group of pesticides is somewhat arbitrary. Some compounds, however, such as Aramite and Kelthane are made to be used specifically against mites.

Repellents. Repellents repulse animals rather than kill them. They are used mainly against mosquitoes, blood-sucking flies, and chiggers. Dimethyl phthalate and deet are examples of repellents that are used to protect humans from insect attack. Repellents that can be used for protecting plants from similar insect attacks have not yet been developed.

Chemosterilants. Chemosterilants inhibit reproduction of insect pests rather than killing them directly. Three general types of chemosterilants are usually recognized: alkylating agents (apholate, TEPA, metepa); antimetabolites (aminopterin, amethopterin); and a miscellaneous group of compounds that will induce sterilization but are important primarily for other reasons.

Antibiotics. Although considerable effort has been expended in the development of antibiotics as fungicides and bactericides, they have not been used commercially, largely because they may kill desirable vegetation. Some potential fungicides in this class are streptomycin, cycloheximide, and griseofulvin.

Compounds used as fungicides

Fungicides. Most fungicides are protectants and must be applied before diseases develop although some can arrest the disease if it is already present. Several groups of compounds can be used as fungicides: (1) the inorganics (see Table 1) are still widely used; (2) organic mercury compounds are used for the treatment of seeds, bulbs, and corms; (3) dithiocarbamates can be used as a spray or as a soil drench; (4) chlorine-based compounds are used against a wide range of crop diseases. Systemic fungicides are at an early stage of development and those available are very restricted in use.

Herbicides. Herbicides can be broadly divided into nonselective, which kill all plants, and selective, which kill only certain groups of plants. Herbicides are derived from a wide range of chemical groups: inorganic compounds, metal-organic compounds, carboxylic aromatic compounds, aliphatic acid compounds, heterocyclic nitrogen derivatives, aliphatic organic nitrogen compounds, and others (see WEED CONTROL).

Rodenticides and other vertebrate poisons. Because rodents are probably the chief vertebrate pests, many chemicals have been developed to combat them; however, many rodenticides may be used against other vertebrate pests as well. The highly toxic rodenticides are: (1) the inorganics; (2) the botanicals; (3) the anticoagulants; and (4) the fluorides. The insecticide rotenone is also a potent fish poison. Specific poisons have been developed for certain rats and for the lamprey. Some avicides act as repellents by affecting a few birds, who in turn frighten the others away.

The development, production, and application of chemical pesticides. Although there is a constant demand for new and better pesticides, these are developed only by a time-consuming and costly process. A new chemical is made commercially available for use as an insecticide, for example, by progression through several stages of selection that become increasingly severe and complex. In the laboratory, newly synthesized compounds are tested for biological activity against cultures of standard insect species. They must then be formulated into materials suitable for application as sprays or dusts. Next they must be compared with standard insecticides to establish any unusual properties they may have. Field evaluations are then conducted, first by using small plots and then extensive areas. Finally, if the compound appears promising, a pilot plant is built to provide small amounts for limited testing throughout the area in which the compound is expected to be used. In the meantime, other data must be accumulated relating to residue tolerances, methods of analyzing the compound and its metabolites, and other information required to register the chemical with government agencies. When all requirements for efficiency and safety of the new insecticide have been satisfied and it appears profitable to produce it, a manufacturing plant is constructed and production and marketing are initiated on a major scale.

This process can take from six to eight years. The chance of a compound passing from initial screening to commercial use is very small, as low as 1 in 36,000. The cost of developing a new pesticide may be about $5,000,000, including the cost of those compounds that are rejected.

Pesticide formulations

Pesticides are usually formulated as liquids or solids. In liquid formulations the pesticide may be dissolved in water or in a solvent such as xylene or petroleum oil. In solid formulations the pesticide is prepared as a powder or dust, or as granules. Wettable powders, the most widely used of

the various formulations, consist of the toxicant together with an absorptive carrier such as clay. Dusts are made by mixing the toxicant with a carrier such as a talc, bentonite, or clay. Granular formulations are similar to dusts but they have a larger particle size.

It is necessary to add a number of compounds that, in themselves, are not toxic but that make the toxicant more effective. For example, emulsifiers disperse pesticide solutions insoluble in water, wetting and suspending agents prevent the settling of a wettable powder in water, and stickers improve the retention of spray and dust deposits. Inert carriers aid in obtaining uniform dispersal of small quantities of the active pesticide over large areas. When two pesticides; *e.g.*, an insecticide and a fungicide, are applied together, it is essential that the various additives for each toxicant be compatible. Two toxicants may also act together to produce an effect greater than the sum of each separately.

Applying pesticides

Sprays are applied by hydraulic sprayers or by air-blast sprayers. There are three types of hydraulic sprayers: (1) low-pressure, low-volume sprayers, using pressures up to 100 pounds per square inch, which are suitable for covering plants such as weeds and certain vegetable crops; (2) high-pressure, high-volume sprayers, using pressures of 100–600 pounds per square inch, which are required for thorough coverage of fruit and truck crops; and (3) multipurpose sprayers, which have a relatively wide range of pressures. In air-blast sprayers a blast of air propels the spray and thus large volumes of water are not required. Machines also have been developed to apply the toxicants in a foam. Dusters also use an air stream to carry the pesticide to the plant—particularly to the underside of the leaves. Aerosol fogs and smokes are produced by a pressurized container or by heat-induced volatilization of the pesticides. Aerosol applications are used for treating enclosed spaces and are effective against flying insects. Aircraft, either planes or helicopters, distribute pesticides by using modifications of the low-pressure, low-volume spray, dust, or granular application procedures. For domestic purposes, there is a range of hand-operated sprayers, dusters, and aerosol generators. Animal pesticides may be applied by submerging the animals in solutions held in dipping vats. Soil fumigants are applied by a variety of techniques, each depending on the property of the fumigant being used. Indoor areas may be fumigated with methyl bromide simply by inserting a plastic tube from a cylinder of the gas into the building.

An evaluation of chemical control. *Benefits.* In view of the damage that pests can inflict, it is clear that pesticides are essential for the production of food and fibre. From 1945 to 1965 crop yields have increased by approximately 50 percent, for which pesticides are mainly responsible. Herbicides are estimated to increase yield by 10–15 percent, and fungicides have increased yield by as much as 100 percent. In addition to increasing yield quantitatively, pesticides have increased the quality of crops and the efficiency of crop production and have ensured more stable levels of production.

Insecticides have also been important in improving man's health. They have greatly reduced or eliminated about 30 carriers of disease. Malaria, yellow fever, and typhus have been either eliminated or severely limited in most areas of the world by pesticide use.

Limitations. Although pesticides have proven to be essential in the control of agricultural and medical pests, their extensive use has resulted in a number of serious problems including the development of resistance in pest species, increase in pest abundance, and environmental contamination.

When a formerly susceptible species can no longer be controlled by a pesticide applied at normally recommended rates, it is said to be resistant. Resistance appeared early in the 1900s in scale insects treated with lime sulfur (calcium polysulfides) and hydrogen cyanide, and in 1928 in the codling moth treated with lead arsenate. With the advent of synthetic organic pesticides, resistance increased rapidly; more than 200 species of insects and mites are known to be resistant to some type of insecticide. Rodents, particularly the brown rat and house mouse, have shown resistance to anticoagulants. To oversimplify the situation somewhat, strains that become resistant to one insecticide also may be resistant to a second that has a mode of action similar to the first (cross resistance), or they may become resistant to a number of pesticides with different modes of action when they are exposed to these pesticides simultaneously or in sequence (multiresistance). Resistant strains of a species, for example, the housefly, may arise independently in many areas of the world. Once resistance has developed it tends to persist in the absence of the pesticide for varying periods of time, depending on the type of resistance and the species of pest; this resistance will gradually decline but can be very quickly restored by renewed exposure to the pesticide.

Kinds of resistance

The following may be factors in resistance: body coverings or digestive systems that are not penetrated by poisons, extra fatty tissues that can store pesticides, excretory systems that eliminate pesticides or their toxic metabolites, behaviour patterns that avoid spray deposits, and general vigour. The chief physiological basis of resistance, however, lies in enzymes that detoxify the poisons. In DDT-resistant houseflies, for example, DDT is broken down into nontoxic compounds by an enzyme.

Treatment with wide-spectrum chemicals has often unexpectedly resulted in the rise of surviving populations, which then cause more serious damage than the pest that was eliminated. The majority of pest problems arising from pesticide usage occur from outbreaks of species other than those against which the chemicals were used. In these cases the pesticide eliminates the natural enemies that previously held undesirable species in check. Thus, at times, chemical treatment trades one set of pests for another. This situation arose in the classic case of the control of the cottony-cushion scale (*Icerya purchasi*) by the vedalia beetle (*Rodolia cardinalis*). This predator maintained the pest at a noneconomic status from its introduction in 1888 until about 1947 when the use of DDT on citrus became widespread to control other pests. The vedalia beetle was accidentally eliminated by the DDT, and in its absence the cottony-cushion scale increased to such numbers that complete defoliation occurred in some cases.

In spite of heavy mortality at the time of treatment, use of pesticides sometimes results in a subsequent increase in the density of the pest against which the pesticide was applied. This phenomenon is called flareback, or pest resurgence. Flareback has been recorded for at least 50 species of pests. It can arise for a number of reasons but, in most cases, the pest's natural enemies are eliminated either by the direct effects of the pesticide or by the suppression of the pest itself. In the absence of their natural enemies, the surviving pests can increase their numbers very rapidly.

Pest flareback

Natural enemies of pests can be eliminated from untreated areas if adjacent ones are sprayed. The pesticide sometimes drifts at the time of application so that untreated areas receive dosages often toxic to the natural enemies but not to the pest. Or natural enemies may move through treated areas and become exposed to residual insecticides, thus being "trapped out" and killed.

There are other possible ways in which pesticides may alter the abundance of a species. Although most research shows that the effects of sublethal doses of chemicals are detrimental to pests, such doses appear to have increased the fertility of target species in some instances. Sublethal doses may also reduce the reproduction and development of the natural enemies of pests.

Because pesticides are poisonous compounds, they may adversely affect organisms other than those against which they were originally applied. These hazards to nontarget species or to man can be separated into two classes: the immediate, toxic effects at the time of ingestion or absorption, and the long-term, chronic effects from the gradual assimilation of minute quantities of persistent materials.

Pesticide toxicity varies among the higher animals. Some pesticides, such as the arsenicals, are true biocides, harmful to all living things; many possess varying degrees of

selectivity in insects, birds, and mammals depending on the species and the compound. Some, such as malathion and carbaryl, pose a slight or nonexistent danger to higher terrestrial animals. Pesticides can destroy beneficial insects such as pollinators, as well as pest species. Seeds treated with insecticides have killed birds that consumed them. Some insecticides, such as the organophosphates, are extremely toxic to man and require great care in their manufacture, handling, and application. The ingestion of contaminated food also presents a hazard to man and domestic animals. The protection of food quality is maintained by two means. First, in most countries spray calendars are prepared that indicate the type, amount, and time of application of pesticides for each crop. Second, the maximum permissible amount of residue of each pesticide on various foods has been established in many countries. Some injury to wildlife may occur at the time of application, and incorrect procedures may lead to fatalities among humans, but pesticides cause far fewer direct deaths than many of the other products of advanced technology.

Accumulation of pesticides in the environment poses a much more serious threat to the welfare of man and wildlife. When a pesticide is applied, much of it reaches the soil. Water becomes contaminated from direct application, drift, runoff from treated areas, or waste materials. Minute quantities of pesticides become airborne from volatilization and wind erosion. Many of these chemicals are short-lived or are metabolized by the animals that ingest them. But some are persistent and when applied in large amounts they pervade the environment. The main soil contaminants are arsenic and the chlorinated hydrocarbon insecticides such as DDT, aldrin, dieldrin, heptachlor, and BHC, along with their toxic by-products. The amount that can accumulate is surprisingly large; DDT, for example, has been found in agricultural soils at rates up to 10–100 pounds per acre. Even though they are almost insoluble in water, the persistent organic pesticides are widespread because they are adsorbed to particles of organic matter and become dispersed with them; organic chlorine residues have been found in animal tissues ranging from the Arctic to Antarctic oceans.

Pesticides in food chains

The significance of the persistent pesticides as mortality agents in natural communities is greatly increased because they become associated with food chains. This association has two important consequences. First, some species at the lower end of the chain may be highly susceptible to pesticides and thus the entire chain is destroyed; those species above it lose their food supply. Second, when species at the lower end are resistant, they pass the pesticide to the next species in the chain. This process concentrates the toxicants by a factor of hundreds or perhaps by thousands. Earthworms, being resistant, accumulate DDT without dying but a bird that consumes many such worms may die. Because of the phenomenon of concentration, predators at the top of the food chains are usually the most severely affected.

The usual accumulations of most chlorinated hydrocarbons are harmless to plants and soil micro-organisms, but many organic phosphates, when abundant, reduce plant growth. Minute traces (often measured as parts per thousand million) of chlorinated hydrocarbons can reduce the photosynthesis of marine algae. Marine invertebrates inhabiting coastal waters are particularly susceptible to the organic chlorines; concentrations as low as 0.1 part per thousand million have destroyed mollusks, and other low dosages have killed crabs and shrimp. Many fishes are susceptible to persistent insecticides.

More is known, probably, about the effects of the persistent insecticides on birds than on any other group. Some may die from eating insecticide-treated seeds. Serious population declines or near extinction have been observed in some carnivores at the top of the food chain—the peregrine falcon, bald eagle, osprey, Cooper's hawk, golden eagle, and kestrel. The declines are traceable to a drop in reproduction rather than to the death of adults. Reproductive failures result from delayed breeding, failure to lay eggs, and fragility and breakage of eggs that are laid.

Because the persistent pesticides are so widespread, it is not surprising that they are found in human tissues. The concentration of DDT and its metabolites in fatty tissue varies with person, place, and time. Its average concentration appears to vary from about 4.5 to 8 parts per million for whites and from about 6 to over 16 parts per million for nonwhites. Dieldrin concentrations, on the other hand, are more uniform and much lower (0.04–0.17 parts per million). Lead and arsenic can also accumulate in mammalian tissues; there are 30 atoms of lead for every molecule of DDT circulating in a unit volume of blood.

Because of the clear danger to wildlife and the possible hazard to humans, the use of DDT and other chlorinated hydrocarbons has been restricted by many governments. The amount of DDT used in Canada in 1970, for example, is estimated to be 10 percent of that used in 1967; it is being restricted in the United States, and it has been banned in Sweden. The pattern of legislative restriction continues to change rapidly, however, and a definitive statement cannot be made on this aspect of the pesticide problem.

BIOLOGICAL PEST CONTROL

Although pesticides have proven to be necessary in control programs, they add to the cost of crop production and they give rise to a number of undesirable side effects. Consequently, there has been a renewed interest in previously used nonchemical control methods and in the development of new alternative approaches to pest control in recent years.

The principal methods of biological control. *Cultural control.* Cultural control is concerned with manipulating the environment to make it less favourable for the pests. Because the most vulnerable stage of a pest must be determined, cultural control requires a thorough knowledge of the pest's life history and habits. Some methods are designed to prevent injury to the crop and therefore they must be employed before pest populations become apparent. Others destroy part of the population when the pest is in a susceptible stage of its life cycle. These methods are usually slight modifications in the normal manner of carrying out farming practices. Just as there are many farming practices, there are many cultural methods of control.

Crop rotation

Rotation of plants in cultivated lands or of animals on grazing areas is an important method of limiting pest damage. Crop rotation is the practice of growing a succession of different crops on the same land; it is most effective against pests having limited host range and limited powers of dispersal. Within the rotation there should be some crops that are not suitable for the development of the pest, and the rotation should be sufficiently long so that the pest cannot overwinter for more years than the length of the rotation. Crop rotations can be used to reduce the damage from both insects and disease.

Tillage of the soil at times when crops are not grown will remove weeds, stubble, and volunteer plants that would otherwise provide a reservoir of the pest population and lead to later infestation of newly planted crops. Cultivation will help by either burying pests so deep that they cannot emerge or by bringing them to the surface where they will be destroyed by weather or natural enemies. Clean summer fallowing is often used to reduce plant pests by eliminating their food supply for a season.

Sanitation is an important cultural control method. The destruction of plant refuse often destroys pest populations and so reduces injury to future crops. Many pest populations can develop or overwinter on native vegetation growing on the borders of cultivated fields and later migrate into the cultivated crops. Plowing or burning these areas destroys this source of infestation. A special case arises with alternate host plants. In rust fungi, for example, there is a very complicated life cycle, in which two hosts are necessary. With stem rust of cereals (*Puccinia graminis*) the alternate host is barberry (*Berberis* species) so that the life cycle of this pest may be broken by eradication of the barberry.

Other cultural methods include use of trap crops that are destroyed before the pest can develop, modification of

planting times or spacing, and modification of harvesting techniques.

The use of resistant plant and animal species. Through the normal genetic processes of breeding and selection, man has been able to modify many natural varieties of plants and animals to produce characteristics that he desires—quality of produce, quantity of yield, time of maturing, etc. The same techniques make it possible to develop varieties that also exhibit resistance to attack by pathogens and insects.

There are a number of factors that help plants to avoid attack by fungi. Spores may not find conditions suitable for germination in some plant varieties. If they germinate, they may not be able to secure penetration into the leaf. The pathogen may find growth conditions unsuitable within the host, or the invader's development may result in a response by the host plant to produce structural changes or toxins against the invader.

There are usually three components of resistance to attack by insects: nonpreference, antibiosis, and tolerance. Nonpreference is exhibited by insects that will die before feeding on a plant that otherwise would nourish them. Adults may refuse to oviposit on nonpreferred plant hosts. Nonpreference may also play a role in the resistance of some domestic animals to insects such as flies. Antibiosis is the term used to describe the detrimental effects on an individual pest that result from feeding on a resistant host plant. The results may be death, lowered reproduction, smaller size, or modification of length of life. The tolerance component of resistance operates when the plant withstands populations of the pest that would ordinarily injure a susceptible host. It is found mostly in plants that are attacked by insects with sucking mouthparts, such as aphids, and in which damage results from loss of fluid, or other processes that result in stunted growth.

To develop resistant biotypes of a species, cultures of the pest against which resistance is to be used are maintained. Next, a search for resistance is undertaken by looking for plants or varieties that exhibit little damage under pest attack, compared to common varieties of the species. Because the search for resistance is really a search for genes conferring resistance, it is necessary to prove the genetic basis of the resistant strains by comparing their progeny with standard varieties. Tests for resistance are then carried out in various environments to prove inheritability. When proven resistance is found in combination with other desirable features in a plant, it is then propagated until commercial quantities are available. The whole process may take years to accomplish and is a very complex procedure.

The use of natural enemies. Many pests are species that have been transferred unintentionally to an area from other regions of the world. In their original habitat, they are often harmless, being regulated at low and uneconomic densities by the physical and biotic factors of their environment. Their pest status in the new region might result from the absence of natural enemies that were effective in the original habitat or from the ineffectiveness of native enemies in the new location. An obvious method of pest control consists, therefore, of seeking out the natural enemies of a pest in its original habitat, or in an area similar to the one in which the species is to be controlled, and introducing it into the new area. Attempts are also made to conserve and augment natural enemies to make them more effective. This method of managing natural control is sometimes referred to as biological control. Many types of organisms have been used to combat pests, but the principal ones have been parasites, predators, and pathogens (fungi, bacteria, viruses, protozoa, nematodes), which attack insect and mite pests, and insects, which attack noxious weeds.

Parasites and predators

Insect parasites and predators develop at the expense of the pest population against which they are employed. Insect parasites develop either internally or externally on the host; each successful contact by an ovipositing female usually is sufficient to ensure the death of the individual host and the production of a new parasite in the next generation. Predators, on the other hand, usually consume more than one pest in the course of their development and reproduction. The interactions between natural enemies and the pest population results in self-regulating mechanisms that, when successful from man's point of view, lead to the more or less permanent regulation of the pest population at noneconomic levels of abundance.

Parasites and predators are highly specialized and are especially sensitive to the environmental conditions in which they operate. Probably any natural enemy could regulate pest abundance at low densities if conditions were always ideal. Conditions in the area of introduction, however, are not usually completely identical with the original habitat (where the natural enemy was very effective). The results of this type of biological control are variable: some are completely and permanently successful, some are substantially so, some are partially successful, and the remainder are failures.

When control by parasites and predators is not complete and permanent, it may be possible to make these agents more effective either by conservation or by augmentation. Conservation of already established natural enemies is accomplished by making slight modifications in the environment through the use of suitable insecticides, use of alternate host plants or host insects, diversification of plants and animals, provision of suitable shelters and nesting sites, and adjustment of agricultural practices. Augmentation is necessary when seasonal weather extremes are deleterious to the natural enemies. They can then be recolonized periodically by rearing and releasing them at the proper stage of pest development. They may also be reared in very large numbers, if this can be done cheaply, and used to inundate the area occupied by the pest, just as would be done with a chemical pesticide.

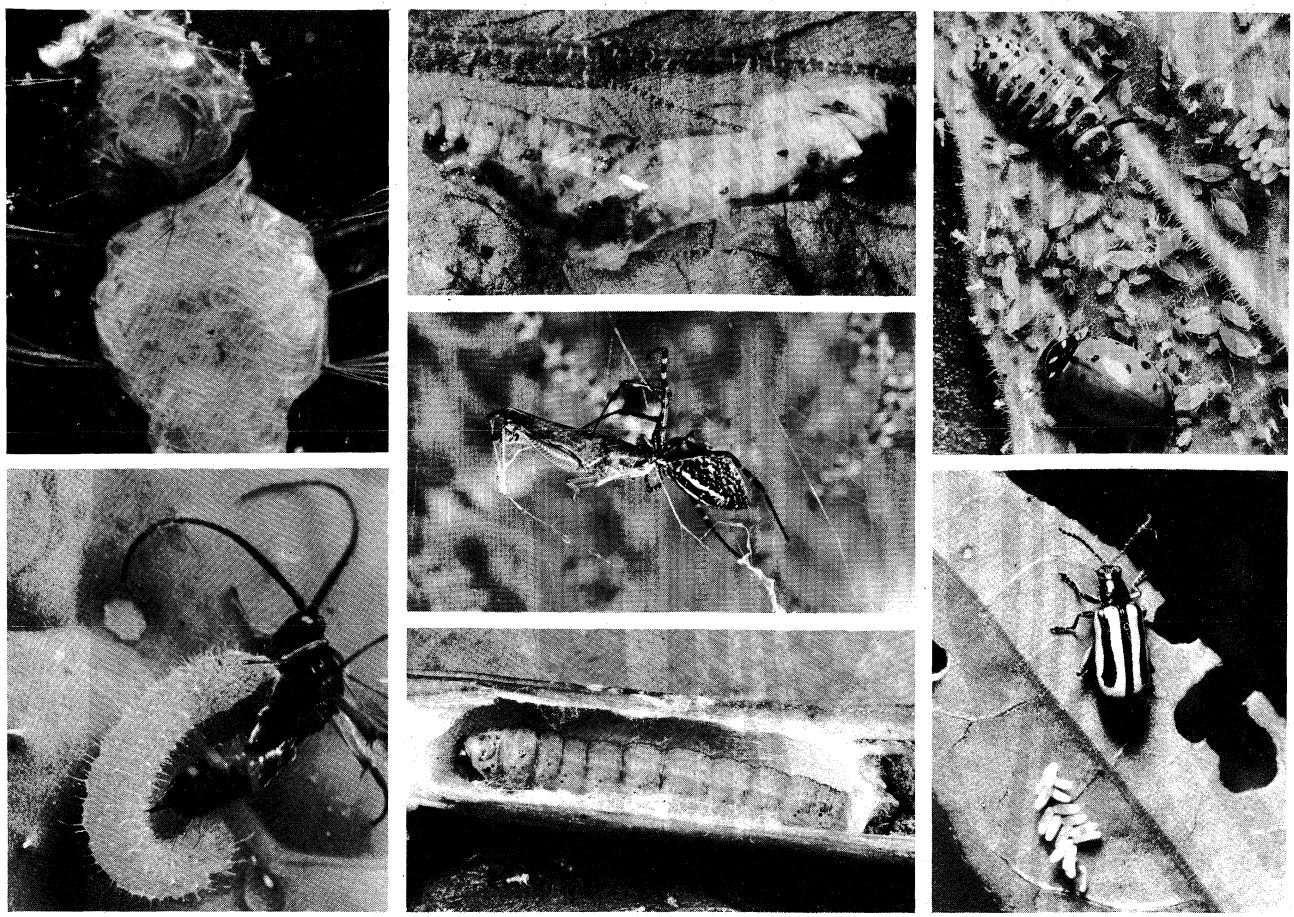
Viruses and bacteria are the most successful and widely used control agents among the more than 1,000 species of micro-organisms found associated with insects. Two important groups of bacteria are the crystalliferous spore-formers and those causing the milky diseases; in the former group, protein crystals toxic to certain insects are formed at the time of spore formation. Like stomach insecticides, both viruses and bacteria must be ingested in order to be effective as pesticides.

In many laboratory studies on competition, one species has been observed to displace and eliminate a second when both were using the same food supply. This same phenomenon can occur in nature. It seems possible, therefore, that man might attempt to eliminate a pest from a given habitat by displacing it with a competitor that, from man's point of view, is relatively harmless. This concept, called competitive displacement, has great potential in a limited number of cases. The larvae of certain blackflies, for example, might be displaced from their aquatic habitats by species that are not so objectionable to man.

Competitive displacement

When crops are grown under intensive cultivation, weeds may be satisfactorily and economically controlled by physical, chemical, and cultural methods. On open rangeland, however, these methods are not practicable, and an alternative method of controlling weeds, particularly those that are imported from another region, is required. Organisms attacking plants can be used to reduce this form of pest to a noneconomic status but not to eradicate it. As with insect pests, the natural enemies of weeds are sought in areas in which they maintain their hosts at noneconomic levels of abundance. The chief natural enemies used for this type of biological control are insects, but other organisms such as mites, vertebrates, microorganisms, and fish (for aquatic weeds) have been used. Insects affect weeds in two ways: either they attack and destroy some vital part of the plant, resulting in its quick or delayed death, or their attack on the plant causes infection by micro-organisms or reduces the weed's competitive advantages over more desirable vegetation.

Probably the most difficult part of a program of managed biological control is deciding what natural enemies are to be used in a particular situation. The identity of the pest must be established precisely, and its enemies that may be scattered throughout the world must be determined.



Biological control of pests.
Microscopic and submicroscopic organisms control the growth of insects by infecting and killing: (top left) the head and thorax of a mosquito larva filled with nematodes; (top centre) the larva of the cabbage looper (*Trichoplusia ni*) killed by a virus. Some insects are predatory or parasitic on others: (top right) adult and larval ladybird beetles feeding on aphids; (centre) a spider eating a grasshopper caught in its web; (bottom left) a wasp (*Apanteles rubecula*) laying eggs in the destructive cabbageworm larva. Insects check the growth of some plants: (bottom centre) the stem-boring caterpillar of the moth *Vogtia malloi* constructing its cocoon in the stem of alligatorweed; (bottom right) the alligatorweed flea beetle (*Agasicles*, adult and eggs) feeding on its leaves.
By courtesy of (top left) Research Institute, Belleville, Ontario, (top right and bottom left) the U.S. Department of Agriculture; photographs, (centre top, centre bottom, bottom right) William E. Ferguson, (centre) Ken Short

Once the required species of natural enemies, and the locations from which they are to be collected, have been determined, the organisms are gathered and sent to the country or region where they are to be used. Because there is a great danger of inadvertently releasing unwanted species, the imported material is always handled under strict quarantine conditions. The natural enemies may be held or reared from their hosts when required and then shipped to the area of pest infestation for release at the appropriate stage of pest development. Large-scale propagation requires mass production of the appropriate hosts, which, in turn, demands extensive laboratory facilities, a detailed knowledge of the biology of all species involved, and development of techniques for handling them. Once the natural enemies are released among the pest population, it is necessary to determine whether they have become established and, if so, what effect they exercise in reducing pest damage.

Insect pathogens

Insect pathogens are produced in a number of ways. Because viruses develop only in living tissues, they can be propagated in laboratory reared insects or from infected insects collected in the field. Artificial diets are used in laboratory rearing, where possible, to increase the production efficiency of the host insects and to limit the possibility of competitive diseases developing—one of the main problems in the propagation of insect pathogens. After the viruses are produced, appropriate dosages must be standardized. Some spore-forming bacteria are propagated in living insects, which are later ground up to yield enormous numbers of bacteria. Bacteria produced on an artificial substance are obtained by drying the medium so they can be held in storage until required. Although some fungi need to be propagated on their hosts, most can be produced on artificial media.

One of the many advantages of the spore-forming bacteria and many viruses is that they can be stored without loss of virulence for long periods before use. They can be applied as a spray or extended with carriers and used as a dust. These pathogens may be used in two ways. First, they can be applied to a local infestation and, if the biology of the pest insect is suitable, they will spread throughout the population; this can lead to permanent control. Second, they may be applied like chemical insecticides, for short-term control, particularly if a relatively high level of insect damage can be sustained by the crop.

The selection, importation, propagation, and release of plant-eating insects to control weeds is, in principle, similar to the use of insect parasites and predators to control pest insects. It is important, however, to conduct detailed safety tests before this type of natural enemy is released, to be certain that only the target plant will be damaged.

Sterilization. The sterility technique in pest control causes an insect to bring about the elimination of its own populations. The basic idea is that sexually sterilized insects will mate freely with individuals in the natural population and thus greatly reduce the numbers of fertile eggs laid. To do this, sterile matings must exceed normal matings considerably to ensure continued population decline in successive generations.

Insects may be sterilized by irradiation or by certain

chemicals. The principal ionizing radiations are X-rays and gamma rays. Radiation induces sterility by destroying or inactivating the sex cells (egg or sperm), or by inducing fatal genetic changes (lethal mutations) in those cells. The action of chemosterilants results in either the failure to produce eggs or the death of eggs and sperm after they are produced, or in the induction of dominant lethal mutations.

The sterility principle can be applied in two ways. First, the pest may be reared in very large numbers, sterilized, and released among the natural population. Second, sterility may be induced in natural populations by using chemosterilants, usually in baits, to sterilize males and females in their native habitat. The use of chemosterilants in nature leads to two types of sterility: that of females that consume the chemosterilant and that of normal females that happen to escape sterilization but mate with sterilized males.

Each pest problem must be investigated in great detail before it is decided that the sterile-insect approach is practicable. The rearing of and the induction of sterility in the pest species must not reduce the competitiveness of the sterilized insects. The treated species must be easily and inexpensively reared. The released individuals must not become pests themselves or increase the extent of the damages inflicted by their normal counterparts. Normal females must not be exceptionally prolific.

Attractants. Most insects respond instinctively to numerous stimuli as they develop into adults and reproduce. Stimuli may control movement, feeding, copulation, and egglaying responses of insects. From the point of view of control, however, interest has centred largely on stimuli produced by chemical attractants.

Two important classes of naturally occurring attractants are the pheromones and the food lures. A pheromone is a chemical that is secreted into the external environment by an animal and elicits specific response in another individual of the same species. The sex pheromones (sometimes called sex attractants or sex lures) cause orientation and approach of a sexual partner from a distance. The amount of material required to elicit response is extremely small. Sex pheromones are usually produced by the female insect but may originate with the male. Sex pheromones have been isolated and identified in a number of insects, including the silkworm, gypsy moth, pink bollworm, cabbage looper, and honeybee queen. A second type of pheromone is the aggregation pheromone. The feces and frass produced by male bark beetles that have been feeding on fir trees, for example, contains a volatile attractant that orients flying populations of males and females to the area.

Food lures

Food lures are generally chemical substances emanating from plants, and they have been identified in many instances. Thus floral scents attract insects seeking nectar, essential oils attract plant-eating insects, scavengers respond to decomposition products, and bloodsucking insects respond to animal products when the levels of moisture and carbon dioxide are high.

Even though natural powerful attractants such as the sex pheromones can be isolated from insects, a great amount of labour yields very little. An obvious solution to this problem has been to make some of them synthetically. Methyleugenol for the Oriental fruit fly; anisylacetone and cue-lure for the melon fly; siglure, medlure, and trimedlure for the Mediterranean fruit fly; gyplure for the gypsy moth; and bombykol for the silkworm are examples of synthetic attractants.

The synthetic and food lures may be used in pest control either as a means of detecting insects or as part of a control program. In survey and detection work, a lure specific for one species permits the determination of the range of a new pest and allows insecticides to be applied only where the species is caught. In control programs, traps baited with attractants may be used to bring insects into contact with insecticides, pathogens, or chemosterilants.

Other methods. An important aspect of control is the prevention of entry of a pest species into a country or area where it does not already exist. This is difficult to do in view of the great expansion in commercial traffic; nevertheless, quarantine procedures—*i.e.*, the interception of potential pests by inspection at points of entry—have greatly restricted the spread of noxious insects, diseases, and weeds.

An evaluation of biological control. *Benefits.* The attractiveness of the biological approach to pest control arises from the fact that each individual technique is developed to control one or a few related types of pest. Its use reduces the amount and distribution of pesticides, particularly the persistent ones, in the environment. These methods, therefore, reduce the risk of damage to nontarget organisms and to man, avoid the development of resistant strains of pests, and permit the normal ecological processes to operate much more freely. The cost of pest control may be greatly reduced in those cases in which the biological agents remain effective over a period of years.

The procedures of cultural control are associated often with normal farm practices and can be performed routinely before a pest problem arises. They are, therefore, both effective and inexpensive.

Selective breeding of plant resistance probably has been the most successful of all the biological means of controlling insect and disease damage. For insect pests, effects of resistant crop varieties are both specific and cumulative; resistance need not be complete because its persistence will ensure the elimination of the pest within a few generations. The cost of using resistant varieties is quite small and they can be used along with other measures. This compatibility with other measures permits use of varieties with a relatively low level of resistance that possess some tolerance to attack. The use of a resistant variety of one crop may reduce the population level of the same insect on a second crop. Because of their more normal growth, resistant plants often show an improvement in yield and quality. Strains of alfalfa, cotton, tobacco, beans, potatoes, and tomatoes have been bred successfully for resistance to specific diseases caused by fungi, bacteria, viruses, and nematodes. Resistance to the wheat stem sawfly (*Cephus cinctus*) and to the Hessian fly (*Mayetiola destructor*) has been developed in wheats, and resistance to leafhoppers has been produced in cotton.

More than 100 species of pests are either completely, substantially, or partially controlled by their natural enemies in over 60 countries throughout the world. At least 650 species of insect parasites and predators have been imported into the United States, and about 100 of them have become established and exert some control over about 50 species of pests. In Canada about a dozen species of pests are successfully controlled by this means.

Economic benefits

The economic benefits derived from the use of insect parasites and predators have been substantial. In California the net savings from major projects in biological control have been estimated at $110,000,000 over a 36-year period. The control of the coconut scale (*Aspidiotus destructor*) on the island of Príncipe gave a return over cost of 1,800 percent per annum. Control of the sugarcane borers (*Diatraea* species) gave returns over cost ranging from 200 to 25,000 percent per annum. Use of the vedalia beetle against the cottony-cushion scale in St. Kitts, Nevis, and Montserrat resulted in returns of about 450 percent per annum.

The control of the Japanese beetle by the bacterium *Bacillus popilliae* and of many moths by the crystalliferous bacterium *Bacillus thuringiensis* are examples of the successful use of the bacterial diseases. At least 20 cases involving control of insect pests by viruses have been recorded. Microbial pathogens are usually specific and therefore represent little danger to man, animals, or beneficial insects; and susceptible hosts are very slow to develop resistance.

The classic example of the "rear-and-release" technique, using radiation for sterilization, is the eradication, in the southern areas of the United States, of the screwworm (*Cochliomyia hominivorax*), a parasite of cattle. After eradication, barriers to prevent re-entry of this pest are maintained by periodic releases. More or less extensive field trials on additional species have been conducted; these include the Mediterranean fruit fly (*Ceratitis capita-*

ta), the Oriental fruit fly (*Dacus dorsalis*), the melon fly (*Dacus cucurbitae*), and the codling moth (*Carpocapsa pomonella*). Some limited field experiments have also been conducted using chemicals as the sterilizing agent.

Attractants in traps are used annually to survey for infestations of the Japanese beetle and the gypsy moth. An example of attractant use in a control program is the eradication of the Oriental fruit fly from the Pacific island of Rota; fibreboard squares, saturated with the attractant methyleugenol and the insecticide naled, were distributed in 15 biweekly treatments. The number of males removed from the population was so great that females were unable to reproduce.

Limitations. The great advantage of biological control methods—their specificity—is also a source of weakness. Usually several species of pests attack a given host plant. If biological methods can deal satisfactorily with perhaps three out of five pest species, two pests will still be free of restraint. In this case, it will be necessary to use a pesticide that could have been used initially to control all five species.

Biological methods may be very expensive, especially if they have to be repeated periodically. In the case of resistant varieties, for example, the time required to develop them, if they are not available when a new pest problem arises, may be excessive. For a wheat variety it may require 15–20 years and for alfalfa, 3–5 years. The development of resistant varieties may become an endless process because of the ability of pest species to change into new forms (biotypes) capable of attacking a previously resistant variety. The development of new biotypes appears to occur more rapidly among the pathogens, such as wheat rust, than among the insects.

Limitations of natural enemies

Although insect parasites and predators have been effective against a number of important pests, their use presents some difficulties. Many attempts to use them have been failures: whereas only 95 species of parasites and predators were successfully established in the continental United States between 1875 and 1951, about 390 other species failed to become established. The population dynamics of parasites and predators are dependent on the environment, so that a natural enemy that is effective in one region or during one year may not be in another. Even when successful, the growth of many parasite and predator populations lags behind that of the pest populations, so that damage may be inflicted before the natural enemies can become effective controlling agents. Usually, parasites and predators are more effective against pests attacking perennial plants than against pests of annual plants such as truck crops. Many parasites and predators are themselves attacked by carnivores.

There are also some difficulties attending the use of microbial control methods. The timing of the application must be correct. Usually, the pathogen is effective against only one pest out of all that are present. The pathogen must remain viable between the time of application and its actual contact with the insect pest; in some cases, its viability can be affected by the physical environment. Some pathogens are difficult to produce in large quantities. Even when microbial control methods succeed, other difficulties related to those methods may arise. For example, diseased insects that remain attached to crops are often objectionable.

In the development of pest control methods, the use of chemicals and biotic agents tended to be in conflict with each other. The problems that followed the introduction of the organic pesticides, however, led to the conclusion that both methods must be combined. Integrated control, as this approach is called, was, therefore, originally concerned with developing procedures in which chemicals and natural enemies could be used compatibly.

INTEGRATED PEST CONTROL

The concept of pest management. The term pest management is a summary definition of integrated control practices. Pest management implies both the manipulation of the populations to maintain them at noneconomic levels of abundance and the eradication of a pest species when this is attainable and desirable. Both control strategies are based on a thorough knowledge of the pest's biology and ecology before a control program is undertaken.

In developing an integrated program there are two groups of facts that must be determined in some detail. These are the structure of the agricultural ecosystem and the population densities of the pests below which little economic damage will occur.

Agricultural ecosystems. If pest control depends on manipulation of pest populations, it becomes necessary to develop techniques for analyzing the multitude of factors that influence pest density. The identification of the processes that determine community structure is one of the primary problems of modern ecology. The magnitude of the task may be appreciated from the example of Wytham Woods, where an area of less than two square miles of English countryside contains at least 2,500 species of animals.

Agricultural ecosystems, however, are usually based on one or a few species of plants and so are relatively simple compared with their natural counterparts. This simplicity, of course, is only relative: even with a single crop, alfalfa, about 1,000 species of arthropods have been found in association with it. An agricultural ecosystem comprises all of the organisms in a given farm location, together with the physical environment that prevails there. Usually, precise geographic limitations cannot be placed on the extent of such an ecosystem, but a sufficiently large area must be considered so that all of the biotic elements having an effect on pest density are capable of being evaluated. Thus the area surrounding a group of cultivated fields cannot be ignored, for it may harbour reservoirs of pests and their natural enemies. If some of these species are wide-ranging, their biology may require investigation.

Detailing quantitatively all of the factors that determine the abundance of each species in a simplified agricultural ecosystem is a formidable undertaking. In most agricultural and forestry habitats, however, there are only a few species of insects that are pests and often there is only one of major significance. It is thus possible to consider only one or a few subsystems that affect the target species in a significant manner. Once the major pest has been dealt with, consideration may be turned to other pests if necessary.

Economic thresholds. An integrated program requires a determination of the density of each pest that will result in economic damage to the crop. This is a difficult task because it involves estimating a series of factors, some biological and some economic.

Extent of damage inflicted depends, in part, on the type of pest. Major pests are serious, continuous, and persistent; they dominate control practices because they remain above economically damaging levels of abundance. Few insects in agricultural ecosystems are major pests; of the approximately 100 potential pest species in orchards, for instance, only about five can be considered major. Occasional pests cause damage only at certain places or times; they are suitable objects for integrated control because they permit corrective treatment when necessary rather than continual preventive measures.

Evaluating pest damage

Many biological factors may alter the evaluation of pest density in relation to pest damage. The susceptibility of a given plant to attack can be modified by variation in the physical environment, stage of development of the plant, irrigation, fertilization, and variety of crop. The type of injury that is sustained by the plant is quite important in determining to what extent a pest may be tolerated before economic injury results. Many plants can withstand some loss of leaves, stems, roots, or even fruits before appreciable damage is inflicted. But when pests attack and blemish harvestable crops or are carriers of diseases, even small numbers of them can inflict severe damage.

One important consequence of the concept of an economic threshold of pest damage is the acceptance by growers of noneconomic levels of pest populations in the agricultural ecosystem. The functioning of an integrated control program depends on the presence of at least some pest population as insurance that the natural enemy of

that pest will continue to be maintained. Integrated programs differ in this respect from programs that attempt to eradicate or maintain pests at the minimum densities attainable.

The combination methods of integrated control. Pesticides are, without doubt, the main tools in pest control, and an integrated program attempts to minimize their disruptive effects on the agricultural ecosystem by making them as selective as possible. Unfortunately, pesticides are not very selective, physiologically, among the arthropods, for they usually kill broad groups of organisms—including many species of natural enemies, which are generally more susceptible to pesticides than the pest species themselves.

Selectivity of even the broad-spectrum insecticides may be increased by the manner in which they are used. A limited area of pest infestation can be treated. A degree of selectivity can be achieved by applying pesticides at a time when beneficial insects are elsewhere or are protected. The success of timing can be greatly enhanced through use of pesticides with short residual action. Selectivity can be increased by lowering the dosage of a pesticide; at low dosages, some insecticides can decrease pest abundance but are not detrimental to all of the pest's natural enemies. Systemic insecticides absorbed by the plant and translocated to the leaves are effective against plant feeders such as aphids, mites, and thrips but are not usually harmful to their natural enemies. The use of attractants and seed treatments are means of localizing the distribution of pesticides.

Encouraging natural enemies of pests

Natural enemies should be fostered as much as possible, and the use of selective pesticides, when available, is an important means of accomplishing this. Food for natural enemies has been provided, in some cases, by maintaining an acceptable population level of pests in special plantings of alternate crops that support the pest or other hosts. Thus the supply of natural enemies is maintained and is available when needed. In other cases a host or an alternate host will develop on weeds or cover crops. Strip cutting of alfalfa in California allowed both the pea aphid and its parasite to persist through unfavourable hot, dry, summer periods. Provision of nesting sites also increases the density of beneficial species such as birds and wasps.

In addition to pesticide selectivity and enhancement of natural enemies, other components may be added to an integrated program. Resistant varieties may be used to augment control by natural enemies, which may not be completely effective alone. Insecticides can be combined with microbial agents such as viruses and *Bacillus thuringiensis*; when this can be done, dosages of each agent can be reduced. Chemicals can be combined with attractants to reduce populations yet limit pesticide distribution. The sterile-insect method works best when pest densities are low; an initial application of an insecticide can produce the necessary low densities.

An evaluation of integrated control. *Benefits.* Outstanding examples of successful integrated programs have been developed in three areas of the world.

In Canada control of apple and pear pests was accomplished through use of a modified spray program. At one time, over 80 percent of apple and pear acreage in Nova Scotia was under this type of program. In Quebec apple orchards, a modified spray program gave as clean a crop for approximately half the cost as that obtained with the normal spray program.

In California the spotted alfalfa aphid (*Therioaphis maculata*) was a pest until biological control was augmented by a selective chemical control program. After the successful integration of chemicals and natural enemies, the development of aphid-resistant varieties reduced the status of the aphid to minor proportions.

In the Cañete valley of Peru the production of cotton was seriously threatened while broad-spectrum insecticides were used in the control program. The insecticides were eliminating natural enemies and developing resistance in pest species. In the face of disaster, a series of changes in control programs, including the adoption of arsenates and nicotine sulfate, was made. The new program resulted in the return of many pests to their former harmless state, the diminution of key pest problems, an increase in cotton yields, and reduction of control costs.

Limitations. In spite of the great progress that has been made in developing integrated control programs, the approach has some limitations and is not always possible to apply. Since there are so many variable factors, in any given situation one or more of the central components of integrated control may be invalidated by specific circumstances; for example, natural enemies are not effective in all crops. Or a pest may be so difficult to control as to require a highly toxic broad-spectrum insecticide. Although a pest may be susceptible to low dosages of a highly selective insecticide, residues of this compound may be very dangerous to other organisms, including man, and therefore a more broadly toxic chemical with shorter residual effect must be used. Integrated programs are successful only when supervised by highly qualified specialists, and the techniques employed may sometimes be more costly than normal control programs. When control cannot be obtained, the best solution is to grow crops less susceptible to pest infestation.

Integrated control limitations

PRODUCTION AND CONSUMPTION OF PESTICIDES

The United States produces from 50 to 75 percent of the world's pesticides, an appreciable portion of which is exported. As other countries develop manufacturing facilities, this percentage is expected to decline.

The use of pesticides throughout the world is indicated in Table 2. Pesticides tend to be used against pests attacking the large monocultures; in the early 1970s in the United States about two-thirds of all insecticides, for example, were applied on three crops: cotton, corn, and apples. Although use of DDT in American agriculture is declining, this pesticide is still used extensively in foreign malaria eradication programs.

Table 2: Estimated Quantities of Biocides Used in or Sold to Agriculture, 1967
(active ingredients, in tons)

biocide	U.S.	Canada	Italy	Egypt	India	Japan
Pesticides						
DDT	20,128	842	2,007	2,506	3,165	794
Benzene hexachloride	...	40	4,901	...	25,637	3,468
Aldrin	...	216	4,467	...	337	253
Endrin	43,144	33	37	573	827	140
Organophosphates	54,660	934	4,736	2,390	2,340	2,516
Arsenicals	4,240	218	627	3	...	766
Oils	77	2,752	11,381	3,580	...	1,020
Botanicals	173	1	202	...	...	106
Fungicides						
Copper compounds	16,996	213	29,588	172	5,269	8,851
Dithio-carbamates	3,373	2,080	15,418	...	1,028	2,898
Fumigants	12,452	...	3,373	...	445	10,299
Herbicides	15,000	12,790	4,321	106	587	27,589

Source: *Production Yearbook*, vol. 22, Food and Agriculture Organization, United Nations, 1967 and 1968.

The production and use of pesticides will, in all probability, increase even more rapidly in the future than it has in the past. In the United States production of synthetic organic pesticides is expected to increase at an annual rate of 15 percent and to reach sales of $3,000,000,000 by 1975. Use of insecticides is expected to at least double by 1975, with sales reaching $600,000,000.

Although some insect parasites and predators, and microbial pathogens such as *Bacillus thuringiensis* are available commercially, the development and execution of biological control programs has, until now, been largely the responsibility of government institutions.

BIBLIOGRAPHY. C.S. ELTON, *The Ecology of Invasions by Animals and Plants* (1958), a treatment of the movement of pests into areas of food production; H.H. CRAMER, *Plant Protection and World Crop Production* (1967), a book about the effects of pests on food production; A.W.A. BROWN, *Insect Control by Chemicals* (1951), and H. MARTIN, *The Scientific Principles of Crop Protection* (1964), important textbooks on the chemistry, use, and effects of pesticides; C.M. WILLIAMS, "Third Generation Pesticides," *Scient. Am.*, 217: 13–17 (1967), an article discussing the value of hormones in

pest control; F.A. GUNTHER and L.R. JEPPSON, *Modern Insecticides and World Food Production* (1960), a treatment of the chemistry, toxicology, and application of pesticides; R. CARSON, *Silent Spring* (1962), a popular and important account of the dangers of improper pesticide use; U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE, *Report of the Secretary's Commission on Pesticides and their Relationship to Environmental Health*, 2 pt. (1969), an account of the effects of pesticides on man and his environment; D.B. PEAKALL, "Pesticides and the Reproduction of Birds," *Scient. Am.*, 222:72–78 (1970), a good recent article treating one of the side effects of pesticide use; G. ORDISH, *Biological Methods in Crop Pest Control* (1967), a treatment of nonchemical pest control techniques; P. DEBACH and E.T. SCHLINGER (eds.), *Biological Control of Insect Pests and Weeds* (1964), an account of nonchemical pest control methods; W.W. KILGORE and R.L. DOUTT (eds.), *Pest Control* (1967), good accounts of nonchemical pest control methods; H. MARTIN (ed.), *Pesticide Manual*, 3rd ed. (1969); J.H. STAPELY and P.C.H. GAYNER, *World Crop Protection* (1969); A. MALLIS, *Handbook of Pest Control*, 3rd ed. (1960); P.P. PIRONE, B.O. DODGE, and H.W. RICKETT, *Diseases and Pests of Ornamental Plants*, 4th ed. (1970), an excellent work on the pest problems of house, landscaping, and other horticultural plants; C.L. METCALF and W.P. FLINT, *Destructive and Useful Insects*, 4th ed. (1962); G.W. IRVING, "Agricultural Pest Control and the Environment," *Science*, 168:1419–1424 (1970), a general review of modern pest control technology; NATIONAL ACADEMY OF SCIENCES, *Principles of Plant and Animal Pest Control*, 6 vol. (1968–69), an extensive work on pests and their management, including volumes on weeds, plant diseases, insects, nematodes, and vertebrate pests.

(T.B.)

Pet

Generally, a pet is an animal tamed and kept as a favourite or treated with fondness; however, the range of animals kept by modern man is so wide that the words tamed and fondness must be qualified.

Most pets have been selected from animals that were domesticated early in man's history and have entered into a special, and more or less permanent, relationship with man. They differ from the original forms not only in size, colour, and shape but also in temperament. This inherited capacity to live cooperatively with man may be only potential in some animals, as with horses, which must be individually tamed and trained; others—in particular, the cat—are usually prepared to accept at once their place in the household. In contrast, numerous wild animals are brought into captivity and spend the rest of their lives as captives, some more willingly than others by reason of their temperament or the patience of their masters.

A pet is kept for its own sake and for the pleasure that it gives its owner. For a wide range of higher animals, especially horses, dogs, cats, and primates, this pleasure appears to be mutual. Pet keeping, therefore, is essentially a symbiotic, or mutually beneficial, animal–human relationship. Keeping pets satisfies a deep, universal human need, and pets are found at every cultural level.

ORIGIN AND HISTORY OF PETS

The dog as the first pet

The dog, which became domesticated long before any other species, must have been the first true pet. A working relationship probably formed between early man and the wolves that lived around his settlements and gradually evolved until certain individuals became tame and were then regarded as settlement or household dependents as well as partners in hunting and guarding duties. True domestication slowly followed, and the dog had arrived; later came the deliberate selection of a range of breeds illustrated in many ancient carvings and paintings. In Mesopotamia the giant mastiffs were used in the lion hunt; in Egypt the tomb paintings depict greyhound-like dogs and short-legged terriers resembling those of today. In some periods of history man seems to have had a special affinity for animals; this was markedly true in the 12th dynasty in Egypt (1991–1786 BC) when hyenas and lions were tamed and trained and the cat began to be domesticated. This faculty for keeping animals was gradually lost, to be revived some 200 years later, in the time of Queen Hatshepsut.

Of classical references to pet keeping, two in the Bible are well known. In II Sam. 12 Nathan speaks of a ewe lamb brought up in the family of a poor farmer, a practice that is still common. Because of their scavenging habits, house dogs were not kept by the Jews, but in Mark 7:28 a foreign woman spoke of her little pet dogs that sat under the table.

THE RELATIONSHIP OF PETS TO MAN

Companionship. A wide range of pet–owner relationships is found, the worst of which is little but exploitation. An extreme example of this is the overfed, underexercised dog that has been taught no discipline and whose life contains little of interest or excitement; the selfish owner may keep it alive long after its existence becomes a burden because he or she cannot face losing it. Equally unsatisfactory is a relationship so loose that an animal hardly regards itself as attached to a family, which may feed it but takes little other interest in it.

The emotional relationship between owner and pet is not easy to define, but it seems that most human beings have a latent capacity to seek contact with animals, mostly but not only with furred animals. Modern experience in the pet trade and in zoos suggests that this desire is heightened by living in urban surroundings.

The instructive value of pets

Keeping pets offers the opportunity to teach children the close dependence of privilege on responsibility and also something about sex; the process of mating is soon noticed, followed by such matters as gestation periods and the varied problems involved in the birth and care of young.

Almost by definition, pets kept as companions must also give pleasure and satisfaction. In addition, many animals are kept as personal pets, in the broader sense, to which the description "companion" cannot possibly be given. They are scarcely aware of any relationship with man, for they live in reasonable replicas of their habitats. All invertebrates fall into this class, and, because of their different physiological needs, all fishes and amphibians. This indifference to man is generally true of reptiles as well, but a few snakes, especially captive pythons, appear to seek and enjoy the proximity of humans.

A contrast exists between mammals and birds of domesticated stock and those whose immediate ancestors are truly wild. With the former a more satisfying, complete, and lasting relationship is likely. Such is possible in the latter groups also, but the term companion can be applied properly only to the exceptions. In such cases, which may occur in widely varying groups of mammals and birds, the human, in effect, replaces the mate or, at least, a companion of the same species. Many such animals revert to the wild when they become adult, a pattern particularly true of the large apes. In others, when a mate is accepted, the temperament changes radically, and they can no longer be handled; if the introduction is not made, however, until one or other has long been adult, mate acceptance is unlikely. Moreover, if both animals have been tame, the chance of mating is small and of breeding even less; this is the most likely explanation of the failure to mate many long-term inmates of zoos. The rather vague term humanized is often applied to such animals. From the human side there is some support for this substitute theory; the most extravagant examples of lavish coddling, especially of cats or toy dogs, usually involve unmarried or childless women, often in later life.

Partnership. The relationship between an animal trainer and his animals is a special one, for they form a team to earn money; yet the bond joining them is close, and, although these animals often live in cramped quarters and under artificial conditions, many seem to enjoy a more interesting and satisfying life than the average household pet. This does not constitute an endorsement of all that goes on, especially in the training, but it is clear that most trainers have a deep concern for their animals, which can truly be described as their pets. They are indeed tame animals, whether of wild or domestic stock, in which obedience training has been pursued to an extreme. The training consists largely of noting what individual animals enjoy doing, and developing that habit in an interesting way.

Scientific observation. The margin between keeping pets for their own sake and as objects for serious study is narrow. The wild animals forming the subject of many books often come into human care by accident. The immediate objective, in such instances, is a humanitarian one, providing its own unique satisfaction; this phase is followed by the pleasure of companionship and then the recording of valuable facts about animals hard to observe closely in the wild.

Pets in studies of animal behaviour

It is doubtful whether any animals used solely for scientific study can truly qualify as pets. Many are subjected to observations and experiments and are unaware that they are being studied (in many cases the development of a personal relationship would invalidate the results).

Some aspects of animal behaviour are best studied when the subjects are living free, yet secure and approachable because a special relationship has been built up with the observer. Foremost in this field of ethology is the German zoologist Konrad Lorenz, whose pioneer work is recorded in several popular books (see bibliography). His deep understanding of animal psychology made a family of jackdaws regard him as one of themselves and a flock of greylag geese take him as their leader. Other persons have come to the same terms with wolves, chimpanzees, and gorillas.

KINDS OF PETS AND THEIR NEEDS

Basic requirements. It is important to realize that human standards are not of much use in determining the best way to meet the varied needs of the different pets kept by man. This is best illustrated by mentioning certain basic needs for all kinds of pets. Variety in food is largely an invention of civilized man. For most animals, and also for man in many parts of the world, change comes only with the season. A little variety may be useful in keeping some sensitive higher animals interested in food, but the first requirement is a balanced diet appropriate to each species, adjusted when necessary to individual need or whim. Overfeeding is never good; as with humans, it often results in obesity, lethargy, and shortening of the life-span.

Correct design of pet quarters is usually more important than the total area enclosed. Any group of animals, tame or wild, includes bullies and the bullied; a good plan lets the latter keep safely out of the former's way. Many nocturnal pets find security by day in dark holes that other animals would regard with mistrust, and they may be badly upset if disturbed. A dog's kennel gives him a safe feeling; with his rear protected he can be watchful of the area before him.

Some animals carefully deposit waste products away from their immediate living quarters. Others, however, including lorises and bush babies, put urine on their feet to mark their territory; such an odorous condition is necessary to give a sense of security.

Long walks are desirable for some pet animals, such as dogs, whose makeup has been modified by generations of living with man, but few adult wild animals take exercise for its own sake; they run hard when they must—to kill or avoid being killed.

Proper handling is basic in establishing human–pet relationships; some animals enjoy much physical contact, while others dislike even being touched.

Few pets are kept whose needs are not well known; careful study of these should show whether or not an animal will fit into a particular household. Most pets cannot fend for themselves, even for a weekend, and any decision about keeping an animal must take note of the degree of care it requires.

Household pets. *Dog.* The dog is the most variable domestic animal, with more than 100 breeds, ranging in size from two to four pounds (one to two kilograms) for the Chihuahua to over 200 pounds (100 kilograms) for the St. Bernard and the mastiff. Many crossbred, or mongrel, dogs make delightful pets, but their adult size and shape are often unpredictable, and, contrary to popular belief, they are not necessarily of better temperament than a pedigreed dog. As a general rule, a dog becomes attached to people, a cat to premises (see DOG).

Cat. Partly because of its independent nature and nocturnal wanderings, the cat has comparatively few breeds. Shape and size vary rather little, the differences being mainly in coat length and colour. All cats make good pets, but long-haired breeds should be kept only in homes where they will be groomed daily (see CAT).

Birds. First kept widely in Europe in the 16th century, the canary is still a favourite singing pet. There is much variation of plumage in the show breeds, in which singing is less important. The canary is ideal for a room pet. The wild budgerigar, from Australia, is green, but variants have been selected in many shades and combinations of greens, blues, and yellows, with white and gray. Since they are clean, easily looked after, and cheap to feed, budgerigars (often called parrakeets) make excellent pets, especially for older people living alone. Most individuals of both sexes can be taught to repeat simple phrases. Of the 300 or so parrot species seen in aviaries, the gray parrot of central Africa and the Amazon parrots of South America are the only ones commonly kept as house pets. They are, however, expensive and often noisy, and their powerful beaks make them destructive and capable of severe bites. A member of the starling family, the mynah, is one of the best talking birds. It requires large quantities of soft fruit and may often be messy in its habits.

Other birds kept as house pets include pigeons and doves. There are many varieties of these birds, some of them highly ornamental; they are often kept at liberty, with access to shelter but free to fly in the neighbourhood, where they unfortunately may cause damage.

Poultry and waterfowl are kept primarily for commercial and ornamental purposes, but a few individuals—notably among ducks, geese, and chickens—become unusually tame and are adopted as members of the household.

Vivarium pets. When kept as pets, reptiles and amphibians frequently require special conditions of heat and humidity. For this reason they are best kept in glassed enclosures called vivaria. These animals have no internal temperature regulation and in nature their activity varies with the season. Species of temperate climates can do well at room temperatures and, in the case of some, even outdoors, but they must be allowed to hibernate as winter approaches. The vivaria can be put outdoors in autumn and filled with leaf mold or light soil; when the weather has turned cold, they can be taken in and stored in a cold but frost-free location. Alternatively, early in autumn the animals can be released in suitable places to hibernate.

Tropical snakes, lizards, turtles, frogs, and toads require special treatment. These animals are imported from warm countries, and all are used to being active throughout the year, though not always in the same degree. They need warm quarters at all times, preferably with day and night variation that simulates their native environment. Furthermore, for many people a big difficulty is that snakes, lizards, frogs, and toads eat live animal food. Only experts should keep venomous snakes.

Aquarium pets. Fishes constitute a completely separate section of the pet world, and a worldwide industry exists for catching, breeding, transporting, and supplying stock, as well as for making a great range of equipment and formulating special food. Fishes can be divided into two main groups: cold-water and tropical. Each of these categories may be divided further into freshwater and marine, the largest being the freshwater tropical. Many hundreds of species are available, and there is a rich literature on the subject, with many specialist magazines.

Insect pets. Although enthusiasts in many countries study live insects as a hobby, only a few kinds of insects are of widespread general acceptance as house pets. Stick insects are the easiest to keep; they can be fed privet leaves and do well in simple containers at room temperature. The larvae of silk moths, the semidomesticated form of which yields real silk, can be fed on mulberry or lettuce leaves; they grow to full size before pupating and emerging the following year as silk moths.

It is possible to establish colonies of several kinds of ants in artificial nests in which they can easily be observed under glass.

Hutch pets. Of the common hutch or cage pets, rabbits and guinea pigs perhaps do better when kept in a protected situation out-of-doors, while rats, mice, hamsters, and gerbils thrive under room temperature conditions.

Rabbit. Rabbits are available in many sizes—from the 2½-pound (1-kilogram) Netherland dwarf to the New Zealand and Flemish giant of 12 pounds (5½ kilograms)—and in many colours and patterns, with coats of various lengths and textures. The small and medium sizes are best as pets (the larger are often raised for fur and meat). Rabbits are heavy drinkers and feeders, taking large amounts of roots, leafy shoots, grain products, and hay (modern feeding is based on high protein dried grass pellets and hay).

Guinea pig. There is little variation in size or shape of guinea pigs, but a wide range of colours and three types of coat—short, rosette (Abyssinian), and long (Peruvian). Guinea pigs feed on the same food as rabbits, but not so heavily. They are notably prolific, bearing litter upon litter of young that are fully mobile at birth.

Rat and mouse. Although used mainly in research work, the tame rat, usually white or hooded, makes an excellent and docile pet. It has clean habits and is without a distinctive mousy smell. It tolerates a range of indoor temperatures and is an ideal children's pet that is not easily hurt by handling. Tame mice, domesticated from the house mouse, have certain disadvantages. The adult male's smell is regarded as unpleasant by some. Mice breed rapidly and their small size makes them hard to handle and confine.

Hamster. This small rodent, first tamed from the wild golden variety about 1930, is now available in many colours. The hamster has largely replaced the mouse as a house pet because of its clean habits, lack of odour, and convenient size (six inches long, or about 15 centimetres). Disadvantages are that it is easily injured if dropped and is primarily nocturnal and often bad tempered by day.

Gerbil. Several species of gerbil, mostly from southwestern Asia, became popular pets after having been established as laboratory animals. They are as clean as hamsters and of similar size; however, they are much more active and not as easily handled.

Chinchilla. Originally kept as a fur-producing animal, the chinchilla also makes a charming pet. They enjoy attention, but not all are fond of being picked up and handled.

Paddock pets. This category includes such popular animals as horses and ponies. The high costs of providing for these animals make it increasingly difficult to keep them as personal pets. Space is needed for paddock and stabling, with regular care at all times (see HORSE). In some areas goats are important for their milk and for their wool, and many breeds are kept for show purposes. Some may be suitable as household pets, but they need grazing areas and stabling, with attention twice a day when in milk. The adult male goat is undesirable because of its powerful scent. In farming communities an orphan calf, piglet, or lamb may be adopted by a family, but usually only long enough to rear it safely.

Aviary birds. Many hundreds of species of birds are kept in aviaries, or large enclosures with space for flight. They range from cranes and flamingos to sunbirds and hummingbirds. Jays, magpies, jackdaws, and other members of the crow family are good mimics and make interesting pets. Many are now bred in captivity. The catching of many birds is regulated by governmental agencies.

Wild animals as pets

Unusual mammals. Every year numbers of smaller wild mammals, such as badgers, foxes, otters, hares, and squirrels, are rescued as helpless or wounded young and become hand tamed for varying periods. Only a small minority remain pets all their lives; most of them become difficult to handle after a few months and must be released or sent to the safety of a zoo.

There is much impulse buying of exotic pets; many are offered to zoos after the owners have realized the difficulties of caring for them. The following are some of the most important groups of exotic pets.

Primates. Although apes and monkeys can be kept successfully as house pets, these primates are in general unsuitable for this purpose. With rare exceptions, they cannot be house-trained, they require a great deal of time and attention, and they are potentially dangerous because of their erratic temperament. On the other hand, such primates as bush babies and lorises and other lemurs make interesting pets in warm climates where outside accommodation with some degree of freedom is possible. Their pungent scent, used for marking territory, makes them generally unsuitable as house pets.

Carnivores. Kinkajous, raccoons, coatis, skunks, mongooses, and other small carnivores from many parts of the world are kept in varying numbers, often in large cages, sometimes partly at liberty. Feeding and accommodation are simple, but their temperaments vary widely and only some settle well as pets.

Rodents. Flying squirrels are among the most attractive small rodents. They can do well in pairs in a cage kept indoors. Some enjoy being handled and having occasional freedom of the house so that they may climb and glide.

Jerboas, kangaroo rats, dormice, and other such small rodents make unusual pets and are easy to keep. Because many feed only on seeds and produce dry droppings, they are eminently suitable for keeping indoors.

THE SOURCES OF PETS

Domestic animal sources. Most pets come from domestic stock supplied through breeding and then offered for sale. Small pets going into homes—rodents, budgerigars, etc.—are normally sold through pet shops. The "fancy," or group of enthusiasts interested in a particular pet animal, does most of its business through trade and specialist papers; many dogs and cats are bought and sold in this manner.

Rising costs compel breeders, especially of cats and dogs, to confine themselves to first-class pedigree stock that commands good prices; many buyers go directly to such breeders for their pets. Unfortunately, the standards evolved for some breeds are largely artificial. Attention may be so focussed on details of colour pattern or on small size, for example, that other considerations of temperament and stamina are ignored and the breed may suffer. This has been particularly marked in dog breeding, in which several breeds have declined in quality since World War I.

Native animal sources. Some local animals are deliberately trapped, netted, or taken from the nest. This is especially true of birds, particularly members of the finch and thrush families, that are valued for their song. Most countries now have strict laws that make it illegal to catch birds other than a few regarded as potential pests. The prohibited species may be kept as pets only when "closed ringed"; *i.e.,* when they have been bred in captivity and ringed as nestlings.

Exotic animal sources. A few exotic animals are brought in by returning travellers, but most are obtained through the normal trade channels, which are usually pet stores stocking a wide range of pets and obtaining less common ones on order.

Fishes, amphibians, snakes, invertebrates, etc. are caught by collectors in many parts of the world. Most mammals and birds are found or caught when young, and are much more likely to survive than wild-caught adults; they are also easier and cheaper to ship to markets.

Attempts to limit animal exports

Restrictions are rightly placed on the traffic in animals. The export of certain animals listed in the Red Data Book (list of species in danger of extinction) of the International Union for the Conservation of Nature is totally forbidden; some countries control trading of additional animals and allow export only to approved institutions such as zoos. It will thus become increasingly difficult to import exotic animals, especially mammals and birds, as pets; other classes of animals are less affected. Increasing water pollution, however, could seriously reduce the supply of wild-caught tropical fish.

DISEASES OF PETS AND THEIR HAZARDS TO MAN

Many diseases of pets can be transmitted to humans. Some of these ailments are mild and often pass unnoticed

for what they are; respiratory diseases and skin rashes are among these zoonoses (diseases common to man and other animals).

Cage birds may harbour trachoma, which may damage eye tissue in man, and psittacosis, a viral disease that may be serious in man. Turtles and perhaps other reptilian pets have been implicated as carriers of viruses and fungi that may infect man. Dogs and cats frequently contract respiratory infections from man and in turn can transmit them to man (see also DISEASES OF ANIMALS).

Rabies is one of the most serious animal diseases that can be transmitted to man. It occurs naturally in wild canines such as jackals, foxes, and wolves, and also among the weasels and hyenas. It is widespread in many Asian and European countries where there is always a wild source from which domestic animals may be infected (see INFECTIOUS DISEASES).

External parasites are specific to fairly narrow groups of animals and soon die on other hosts, but some fleas found on cats and dogs may be transferred to man. Suitable insecticides are available to eradicate these pests. Several serious skin diseases are caused by organisms able to live on various pets as well as on man: most important are sarcoptic mange (caused by the mite *Sarcoptes scabei*) and ringworm (caused by the fungus *Microsporum canis*). Among internal parasites, several kinds of parasitic worms of the dog and the cat can be transmitted to man.

Pets obtained from fully reliable sources and well looked after should not endanger human health. Proper care of pets includes sound hygiene, with regular and thorough cleaning of feeding and water bowls, cages or enclosures, and bedding or sleeping area.

Quarantines

It is good practice always to use special bowls for feeding that are kept separate from the utensils used by human members of the household.

PETS IN LAW

Some groups of imported animals are put in quarantine to prevent the spread of serious diseases; in addition, the provision of health certificates from the country of origin may be required. Although details vary with the country and are modified with changing conditions, the most important diseases involving quarantine generally are rabies, foot-and-mouth disease, and psittacosis and other bird diseases.

During outbreaks of some animal diseases, movements within specified areas may be banned and animals such as dogs may be compelled to wear muzzles. Local laws may forbid the keeping of certain types of pets. Animals may be banned from some forms of public transport, and it may be an offense to allow them to run freely on roads or in parks and other public spaces. In many cities, for example, penalties are imposed for the fouling of sidewalks by dogs. In many localities, annual license fees are required for pet animals, especially dogs.

In many countries, societies have been formed that are expressly concerned with the welfare of animals, particularly those in captivity; part of their work may be to help enforce laws that establish minimum standards of care for pets.

BIBLIOGRAPHY. G.S. CANSDALE, *Animals and Man* (1952), a consideration of all aspects of man's dealings with animals from earliest times, and *Bush Baby Book* (1960), a description of six species of bush babies and other lemuroids kept by the author in Africa and Great Britain, including a chapter on the problems of keeping these animals as house pets; H.F. HEDIGER, *Skizzen zu einer Tierpsychologie im Zoo und in Zirkus* (1954; Eng. trans., *Studies of the Psychology and Behavior of Captive Animals in Zoos and Circuses*, 1955, reprinted 1968), and *Mensch und Tier im Zoo* (1965; Eng. trans., *Man and Animal in the Zoo*, 1969), two books by a foremost animal psychologist that deal mainly with captive animals but also include much information about man's personal relationship to them; G. LOISEL, *Histoire des ménageries de l'antiquité à nos jours*, 3 vol. (1912), published for the Universities Federation for Animal Welfare (UFAW), chiefly valuable for its treatment of rare animals kept in captivity and as pets since ancient Egypt; KONRAD LORENZ, *Er redete mit dem Vieh, den Vögeln und den Fischen*, 6–8th ed. (1952; Eng. trans., *King Solomon's Ring: New Light on Animal Ways*, 1952; reduced photographic reprint, 1961), and *So Kam der Mensch auf den Hund* (1950; Eng. trans., *Man Meets Dog*, 1954), two popular books by a world-famous authority on animal behaviour; J.P. VOLRATH, *Animals in Schools* (1955, later editions), a useful manual in a series published by the UFAW that covers the complete range of house pets other than cats and dogs.

(Ge.C.)

Pétain, Philippe

The victor of Verdun, the hardest fought battle of World War I, the French general Philippe Pétain was first a national hero and, then, 25 years later in World War II, by his equivocal role in dealing with the German victors as chief of state of unoccupied France, an extremely controversial—and poignant—figure in French politics.

Early years

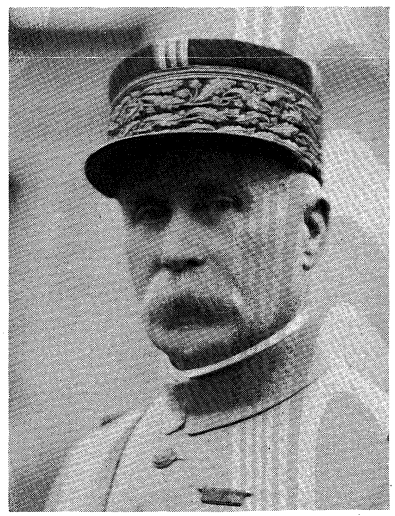
Born into a family of farmers on April 24, 1856, at Cauchy-à-la-Tour in northern France, Pétain, after attending the local village school and a religious secondary school, was admitted to Saint-Cyr, France's principal military academy. As a young second lieutenant in an Alpine regiment, sharing the rough outdoor life of his men, he came to understand the ordinary soldier. The extraordinary popularity he was later to enjoy with the rank and file in World War I is believed to have had its origin there.

His advancement until the outbreak of World War I in 1914—he was 58 when he finally became a general—was slow because as a professor at the War College he had propounded tactical theories opposed to those held by the high command. While the latter favoured the offensive at all costs, Pétain held that a well-organized defensive was sometimes called for and that before any attack the commander must be sure of the superiority of his fire power.

After successively commanding a brigade, a corps, and an army, Pétain in 1916 was charged with stopping the German attack on the fortress city of Verdun. Though the situation was practically hopeless, he masterfully reorganized both the front and the transport systems, made prudent use of the artillery, and was able to inspire in his troops a heroism that became historic. He became a popular hero, and, when serious mutinies erupted in the French Army following the ill-considered offensives of Gen. Robert-Georges Nivelle, then French commander in chief, Pétain was named his successor. He re-established discipline with a minimum of repression by personally explaining his intentions to the soldiers and improving their living conditions. Under him the French armies participated in the victorious offensive of 1918, led by Marshal Ferdinand Foch, generalissimo of the Allied armies. Pétain was made a marshal of France in November 1918 and was subsequently appointed to the highest military offices (vice president of the Supreme War Council and inspector general of the army).

Appointment as commander in chief

EB Inc.



Pétain.

After the German attack of May 1940 in World War II, Paul Reynaud, then head of the government, named Pétain vice premier, and on June 16, at the age of 84, Marshal Pétain was asked to form a new ministry. Seeing the French Army defeated, the "hero of Verdun" asked for an armistice. After it was concluded, the Chamber of Deputies and the Senate, meeting in Vichy, conferred upon him almost absolute powers as "chief of state."

With the German army occupying two-thirds of the country, Pétain believed he could repair the ruin caused by the invasion and the release of the numerous prisoners of war only by cooperating with the Germans. In the southern part of France, left free by the armistice agreement, he set up a paternalistic regime the motto of which was "Work, Family, and Fatherland." Reactionary by temperament and education, he allowed his government to promulgate a law dissolving the Masonic lodges and excluding Jews from certain professions. He was, however, opposed to the policy of close Franco-German collaboration advocated by his prime minister Pierre Laval, whom he dismissed in December 1940, replacing him with Adm. François Darlan. Pétain then attempted to practice a foreign policy of neutrality and delay. He secretly sent an emissary to London, met with the Spanish dictator Francisco Franco, whom he urged to refuse free passage of Hitler's army to North Africa, and maintained a cordial relationship with Adm. William Leahy, the United States ambassador to Vichy until 1942.

Head of Vichy government

When, in April 1942, the Germans forced Pétain to take Laval back as head of the government, he himself withdrew into a purely nominal role. Yet he balked at resigning, convinced that, if he did, Hitler would place all of France directly under German rule. After Allied landings in November 1942 in North Africa, Pétain secretly ordered Admiral Darlan, then in Algeria, to merge the French forces in Africa with those of the Allies. But, at the same time, he published official messages protesting the landing. His double-dealing was to prove his undoing.

In August 1944, after the liberation of Paris by General Charles de Gaulle, Pétain dispatched an emissary to arrange for a peaceful transfer of power. De Gaulle refused to receive the envoy. At the end of August the Germans transferred Pétain from Vichy to Germay. Brought to trial in France for his behaviour after 1940, he was condemned to death in August 1945. His sentence was immediately commuted to solitary confinement for life. He was imprisoned in a fortress on the Île d'Yeu off the Atlantic coast. He died, 95 years old, on July 23, 1951.

BIBLIOGRAPHY. GEORGES BLOND, *Pétain, 1856–1951* (1966), is a recent, definitive biography (in French). On Pétain's long military career prior to 1914, AUGUSTE M.E. LAURE, *Pétain* (1941), while somewhat biassed, furnishes interesting detail. For Pétain's role in World War I, see BERNARD SERRIGNY, *Trente ans avec Pétain* (1959). B.H. LIDDELL HART, *Reputations* (1928); and ALFRED CONQUET, *Autour du maréchal Pétain: Lumières sur l'Histoire* (1963). Interesting familiar details are found in PIERRE BOURGET, *Témoignages inédits sur le Maréchal Pétain* (1960). The first noteworthy attempt at assembling and arranging the documentation on Pétain as chief of state, with a monumental bibliography, was ROBERT ARON, *Histoire de Vichy, 1940–1944* (1954; abridged Eng. trans., *The Vichy Regime, 1940–1944*, 1958). ANDRE BRISSAUD, *La Dernière année de Vichy* (1965), is a well-documented work on this specific period, while his *Pétain à Sigmaringen* (1966), is the most complete book on Pétain's exile and captivity in Germany. Documentation on Pétain's trial may be found in *Le Procès du Maréchal Pétain* (1945); and JACQUES ISORNI, *Souffrances et Mort du Maréchal Pétain* (1951).

(G.Bl.)

Peter the Apostle, Saint

Peter (died *c.* AD 64) was recognized in the early Christian Church as the leader of the disciples of Jesus Christ and by the Roman Catholic Church as the first of its unbroken succession of popes. According to the New Testament, he was designated by Jesus as the "rock" upon which the church would be built, but the meaning of that statement has been the subject of considerable controversy at various times in the history of Christianity up to the present day.

The man and his position among the disciples. The sources of information concerning the life of Peter are limited to the New Testament: the four Gospels, Acts, the letters of Paul, and the two letters that bear the name of Peter. He probably was known originally by the Hebrew name Simeon or the Greek form of that name, Simōn. The former appears only twice in the New Testament; the latter, 49 times. At solemn moments (John 21:15) he was called "Simon, son of John." The Gospel According to John prefers Simon (17 times) or the compound, rarely found elsewhere, of Simon Peter. Though Paul has a distinct preference (eight times out of ten) for the Greek transliteration Kēphas (Latinized as Cephas) of the Aramaic name or title Kepha, meaning "rock," the Gospels and Acts use the Greek translation Petros approximately 150 times. From the Synoptic Gospels (Matt. 8:14) and Paul (I Cor. 9:5), there is indirect evidence that Peter was the son of John and was married, but only later legendary material (*Acts of Nereus and Achilleus; Acts of Philip*) alludes to any children. His family originally came from Bethsaida (John 1:44), but during the period of Jesus' ministry he lived in Capernaum, at the northwest end of the Sea of Galilee, where he and his brother Andrew were in partnership as fishermen with James and John, the sons of Zebedee (Luke 5:10).

Sources of information about Peter

Much can be learned about Peter from the New Testament—either explicitly from the statements made by and about Peter or indirectly from his actions and reactions as revealed in a number of episodes in which he figures prominently. He was at times vacillating and unsure, as in his relations with the church of Antioch when he at first ate with the Gentiles and later refused to do so (Gal. 2:11–14); he could also be resolute (Acts 4:10; 5:1–10). Occasionally he is depicted as rash and hasty (Luke 22:33, etc.) or irritable and capable of great anger (John 18:10). Often he is pictured as gentle but firm and, as in his professions of love to Jesus, capable of great loyalty and love (John 21:15–17).

The New Testament reports that Peter was unlearned in the sense that he was untrained in the Mosaic Law (Acts 4:13), and it is doubtful that he knew Greek. He apparently learned slowly and erred time and time again, but later, when entrusted with responsibility, he demonstrated that he was mature and capable.

The Gospels agree that Peter was called to be a disciple of Jesus at the beginning of his ministry, but when and where the event took place is recorded differently in the several Gospels. Luke (5:1–11) scarcely mentions James and John and omits Andrew while emphasizing the call of Peter. Matthew (4:18–22) and Mark (1:16–20) note the call of the four men and—with Luke—agree that the event took place at the Sea of Galilee. The Gospel According to John places the call in Judaea (1:28) and states that Andrew—who had been a follower of John the Baptist (1:35) and had heard John indicate that Jesus was the Lamb of God—left John and introduced Peter to "the Messiah," who at that time gave him the name (or title) Cephas (*i.e.*, Peter, or "Rock").

The Synoptic Gospels (Matthew, Mark, and Luke) are probably correct in recording that the call to Peter was extended in Galilee when Jesus first began his work in that area. The Gospel According to John is here, as elsewhere, perhaps more theologically than historically motivated; the author of John wishes to stress that Peter recognized Jesus' Messiahship from the beginning and that Jesus had seen Simon as the rock from their first meeting. Several not generally convincing attempts have been made to reconcile the differences between the Synoptics and the Fourth Gospel.

The Synoptic Gospels largely agree in the amount of emphasis each gives to the leadership of Peter among the Twelve Apostles, but there are differences also. For example, in one case Matthew and Luke note that Peter was the speaker in questioning Jesus about a parable, but Mark has attributed these words to the group of disciples (see Matt. 15:15; Luke 8:45; and Mark 7:17). With differing degrees of emphasis, the Synoptic Gospels agree that Peter served as spokesman, the outstanding member of the group, and enjoyed a certain precedence over the

Early prominence



St. Peter, detail of a tempera painting on wood by Nardo di Cione, second half of the 14th century. In the Yale University Art Gallery.

By courtesy of the Yale University Art Gallery, James Jackson Jarves Collection

other disciples. Whenever the disciples are listed, Peter is invariably mentioned first (Matt. 10:2–4; Mark 3:16–19; Luke 6:14–16; Acts 1:13; *cf.* only Gal. 2:9). Although it is not certain whether or not this priority is due primarily to reading back into the Gospel narrative Peter's importance in the apostolic church, his forceful personality was surely a factor.

Those not belonging to the immediate followers of Jesus also recognized the authority of Peter, such as when the collectors of the temple tax approached him for information (Matt. 17:24). Again, with characteristic quickness he sought a clarification from Jesus on behalf of the disciples concerning the meaning of a parable (Matt. 15:15) or of a saying (Matt. 18:21). As both an individual and representative of the Twelve Apostles, he made a plea for personal preference in the Kingdom of Heaven as a reward for faithful service (Matt. 19:27, 28).

On several occasions Peter alone is mentioned by name, and others are indicated as merely accompanying him (Mark 1:36; Luke 8:45). Even when the three disciples closest to Jesus (the "pillars"—Peter, James, and John) figure in a particular incident, it is frequently Peter alone who is named. When the three are named, Peter's name invariably appears first (as in Matt. 17:1, 26:37). It was his home in Capernaum that Jesus visited and cured Peter's mother-in-law (Matt. 8:14); it was Peter's boat that Jesus used when he instructed the crowd (Luke 5:3). It was Peter who possessed remarkable insight and displayed his depth of faith in the confession of Christ as the Son of God (Matt. 16:15–18; Mark 8:29; Luke 9:20); and it was Peter who rebuked and in turn was rebuked by Jesus when the Master prophesied that he would suffer and die (Mark 8:32, 33). It was also Peter who manifested the momentary weakness of even the strongest in the denial of his Lord (Matt. 26:69–75; Mark 14:66–72; Luke 22:54–61). Later, however, with greater maturity, he discovered strength and, as he was charged by Jesus (Luke 22:31, 32), effected the strengthening of others. Finally, Peter, who survived his denial, was permitted to be the first witness of the Resurrection (Luke 24:34).

In the Fourth Gospel, the prominence of Peter is challenged in the person of John, the "Beloved Disciple." Though Peter receives mention in John 37 times (out of a total of 109 times in the four Gospels), one-third of the references are found in the appendix (chapter 21), and he appears in only 9 incidents. The Gospel According to John attempts to show the close relationship between John and Jesus while still reserving to Peter the role of representative and spokesman. The fact that Peter is emphasized in John and charged by Jesus to "tend my sheep" and "feed my lambs" (John 21:15, 16) at the same time that the role of the disciples as a whole is being de-emphasized attests to the prestige of Peter in the apostolic church. But throughout the Fourth Gospel Peter is required to share his prominence with John (13:24; 18:15; 19:26, 27, etc.). Among the purposes of chapter 21 in emphasizing Peter may well be an attempt to restore the disciple who denied his Lord to the position he enjoyed in the Synoptic Gospels.

Incidents important in interpretations of Peter. Out of the many incidents in which Peter figures prominently in the Gospels, three should be separately and carefully considered; for each is important, contains problems of interpretation, and is controversial.

Peter's confession of faith

In Mark (8:29) and Luke (9:20), to a question of Jesus concerning his essential identity, about which he pressed the disciples for an opinion, Peter answered for them all that Jesus is the "Messiah" or "God's Messiah." In adjuring them to be silent, Jesus rejected the response as perhaps too partial, too political. In the Matthean version (16:13), expanding upon the narrative in Mark, Peter answered for himself and presumably for the other disciples, "You are the Christ, the Son of the living God." A new dimension of understanding was thus reached, and this heightened awareness of Jesus' divinity was approved by Jesus and occasioned Peter's "ordination."

In what may be a grouping of Petrine material (Matt. 16:18, 19)—the confession, naming, and receiving of authority—Jesus gave to Simon the title of petros ("rock").

> And I tell you, you are Peter, and on this rock I will build my church, and the powers of death shall not prevail against it. I will give you the keys of the kingdom of heaven, and whatever you bind on earth shall be bound in heaven, and whatever you loose on earth shall be loosed in heaven.

Though in the past some authorities have considered that the term rock refers to Jesus himself or to Peter's faith, the consensus of the great majority of scholars today is that the most obvious and traditional understanding should be construed, namely, that rock refers to the person of Peter. In John the title was granted at what may have been their first meeting (1:42). Thus *when* the name was given is open to question, but *that* the name was given by Jesus to Peter seems fairly certain. Matthew continues that upon this rock—that is, upon Peter—the church will be built. The word church in the 1st-century Gospel According to Matthew is to be understood as referring to the community of the faithful rather than to a definite ecclesiastical organization. The passage continues that to this disciple, upon whom the new community is to be founded, will be given the "keys of the kingdom of heaven"; *i.e.*, Peter is to be granted a position of unique authority within the community. But he is to share with his fellow disciples (Matt. 18:18) the authority of "binding" and "loosing" —*i.e.*, controlling admission to the community.

The authenticity of the uniquely Matthean material (Matt. 16:16–19) of this narrative has been and is widely discussed and has been challenged on the bases (1) that verses 16–19 are found only in Matthew or (2) that the inclusion of the word church suggests a level of organization acquired only at a later period. Though these and other arguments against authenticity are given most careful consideration, the general consensus is that at some time—and more likely at the end of his career—these words were spoken by Jesus.

The denial

If Peter's confession demonstrates his faith and insight, his denial that he knew Jesus demonstrates a weakness of will (even if momentary), capability of inaction, and a tendency toward vacillation, but not a loss of faith. Prior to the denial, out of his deep love for Jesus and his overestimation of his own capabilities, he had sought to overrule Jesus' prophecy of his denial and declared that, even if the other disciples deserted Jesus, he would suffer death rather than disown his Lord (Matt. 26:33–35; Mark 14:29–31; Luke 22:31–34; John 13:37–38). As the drama unfolded, Peter fled when Jesus was arrested but did find his way to the palace of the high priest where Jesus had been taken. When confronted in the courtyard

with the danger of admitting having had association with Jesus, he chose to deny (Matt. 26:69–75; Mark 14:66–72; Luke 22:54–61; John 18:15–18, 25–27). The degree of his shame and the depth of his love were revealed when he later realized that the prophecy had been fulfilled, and he wept bitterly (Matt. 26:75; Mark 14:72).

The fact of Peter's denial did not destroy the love and trust that Jesus felt for him. It was to Peter—who had confessed the Sonship of Jesus (Matt. 16:16), who had been commissioned earlier to "lend strength" to his brothers (Luke 22:32), who had hesitated in his resolution at one crucial point (Mark 14:66–72), and who on the morning of the resurrection"ran to the tomb" (Luke 24:12)—that the Resurrected Christ first appeared. The earliest report of Peter's priority as a witness to the Resurrection is found in the letters of Paul (I Cor. 15:5), and this most probably is the intent of Luke (24:34). The initial appearance to Peter in Galilee may have been included in the original ending of Mark.

The silence concerning this important matter of priority in Matthew and John is remarkable. It may be, however, that Matt. 14:27, 28 represents a misplaced post-Resurrection narrative, and John 21 may contain an echo of the tradition preserved by Paul (I Cor. 15:5).

Whether or not Jesus appeared *first* to Peter after the Resurrection, he was a witness, which Peter declared to be a criterion of apostleship (Acts 1:22).

The position of Peter in the Apostolic Church. Given the information supplied by the Gospels, it is not unexpected that Peter should emerge immediately after Jesus' death as the leader of the earliest church. For approximately 15 years after the Resurrection, the figure of Peter dominated the community. He presided over the appointment of Matthias as an Apostle (Acts 1:23–26) to take the place of Judas, who had betrayed Christ and later died. It was Peter who first "raised his voice" and preached at Pentecost, the day when the church came into being (Acts 1:14–39). It was Peter who served as an advocate for the Apostles before the Jewish religious court in Jerusalem (Acts 4:5–22). And it was he who exercised the role of judge in the disciplining of those who erred within the church (Acts 5:1–10).

Peter as leader of the Apostles

Peter led the Twelve Apostles in extending the church "here and there among them all" (Acts 9:32). He went first to the Samaritans (Acts 8:4–17), "who received the Holy Spirit"; in Samaria he encountered the magician and faith healer Simon Magus (the theme of their meeting is developed in apocryphal literature, not included in the Bible); then he went to Lydda, in the plain of Sharon (Acts 9:32–35), where he healed the paralyzed Aeneas; and then at the Mediterranean coastal town of Joppa (Acts 9:36–43) he effected the cure of Tabitha (Dorcas) in the name of Christ.

He went farther north on the Mediterranean coast to Caesarea (Acts 10:1–11:18), where, through the conversion of Cornelius, "a centurion of what was known as the Italian Cohort" (Acts 10:1), Peter introduced Gentiles into the church. According to Jewish requirements, a Gentile convert must first become a Jew through the rite of circumcision and be acceptable as a proselyte. In accepting Cornelius and the others—who may have had some informal connection with the synagogue (Acts 10:1)—and ordering "them to be baptized in the name of Jesus Christ" (Acts 10:48) without submission to the prior rite of circumcision, Peter introduced an innovation that insured the opposition of the Jewish Christians and others. This independent course set by Peter with the blessing of "the Spirit" (Acts 10:10–15) was possibly a factor in Herod's beheading of James (the brother of John) and in the arrest of Peter (Acts 12:2, 3). In prison (*c.* AD 44) Peter was visited by an "angel of the Lord. . . . And the chains fell off his hands," and he made his escape (Acts 12:1–8). He went immediately to "the house of Mary, the mother of John whose other name was Mark" (Acts 12:12). After asking them to report his escape "to James and to the brethren," he "went to another place" (Acts 12:17).

At this point the unchallenged leadership of Peter in Jerusalem came to an end. It is not at all clear where Peter went, but it is not probable that the words to another place refer to a different home in the same general area that would provide temporary refuge. Though there is no internal evidence to suggest any place in particular, he might have travelled to any or none of those places with which tradition has associated his name—perhaps to Antioch of Syria, where he visited later, perhaps to Corinth, perhaps to Rome. Those who would preserve the ancient tradition of the 25-year episcopate of Peter in Rome are particularly confident that the "place" indicated is Rome. Many others, who are certain that he eventually went there, are equally positive that his visit to Rome could not have been this early.

The later work of Peter is not covered in Acts, perhaps because the author of Luke–Acts had planned a third book that would have included such a discussion, but the book was never written or was written and later lost. Perhaps the events would have included unedifying material such as the internal jealousy within the church alluded to in the *First Letter of Clement* 4–6, or perhaps the author died before completion of his work. Whatever momentary glimpses into the period of the later ministry of Peter remain can only be noted in a discussion of his relationship with the two other outstanding Apostles of the time, James and Paul.

Peter and James

Peter was the most prominent figure in the Jerusalem Church up to the time of his departure from Jerusalem after his imprisonment by King Herod and his subsequent release in the New Testament account (Acts 12:1–17). For example, Paul went up to Jerusalem to consult with Peter three years after he had been converted and remained with him for two weeks (Gal. 1:18, 19). When Peter left Jerusalem, however, it appears clear to many New Testament scholars (although unconvincing to others) that he assumed a missionary role while the actual leadership of the church devolved upon James, "the brother of the Lord." This sequence of authority is suggested by Peter's obedience to the wishes of "certain persons who came from James" and hence his ceasing to eat with Gentile Christians at Antioch (Gal. 2:11–14); by a final "summing up" of decisions made in the so-called apostolic Council of Jerusalem (Acts 15:7) by James; and later, when Peter made his departure from the home of the mother of John whose other name was Mark, by the word of explanation or "report" of his whereabouts left primarily for James (Acts 12:17).

Peter and Paul

Paul first met with Peter at Jerusalem three years after his conversion. In the record of this meeting the name of Cephas (Peter) precedes that of James, although Galatians notes that in another meeting 14 years later the name of James precedes that of Cephas (Gal. 2:9). Paul also emphasized an incident involving himself and Peter at Antioch. Apparently, Paul had achieved some success in the difficult matter of welding the Jewish and Gentile Christians of Antioch into one congregation. The Jewish Christians saw the sharing of food with Gentiles as quite alien to their tradition. In the absence of Paul, Peter, perhaps in his capacity as missionary, visited Antioch and ate with the united group. Later, "certain persons came from James" and opposed the united congregation's custom of eating together. In apparent deference to James, Peter "drew back and began to hold aloof," and the Jewish Christians did likewise. The unity of the group had been destroyed. When Paul returned, he upbraided Peter for what he may have considered Peter's vacillation or perhaps even purposeful disruption (Gal. 2:11–14). This incident may have occasioned the Jerusalem Council (AD 49 or 50), in which it was settled that hereafter Paul should be "entrusted with the gospel to the uncircumcised" (Gal. 2:7) and Peter "for the mission to the circumcised" (Gal. 2:8). A late tradition found in Origen's homily on Saint Luke and in the *Clementine Homilies* and *Recognitions* claims that Peter was the first bishop of Antioch, a tradition that may rest upon the early papal lists of the last half of the 2nd century.

In passing, Paul refers to a party of "Cephas" (Peter) in I Corinthians 1:12 that suggests that a group in the church of Corinth was especially devoted to Peter (leading some to assume a residence of Peter in Corinth) and

in I Corinthians 9:5 to Peter as carrying on missionary activity accompanied by his wife. A missionary journey to Asia Minor may be suggested in the First Letter of Peter 1:1.

Except for the possible references to the later residence and martyrdom of the Apostle noted below, there is no further information to be gathered about his life from the pages of the New Testament.

Tradition of Peter in Rome. The problems surrounding the residence, martyrdom, and burial of Peter are among the most complicated and controversial of all those encountered in the study of the New Testament and the early church. The absence of any reference in Acts or Romans to a residence of Peter in Rome gives pause but is not conclusive. If Peter did write I Peter, the mention of "Babylon" in 5:13 is fairly reliable evidence that Peter resided at some time in the capital city. If Peter was not the author of the first epistle that bears his name, the presence of this cryptic reference witnesses at least to a tradition of the late 1st or early 2nd century. An identification of the "Babylon" in I Peter 5:13 with the Babylons of the Euphrates in Mesopotamia or the Nile Delta near Memphis in Egypt must be discarded for lack of evidence, as must the suggestion that it stands for Jerusalem. In that period the first Babylon was "no more than walls"; the second was a military outpost; and for the third there is lacking any convincing evidence that the term was used in this way. "Babylon" is a cryptic term indicating Rome, and it is the understanding utilized in Revelation 14:8; 16:19; 17:5, 6 and in the works of various Jewish seers, such as Baruch 1:1, 4; Esdras 2; and the Sibylline Oracles 5:143, 158.

It may be said that by the end of the 1st century there existed a tradition that Peter had lived in Rome. Further early evidence for the tradition is found in the *Letter to the Romans* by the early-2nd-century bishop of Antioch Ignatius. Later, less trustworthy, evidence exists in the fragment of a work by the 2nd-century bishop of Hierapolis Papias (Eusebius, *Ecclesiastical History*, III. xxxix. 15) and in the *Letter to the Romans* of Dionysius of Corinth (Eusebius, *Ecclesiastical History*, II.xxv.8). Also to be noted are other works: Irenaeus, *Against Heresies*, III. iii. 3; Tertullian, *Against Marcion*, IV. v, *Against the Jews*; the words of the Roman presbyter Gaius (Eusebius, *Ecclesiastical History*, II.xxv. 7); and Clement of Alexandria, *Commentary on I Peter*. It is probable that the tradition of a 25-year episcopate of Peter in Rome is not earlier than the beginning or the middle of the 3rd century. The claims that the church of Rome was founded by Peter or that he served as its first bishop (in the present meaning of the word) are in dispute and rest on evidence that is not earlier than the middle or late 2nd century.

Martyrdom of Peter

Words of John 21:18,19 clearly allude to the death of Peter and are cast into the literary form of prophecy. The author of this chapter is aware of a tradition concerning the martyrdom of Peter when the Apostle was an old man. And there is a possible reference here to crucifixion as the manner of his death. But as to when or where the death took place there is not so much as a hint. It has been forcefully suggested, though rarely supported, that Revelation 11:3–13 contains a cryptic reference to the deaths of both Peter and Paul.

The strongest evidence to support the thesis that Peter was martyred in Rome is to be found in the *Letter to the Corinthians* (*c.* AD 96; 5:1–6:4) of Clement of Rome:

> Peter, who by reason of wicked jealousy, not only once or twice but frequently endured suffering and thus, bearing his witness, went to the glorious place which he merited (5:4) . . . . To these men [Peter and Paul] who lived such holy lives there was joined a great multitude of the elect who by reason of rivalry were victims of many outrages and tortures and who became outstanding examples among us (6:1).

These sources, plus the suggestions and implications of later works, combine to lead many scholars to accept Rome as the location of the martyrdom and the reign of Nero, between AD 64 and 67, as the time. The tradition that he suffered crucifixion in a head-downward position is late and apocryphal. *Why* he was martyred is a question that must still remain unanswered. Perhaps the answer is somehow related to Nero and the great fire of Rome (as implied in I Clement 6:1) or to a refusal to make ritual offering on behalf of the emperor or to one or a number of charges that an unfriendly community might bring against a group, such as the Christians, that it disliked and distrusted. Only later and less trustworthy sources appear to be concerned with the area in Rome where the martyrdom took place, but among these there is virtual agreement in citing the vicinity of the Vatican.

Burial of Peter

As part of the general question of Peter's residence and martyrdom in Rome, debated since the appearance of the *Defensor pacis* of Marsilius of Padua (*c.* 1275–*c.* 1342), the particular question of where Peter was buried has been argued. There is not the slightest hint at a solution in the New Testament. The earliest evidence (*c.* AD 200) is found in a fragment of a work by Gaius (or Caius) witnessing to a tradition at least a generation earlier (*c.* AD 165) that the "trophy" (*i.e.*, *tropaion*, or monument) of Peter was located at the Vatican. Though difficult to interpret, the use of the word trophy indicates that in this period the Vatican area was associated with either the tomb of the Apostle or simply a monument erected in the area of Peter's victory (*i.e.*, his martyrdom).

Some scholars find support for a tradition that the Apostle was buried "Ad Catacumbas" ("at the catacombs" of San Sebastiano) on the Via Appia in an inscription of Damasus (pope, 366–384), composed in such ambiguous terms that it was certain to foster such misinterpretations as are found in the letter of Gregory the Great to Empress Constantina and the notice of Cornelius in the *Liber Pontificalis*. Apart from the aforementioned, later literary tradition is unanimous in indicating the Vatican Hill as the place of burial. See *Peristephanon*, XII, of Prudentius, various notices in the *Liber Pontificalis*, and *The Salzburg Itinerary*. Liturgical sources such as the *Depositio Martyrum*, *Martyrologium Hieronymianum*, though interesting, add nothing to the literary evidence.

Excavations were begun in the late 19th century in order to substantiate the theory that the burial of Peter and Paul was "Ad Catacumbas." After a half century of investigation, it now seems reasonable to concede that a cult of the Apostles existed there about AD 260, though Christian influence may have been exerted as early as AD 200. None of the excavations, however, in all of the areas indicated at various times as the resting place of the apostolic relics, have produced any evidence whatsoever that the bodies of Peter and Paul were either buried there originally or brought there at a later time after earlier burials elsewhere.

In the early 4th century, the emperor Constantine (died AD 337) with considerable difficulty erected a basilica on the Vatican Hill. The difficulty of the task, combined with the comparative ease with which this great church might have been built on level ground only a slight distance to the south, may support the contention that the Emperor was convinced that the relics of Peter rested beneath the small Aedicula (shrine for a small statue) over which he had erected the basilica. The task before the excavators was to determine whether or not the belief of Constantine accorded with the facts or was based merely upon a misunderstanding.

The excavation of this site, which lies far beneath the high altar of the present Church of St. Peter, was begun in 1939. The problems encountered in excavation and interpretation of what has been discovered are extremely complex. There are some scholars who are convinced that a box found in one of the fairly late sidewalls of the Aedicula contains fragments of the remains of the Apostle, fragments which at an earlier time may have rested in the earth beneath the Aedicula. Others are most definitely not convinced. If a grave of the Apostle did exist in the area of the base of the Aedicula, nothing identifiable of that grave remains today. Furthermore, the remains discovered in the box that until comparatively recently rested in the sidewall do not lead necessarily to a more positive conclusion. Archaeological investigation has not solved with any great degree of certainty the question of the location of the tomb of Peter. If it was not in the area of the Aedicula, perhaps the grave rested elsewhere in the

immediate vicinity, or perhaps the body was never recovered for burial at all.

Peter in the apocryphal literature

Later traditions about Peter. Two works attributed to Peter (I and II Peter) are included in the New Testament. In addition to these, there are several apocryphal works that record supposed incidents in the life of the Apostle and contain further teaching. This apocryphal literature is untrustworthy and too late to offer any information that might be considered reliable and independent. The details offered are only important to the extent that they serve to indicate a trend in the development of the tradition. The *Gospel of Peter* supplies details of a post-Resurrection experience of certain disciples, including Peter. The *Gospel of Bartholomew* reports discussions between Mary, the Mother of Christ, and the Apostle. Mary refers to Peter as the "chief" and the "pillar" and "stone hewn out of the rock." In *The Death and Assumption of the Virgin*, Peter is ordained an "archbishop" by Jesus. An appearance of Mary to Peter and John is promised before her assumption into heaven in the *Discourse of Theodosius*.

The *Acts of Peter* includes curious incidents and details not found in the canonical Gospels: Peter's healing of his daughter and then returning her to her infirmity (a similar narrative in the Letter of Paul to Titus may explain the story of the healing of Peter's daughter); a long narrative (chapters 7–32) relating to the conflict between Peter and Simon Magus in Rome that includes discussions and miraculous cures; the "Quo Vadis" ("where are you going?") legend (chapter 35) involving a dialogue between Christ and Peter in Rome; a prophecy of Peter's crucifixion in that city; and the crucifixion, "head downward" (chapter 37). This last detail is noted also in the *Acts of Andrew and Paul*. Further details of teaching and miracles are found in the brief *Acts of Peter and Andrew*. Additional post-Resurrection experiences with Christ are detailed in the *Apocalypse of Peter*.

The alleged stay of Peter in the Mamertine prison in Rome is presented in the Pseudo-Linus text that is little more than a detailed account of the *Martyrdom of Peter*. Material appearing in the *Martyrdom* is elaborated upon in the 6th-century *Acts of Processus and Martinian*. Other legends are contained in a *Latin Passion of Peter and Paul, Acts of Nereus and Achilleus, Preaching of Peter, Slavonic Acts of Peter*, and *Didascalia Apostolorum*, the *Clementine Homilies* and *Recognitions*, etc.

The feasts of St. Peter

Five festivals in the calendar of the Roman Catholic Church involve honour paid to Peter. And in each, the name of Paul is also associated. First chronologically, on January 18, is celebrated the festival of the "Cathedra Petri" at Rome, and on February 22 at Antioch. June 29 marks the festival of Peter and Paul, ranking among the 12 most important celebrations of the Roman Catholic Church. The escape of Peter from his chains is noted in the feast of August 1. Last, the dedications of the basilicas of Peter and Paul, commemorating their construction by the emperor Constantine, are celebrated in the festival of November 18.

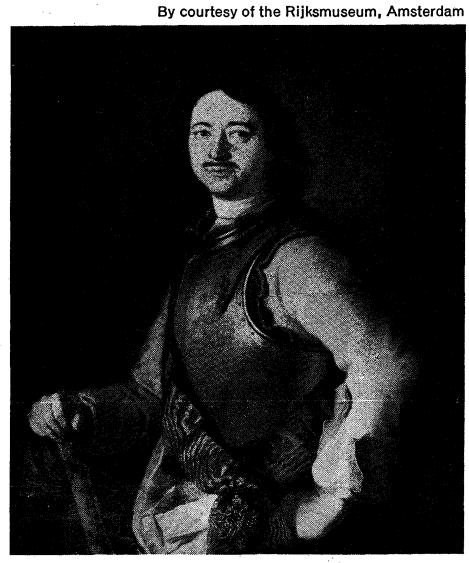
BIBLIOGRAPHY. O. CULLMANN, *Petrus* (1960; Eng. trans., *Peter: Disciple, Apostle, Martyr*, 2nd rev. ed., 1962), a critical discussion of all problems involved in Petrine studies by a Protestant; C. JOURNET, *Primauté de Pierre dans la perspective protestante et dans la perspective catholique* (1953; Eng. trans., *The Primacy of Peter*, 1954), and O. KARRER, *Peter and the Church* (1963), two Roman Catholic appraisals of Cullmann's study; A. DEMARCO, *The Tomb of Saint Peter* (1964), an annotated bibliography of the Vatican and San Sebastiano excavations and related questions, compiled by a Roman Catholic scholar; F.J. FOAKES-JACKSON, *Peter: Prince of Apostles* (1927), a Protestant study of Peter in the New Testament, apocryphal literature, and later tradition; K. HEUSSI, *Die römische Petrustradition in kritischer Sicht* (1958), a Protestant study that seeks to cast doubt on the residence, martyrdom, and burial of Peter in Rome; D.W. O'CONNOR, *Peter in Rome: The Literary, Liturgical and Archeological Evidence* (1969), a recent Protestant discussion of all the evidence that deals with the residence, martyrdom, and burial of Peter in Rome; N. AFANASSIEFF *et al.*, *Der Primat des Petrus in der Orthodoxen Kirche* (1961; Eng. trans., *The Primacy of Peter*, 1963), an Orthodox review of several problems in Petrine studies, especially the primacy of Peter; J. TOYNBEE and J. WARD PERKINS, *The Shrine of St. Peter and the Vatican Excavations* (1956), an excellent Roman Catholic account of the Vatican excavations, dealing at length with the pagan mausolea and the area of the Christian Aedicula; M.M. WINTER, *Saint Peter and the Popes* (1960), a general Roman Catholic review of the various Petrine problems; H. LECLERCQ in F. CABROL, *Dictionnaire d'archéologie chrétienne et de liturgie*, vol. 14, col. 822–981 (1938), an excellent Roman Catholic interpretation in the wide range of Petrine problems—written, however, prior to the Vatican excavations; J. LOWE, *Saint Peter* (1956), a brief Protestant summary of the life, martyrdom, and burial of Peter.

(D.W.O'C.)

Peter I the Great, of Russia

Peter I the Great, tsar of Russia from 1682 and emperor from 1721, was one of his country's greatest statesmen, organizers, and reformers. He was born in Moscow on June 9 (May 30, old style), 1672, the son of Tsar Alexis by his second wife, Natalya Kirillovna Naryshkina. Unlike his half-brothers, sons of his father's first wife, Mariya Ilinichna Miloslavskaya, Peter proved a healthy child, lively and inquisitive. It is probably significant to his development that his mother's former guardian, Artamon Sergeyevich Matveyev, had raised her in an atmosphere open to progressive influences from the West.

By courtesy of the Rijksmuseum, Amsterdam



Peter I the Great, portrait by Aert de Gelder (1645–1727). In the Rijksmuseum, Amsterdam.

Youth and accession. When Alexis died in 1676 Peter was only four years old. His elder half-brother, a sickly youth, then succeeded to the throne as Fyodor III; but, in fact, power fell into the hands of the Miloslavskys, relatives of Fyodor's mother, who deliberately pushed Peter and the Naryshkin circle aside. When Fyodor died childless in 1682, a fierce struggle for power ensued between the Miloslavskys and the Naryshkins: the former wanted to put Fyodor's brother, the delicate and feebleminded Ivan V, on the throne; the Naryshkins stood for the healthy and intelligent Peter. Representatives of the various orders of society, assembled in the Kremlin, declared themselves for Peter, who was then proclaimed tsar; but the Miloslavsky faction exploited a revolt of the Moscow *streltsy*, or musketeers of the sovereign's bodyguard, who killed some of Peter's adherents, including Matveyev. Ivan and Peter were then proclaimed joint tsars; and eventually, because of Ivan's precarious health and Peter's youth, Ivan's 25-year-old sister Sophia was made regent. Clever and influential, Sophia took control of the government; excluded from public affairs, Peter lived with his mother in the village of Preobrazhenskoye, near Moscow, often fearing for his safety. All this left an ineradicable impression on the young tsar and determined his negative attitude toward the *streltsy*.

One result of Sophia's overt exclusion of Peter from the government was that he did not receive the usual educa-

tion of a Russian tsar; he grew up in a free atmosphere instead of being confined within the narrow bounds of a palace. While his first tutor, the former church clerk Nikita Zotov, could give little to satisfy Peter's curiosity, the boy enjoyed noisy outdoor games and took especial interest in military matters, his favourite toys being arms of one sort or another. He also occupied himself with carpentry, joinery, blacksmith's work, and printing.

Boyhood influences

Near Preobrazhenskoye there was a *vemetskaya sloboda* ("colony") where foreigners were allowed to reside. Acquaintance with its inhabitants aroused Peter's interest in the life of other nations, and an English sailboat, found derelict in a shed, whetted his passion for seafaring. Mathematics, fortification, and navigation were the sciences that appealed most strongly to Peter. A model fortress was built for his amusement, and he organized his first "play" troops, from which, in 1687, the Preobrazhensky and Semyonovsky Guards regiments were formed—to become the nucleus of a new Russian Army.

Early in 1689 Natalya Naryshkina arranged Peter's marriage to the beautiful Eudoxia (Yevdokiya Fyodorovna Lopukhina). This was obviously a political act, intended to demonstrate the fact that the 17-year-old Peter was now a grown man, with a right to rule in his own name. The marriage did not last long: Peter soon began to ignore his wife, and in 1698 he relegated her to a convent.

In August 1689 a new revolt of the *streltsy* took place. Sophia and her faction tried to use it to their own advantage for another coup d'état, but events this time turned decisively in Peter's favour. He removed Sophia from power and banished her to the Novodevichy convent, where later in the year, after the suppression of the revolt, she became a nun. Though Ivan V remained nominally joint tsar with Peter, the administration was now largely given over to Peter's kinsmen, the Naryshkins, until Ivan's death in 1696. Peter, meanwhile continuing his military and nautical amusements, sailed the first seaworthy ships to be built in Russia. His games proved to be good training for the tasks ahead.

Peter as sole ruler

External events. At the beginning of Peter's reign, Russia was territorially a huge power, but with no access to the Black Sea, the Caspian, or to the Baltic, and to win such an outlet became the main goal of Peter's foreign policy.

The Azov campaigns (1695–96). The first steps taken in this direction were the campaigns of 1695 and 1696, with the object of capturing Azov from the Crimean Tatar vassals of Turkey. On the one hand, these Azov campaigns could be seen as fulfilling Russia's commitments, undertaken during Sophia's regency, to the anti-Turkish "Holy League" of 1684 (Austria, Poland, and Venice); on the other they were intended to secure the southern frontier against Tatar raids, as well as to approach the Black Sea. The first campaign ended in failure (1695), but this did not discourage Peter: he promptly built a fleet at Voronezh to sail down the Don River and in 1696 Azov was captured. To consolidate this success Taganrog was founded on the northern shore of the Don Estuary, and the building of a large navy was started.

The Grand Embassy (1697–98). Having already sent some young nobles abroad to study nautical matters, Peter, in 1697, went with the so-called Grand Embassy to western Europe. The embassy comprised about 250 people, with the "grand ambassadors" Franz Lefort, F.A. Golovin, and P.B. Voznitsyn at its head. Its chief purposes were to examine the international situation and to strengthen the anti-Turkish coalition, but it was also intended to gather information on the economic and cultural life of Europe. Travelling incognito under the name of Sgt. Pyotr Mikhaylov, Peter familiarized himself with conditions in the advanced countries of the West. For four months he studied shipbuilding, working as a ship's carpenter in the yard of the Dutch East India Company at Saardam; after that he went to Great Britain, where he continued his study of shipbuilding, working in the Royal Navy's dockyard at Deptford, and he also visited factories, arsenals, schools, and museums and even attended a session of Parliament. Meanwhile, the services of foreign experts were engaged for work in Russia.

Familiarization with the West

On the diplomatic side of the Grand Embassy, Peter conducted negotiations with the Dutch and British governments for alliances against Turkey; but the Maritime Powers did not wish to involve themselves with him because they were preoccupied with the problems that were soon to come to a crisis, for them, in the War of the Spanish Succession.

The destruction of the streltsy (1698). From England, Peter went on to Austria; but while he was negotiating in Vienna for a continuance of the anti-Turkish alliance, he received news of a fresh revolt of the *streltsy* in Moscow. In the summer of 1698 he was back in Moscow, where he suppressed the revolt. Hundreds of the *streltsy* were executed, the rest of the rebels were exiled to distant towns, and the corps of the *streltsy* was disbanded.

The Northern War (1700–21). When it became clear that Austria, no less than the Maritime Powers, was preparing to fight for the Spanish Succession and to make peace with Turkey, Peter saw that Russia could not contemplate a war without allies against the Turks, and he abandoned his plans for pushing forward from the Sea of Azov to the Black Sea. By the Russo-Turkish Peace of Constantinople (Istanbul, 1700) he retained possession of Azov. He was now turning his attention to the Baltic instead, following the tradition of his predecessors.

The Swedes occupied Karelia, Ingria, Esthonia, and Livonia and blocked Russia's way to the Baltic coast. To dislodge them, Peter took an active part in forming the great alliance, comprising Russia, Saxony, and Denmark–Norway, which started the Northern War in 1700. This war lasted for 21 years and was Peter's main military enterprise. In planning it and in sustaining it he displayed iron willpower, extraordinary energy, and outstanding gifts of statesmanship, generalship, and diplomacy. Mobilizing all the resources of Russia for the triumph of his cause, constantly keeping himself abreast of events, and actively concerning himself with all important undertakings, often at his personal risk, he could be seen sometimes in a sailor's jacket on a warship, sometimes in an officer's uniform on the battlefield, and sometimes in a labourer's apron and gloves with an axe in a shipyard.

Peter's main military enterprise

The defeat of the Russians at Narva (1700), very early in the war, did not deter Peter and, in fact, he later described it as a blessing: "Necessity drove away sloth and forced me to work night and day." He subsequently took part in the siege that led to the Russian capture of Narva (1704) and in the battles of Lesnaya (1708) and of Poltava (1709). At Poltava, where Charles XII of Sweden suffered a catastrophic defeat, the plan of operations was Peter's own: it was his idea to transform the battlefield by works of his military engineers—the redoubts erected in the path of the Swedish troops to break their combat order, to split them into little groups, and to halt their onslaught. Peter also took part in the naval battle of Gangut (Hanko, or Hangö) in 1714, the first major Russian victory at sea.

The treaties concluded by Russia in the course of the war were made under Peter's personal direction. He also travelled abroad again for diplomatic reasons—*e.g.*, to Pomerania in 1712 and to Denmark, northern Germany, Holland, and France in 1716–17.

In 1703, on the banks of the Neva River, where it flows into the Gulf of Finland, Peter began construction of the city of St. Petersburg (now Leningrad) and established it as the new capital of Russia in 1712. By the Treaty of Nystad (September 10 [August 30, O.S.], 1721) the eastern shores of the Baltic were at last ceded to Russia, Sweden was reduced to a secondary power, and the way was opened for Russian domination over Poland.

In celebration of his triumph, the Senate on November 2 (October 22, O.S.), 1721, changed Peter's title from tsar to that of emperor (imperator) of all the Russias.

The popular revolts (1705–08). The peasant serfs and the poorer urban workers had to bear the greatest hardships in wartime and moreover were intensively exploited in the course of Peter's great work for the modernization and development of Russia (see below *Internal reforms*). Their sufferings, combined with onerous taxation, pro-

Other events during the Northern War

voked a number of revolts, the most important of which were that of Astrakhan (1705–06) and that led by Kondraty Afanasyevich Bulavin in the Don Basin (1707–08). These revolts were cruelly put down.

The Turkish War (1710–13). In the middle of the Northern War, when Peter might have pressed further the advantage won at Poltava, Turkey declared war on Russia. In the summer of 1711 Peter marched against the Turks through Bessarabia into Moldavia, but he was surrounded, with all his forces, on the Prut River. Obliged to sue for peace, he was fortunate to obtain very light terms from the inept Turkish negotiators, who allowed him to retire with no greater sacrifice than the retrocession of Azov. The Turkish government soon decided to renew hostilities; but the Peace of Adrianople (Edirne) was concluded in 1713, leaving Azov to the Turks. From that time on Peter's military effort was concentrated on winning his war against Sweden.

The Tsarevich Alexis and Catherine (to 1718). Peter had a son, the tsarevich Alexis, by his discarded wife Eudoxia. Alexis was his natural heir, but he grew up antipathetic to Peter and receptive to reactionary influences working against Peter's reforms. Peter, meanwhile, had formed a lasting liaison with a low-born woman, the future empress Catherine I, who bore him other children and whom he married in 1712. Pressed finally to mend his ways or to become a monk in renunciation of his hereditary rights (1716), Alexis took refuge in the dominions of the Holy Roman emperor Charles VI, but he was induced to return to Russia in 1718. Thereupon proceedings were brought against him on charges of high treason, and after torture he was condemned to death. He died in prison, presumably by violence, before the formal execution of the sentence.

The Persian campaign (1722–23). Even during the second half of the Northern War, Peter had sent exploratory missions to the East—to the Central Asian steppes in 1714, to the Caspian region in 1715, and to Khiva in 1717. The end of the war left him free to resume a more active policy on his southeastern frontier. In 1722, hearing that the Ottoman Turks would take advantage of Persia's weakness and invade the Caspian region, Peter himself invaded Persian territory. In 1723 Persia ceded the western and southern shores of the Caspian to Russia in return for military aid.

Death. The campaign along the parched shores of the Caspian obviously put a great strain on Peter's health, already undermined by enormous exertions and also by the excesses in which he occasionally indulged himself. In the autumn of 1724, seeing some soldiers in danger of drowning from a ship aground on a sandbank in the Gulf of Finland, he characteristically plunged himself into the icy water to help them. Catching a chill, he became seriously ill in the winter but even so continued to work; indeed, it was at this time that he drew up the instructions for the expedition of Vitus Bering to Kamchatka.

The empire at Peter's death

Peter died in St. Petersburg on February 8 (January 28, O.S.), 1725, leaving an empire that stretched from Arkhangelsk (Archangel) on the White Sea to Mazanderan on the Caspian and from the Baltic Sea to the Pacific Ocean. Though he had in 1722 issued a decree reserving to himself the right to nominate his successor, he did not in fact nominate anyone. His widow Catherine, whom he had crowned as empress in 1724, succeeded him to the temporary exclusion of his grandson, the future Peter II.

Internal reforms. At the beginning of Peter's reign, Russia was backward by comparison with the countries of western Europe. This backwardness inhibited foreign policy and even put Russia's national independence in danger. Peter's aim, therefore, was to overtake the developed countries of western Europe as soon as possible, in order both to promote the national economy and to ensure victory in his wars for access to the seas. Breaking the resistance of the boyars, or members of the ancient landed aristocracy, and of the clergy and severely punishing all other opposition to his projects, he initiated a series of reforms that affected, in the course of 25 years, every field of the national life—administration, industry, commerce, technology, and culture.

The towns. At the beginning of Peter's reign there was already some degree of economic differentiation between the various regions of Russia; and in the towns artisans were establishing small businesses, small-scale production was expanding, and industrial plants and factories were growing up, with both hired workers and serfs employed. There was thus a nascent bourgeoisie, which benefitted considerably from Peter's plans for the development of the national industry and trade. The reform of the urban administration was particularly significant.

Administrative reforms

By a decree of 1699, townspeople (artisans and tradesmen) were released from subjection to the military governors of the provinces and were authorized to elect municipalities of their own, which would be subordinated to the Moscow municipality, or *ratusha*—the council of the great merchant community of the capital. This reform was carried further in 1720, with the establishment of a chief magistracy in St. Petersburg, to which the local town magistracies and the elected municipal officers of the towns (mayors, or *burmistry*; and councillors, or *ratmany*) were subordinated.

All townspeople, meanwhile, were divided between "regulars" and "commons" (inferiors). The regulars were subdivided between two guilds—the first comprising rich merchants and members of the liberal professions (doctors, actors, and artists); the second, artisans (classified according to their vocations) and small tradesmen. A merchant belonged to the first or to the second guild according to the amount of his capital; and those who were also manufacturers had special privileges, coming under the jurisdiction of the College of Manufactures and being exempt from the billeting of troops, from elective rotas of duty, and from military service. The commons were hired labourers, without the privileges of regulars.

Thanks to the reforms, the economic activity and the population of the towns increased. Anyone engaged in trade was legally permitted to settle in a town and to register himself in the appropriate category, and there was a right of "free commerce for people of every rank."

The provinces and the districts. In order to create a more flexible system of control by the central power, Russia was territorially divided in 1708 into eight guberny, or governments, each under a governor appointed by the tsar and vested with administrative, military, and judicial authority. In 1719 these *guberny* were dissolved into 50 *provintsy*, or provinces, which in turn were subdivided into districts. The census of 1722, however, was followed by the substitution of a poll tax for the previous hearth tax; and this provoked a wave of popular discontent, against which Peter decided to distribute the army regiments (released from active service by the Peace of Nystad) in garrisons throughout the country and to make their maintenance obligatory on the local populations. Thus came into being the "regimental districts," which did not coincide with the administrative. The regimental commanders, with their own sphere of jurisdiction and their own requirements, added another layer to the already complex system of local authority.

The central government. In the course of Peter's reign, medieval and obsolescent forms of government gave place to effective autocracy. In 1711 he abolished the *boyarskaya duma*, or boyar council, and established by decree the Senate as the supreme organ of state—to coordinate the action of the various central and local organs, to supervise the collection and expenditure of revenue, and to draft legislation in accordance with his edicts. Martial discipline was extended to civil institutions, and an officer of the guards was always on duty in the Senate. From 1722, moreover, there was a procurator general keeping watch over the daily work of the Senate and its chancellery and acting as "the eye of the sovereign."

When Peter came to power the central departments of state were the *prikazy*, or offices, of which there were about 80, functioning in a confused and fragmented way. To replace most of this outmoded system, Peter in 1718 instituted 9 "colleges" (*kollegy*), or boards, the number of which was by 1722 expanded to 13. Their activities were controlled, on the one hand, by the General Regula-

tion and, on the other, by particular regulations for individual colleges, and indeed there were strict regulations for every branch of the state administration. Crimes against the state came under the jurisdiction of the Preobrazhensky Office, responsible immediately to the tsar.

Europeanization of industry

Industry. A secondary purpose of Peter's Grand Embassy to western Europe in 1697 (see above *The Grand Embassy*) had been to obtain firsthand acquaintance with advanced industrial techniques, and the exigencies of his great war against Sweden, from 1700, made industrial development an urgent matter. In order to provide armaments and to build his navy (Russia had virtually no warships at all), metallurgical and manufacturing industries on a grand scale had to be created; and Peter devoted himself tirelessly to meeting these needs. Large capital investments were made, and numerous privileges were accorded to businessmen and industrialists. These privileges included the right to buy peasant serfs for labour in workshops, with the result that a class of "enlisted" serfs came into existence, living in specified areas and bound to the factories. The methods of other countries were further studied, and foreign experts were invited to Russia. The overall result was satisfactory: the army and the navy were, of course, supplied with their material needs; a great number of manufacturing establishments were founded (mainly with serf labour); the metallurgical industry was so far advanced that by the middle of the 18th century Russia led Europe in this field; and the foreign-trade turnover was increased sevenfold in the course of the reign.

The armed forces. Peter established a regular army on completely modern lines for Russia in the place of the unreliable *streltsy* and the militia of the gentry. While he drew his officers from the nobility, he conscripted peasants and townspeople into the other ranks. Service was for life. The troops were equipped with flintlock firearms and bayonets of Russian make; uniforms were provided; and regular drilling was introduced. For the artillery, obsolete cannons were replaced with new mortars and guns designed by Russian specialists or even by Peter himself (he drew up projects of his own for multicannon warships, fortresses, and ordnance). The Army Regulations of 1716 were particularly important; they required officers to teach their men "how to act in battle," "to know the soldier's business from first principles and not to cling blindly to rules," and to show initiative in the face of the enemy. For the navy, Peter's reign saw the construction, within a few years, of 52 battleships and hundreds of galleys and other craft; thus a powerful Baltic fleet was brought into being. Several special schools prepared their pupils for military or naval service and finally enabled Peter to dispense with foreign experts.

Cultural and educational measures. From January 1, 1700, Peter introduced a new chronology, making the Russian calendar conform to European usage with regard to the year, which in Russia had hitherto been numbered "from the Creation of the World" and had begun on September 1 (he adhered however to the Julian Old Style as opposed to the Gregorian New Style for the days of the month). In 1710 the Old Church Slavonic alphabet was modernized into a secular script.

Peter was the first ruler of Russia to sponsor education on secular lines and to bring an element of state control into that field. Various secular schools were opened; and since too few pupils came from the nobility, the children of soldiers, officials, and churchmen were admitted to them. In many cases, compulsory service to the state was preceded by compulsory education for it. Russians were also permitted to go abroad for their education and indeed were often compelled to do so (at the state's expense). The translation of books from western European languages was actively promoted. The first Russian newspaper, *Vedomosti* ("Records"), appeared in 1703. The Russian Academy of Sciences was instituted in 1724.

Beside his useful measures, Peter often enforced superficial Europeanization rather brutally; for example, when he decreed that beards should be shorn off and Western dress worn. He personally cut the beards of his boyars and the skirts of their long coats (*kaftany*). The Raskolniki (Old Believers) and merchants who insisted on keeping their beards had to pay a special tax, but peasants and the Orthodox clergy were allowed to remain bearded.

Control over church and nobility

The church. In 1721, in order to subject the Orthodox Church of Russia to the state, Peter abolished the Patriarchate of Moscow. Thenceforward the patriach's place as head of the church was taken by a spiritual college, namely the Holy Synod, consisting of representatives of the hierarchy obedient to the tsar's will. A secular official—the *ober-prokuror*, or chief procurator—was appointed by the tsar to supervise the Holy Synod's activities. The Holy Synod ferociously persecuted all dissenters and conducted a censorship of all publications.

Priests officiating in churches were obliged by Peter to deliver sermons and exhortations to make the peasantry "listen to reason" and to teach such prayers to children that everyone would grow up "in fear of God" and in awe of the tsar. The regular clergy were forbidden to allow men under 30 years old or serfs to take vows as monks.

The church was thus transformed into a pillar of the absolutist regime. Partly in the interests of the nobility, the extent of land owned by the church was restricted; Peter disposed of ecclesiastical and monastic property and revenues at his own discretion, for state purposes.

The nobility. Peter's internal policy served to protect the interest of Russia's ruling class—the landowners and the nascent bourgeoisie. The material position of the landed nobility was strengthened considerably under Peter. Almost 100,000 acres of land and 175,000 serfs were allotted to it in the first half of the reign alone. Moreover, a decree of 1714 that instituted succession by primogeniture and so prevented the breaking up of large properties also removed the old distinction between *pomestya* (lands granted by the tsar to the nobility in return for service) and *votchiny* (patrimonial or allodial lands) so that all such property became hereditary.

Moreover, the status of the nobility was modified by Peter's Table of Ranks (1722). This replaced the old system of promotion in the state services, which had been according to ancestry, by one of promotion according to services actually rendered. It classified all functionaries —military, naval, and civilian alike—in 14 categories, the 14th being the lowest and the 1st the highest; and admission to the 8th category conferred hereditary nobility. Factory owners and others who had risen to officer's rank could accede to the nobility, which thus received new blood. The predominance of the boyars ended.

Personality and achievement. Peter was of enormous height, more than six and one-half feet (two metres) tall; he was handsome and of unusual physical strength. Unlike all earlier Russian tsars, whose Byzantine splendours he repudiated, he was very simple in his manners; for example, he enjoyed conversation over a mug of beer with shipwrights and sailors from the foreign ships visiting St. Petersburg. Restless, energetic, and impulsive, he did not like splendid clothes that hindered his movements; often he appeared in worn-out shoes and an old hat, still more often in military or naval uniform. He was fond of merrymaking and knew how to conduct it, though his jokes were frequently crude; and he sometimes drank heavily and forced his guests to do so too. A just man who did not tolerate dishonesty, he was terrible in his anger and could be cruel when he encountered opposition: in such moments only his intimates could soothe him—best of all his beloved second wife, Catherine, whom people frequently asked to intercede with him for them. Sometimes Peter would beat his high officials with his stick, from which even Prince A.D. Menshikov, his closest friend, received many a stroke. One of Peter's great gifts of statesmanship was the ability to pick talented collaborators for the highest appointments, whether from the foremost families of the nobility or from far lower levels of society.

As a ruler, Peter often used the methods of a despotic landlord—the whip and arbitrary rule. He always acted as an autocrat, convinced of the wonder-working power of compulsion by the state. Yet with his insatiable capacity for work he saw himself as the state's servant, and whenever he put himself in a subordinate position he

would perform his duties with the same conscientiousness that he demanded of others. He began his own army service in the lowest rank and required others likewise to master their profession from its elements upward and to expect promotion only for services of real value.

Peter's influence on Russia

Peter's personality left its imprint on the whole history of Russia. A man of original and shrewd intellect, exuberant, courageous, industrious, and iron willed, he could soberly appraise complex and changeable situations so as to uphold consistently the general interests of Russia and his own particular designs. He did not completely bridge the gulf between Russia and the Western countries, but achieved considerable progress in development of the national economy and trade, education, science and culture, and foreign policy. Russia became a great power, without whose concurrence no important European problem could thenceforth be settled. His internal reforms achieved progress to an extent that no earlier innovator could have envisaged (see further RUSSIA AND THE SOVIET UNION, HISTORY OF).

BIBLIOGRAPHY. There are numerous studies in Russian on Peter I and his age. *Письма и бумаги Императора Петра Великаво,* 11 vol. (1887–1962), contains valuable documents on Russian history up to 1711. Other works particularly recommended include: М.М. БОГОСЛОВСКИЙ, *Петр I,* 5 vol. (1940–48), a detailed biography of Peter I until 1700; С.М. СОЛОВЬЕВ, *Публичныя чтения о Петре Великом* (1872), a famous Russian historian on the role of Peter I; И.И. ГОЛИКОВ, *Деяния Петра Великаво,* 30 vol. (1788–97), on his reforms; and Л.А. НИКИФОРОВ, *Русско-английские отношения при Петре I* (1950). Works in English: B.H. SUMNER, *Peter the Great and the Ottoman Empire* (1949, reprinted 1965) and *Peter the Great and the Emergence of Russia* (1950); IAN GREY, *Peter the Great, Emperor of All Russia* (1960). See also R. WITTRAM, *Peter der Grosse: Der Eintritt Russlands in die Neuzeit* (1954).

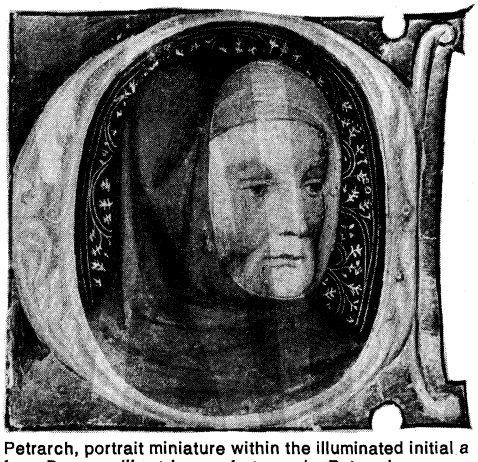
(L.A.N.)

Petrarch

Francesco Petrarch, a scholar, poet, Humanist, and promoter of new ideas, stands between the dying medieval period and the Renaissance, which he ushered in. Although it cannot be claimed that he had much influence on the economic and political forces of his time, the artistic and moral outlook that characterizes the new age can be traced directly to this one man's mind and work. The poet Dante, his older contemporary, belongs entirely to the Middle Ages, and for him the beauty of poetry lay in its *artes*—colours, figures, resonances, verbal and metrical complexities. After Petrarch, beauty was seen to lie in the "classical" virtues of purity, simplicity, and proportion. Morally, Dante was concerned primarily with being theologically right; Petrarch and those who followed him cared for the enrichment of the human mind and conscience.

Education and early poems. Petrarch was a Florentine, but once removed. His father, a lawyer, had been obliged to leave Florence in 1302 and had moved to Arezzo, where Petrarch was born on July 20, 1304. The family eventually moved to Avignon (1312), in the Provence region of southern France, the home of the exiled papal court, at which an Italian lawyer might hope to find employment. Petrarch's first studies were at Carpentras (about 15 miles [24 kilometres] from Avignon itself), and at his father's insistence he was sent to study law at Montpellier (1316) and from there returned to Italy with his younger brother Gherardo to continue these studies at Bologna (1320). But already he was developing what, in a later letter, he described as "an unquenchable thirst for literature." His devotion to it is illustrated by a well-known anecdote: his father, anxious for his son to concentrate on the potentially lucrative study of the law, threw the books that might distract him on the fire; but faced with his son's passionate entreaties he relented sufficiently to rescue copies of Virgil and Cicero.

By courtesy of the Biblioteca Nazionale, Venice



Petrarch, portrait miniature within the illuminated initial *a* from *De remediis utriusque fortunae* by Petrarch, completed 1366. In the Biblioteca Nazionale Marciana, Venice (Cod. Marc. Lat. VI, 86).

Earliest surviving poems

Petrarch's earliest surviving poems, on the death of his mother (1318 or 1319; *Epistolae metricae,* i, 7), date from the Montpellier and Bologna period, though like all Petrarch's work they were heavily revised later. Meanwhile, his knowledge and love of the classical authors increased, he made his acquaintance with the new vernacular poetry that was being written, and perhaps he even experimented in it himself.

After his father's death, in 1326, Petrarch was free to abandon his law studies and pursue his own interests. Returning to Avignon, he took minor ecclesiastical orders (which enabled him to hold benefices and canonries without much attendant responsibility or obligation) and entered the household of the influential Cardinal Giovanni Colonna. Petrarch enjoyed life in Avignon, and there is a famous description of him and his brother as dandies in its polished courtly world. But he was also making a name there for his scholarship and the elegance of his culture.

As well as a love of literature, Petrarch also had during his early youth a deep religious faith, a love of virtue, and an unusually deep perception of the transitory nature of human affairs. There now followed the reaction—a period of dissipation—which also coincided with the beginning of his famous chaste love for a woman known now only as Laura. Vain attempts have been made to identify her, but Petrarch himself kept silent about everything concerning her civil status, as though he thought it unimportant. He first saw her in the Church of St. Clare at Avignon on April 6, 1327, and loved her, although she was outside his reach, almost until his death, in 1374. From this love there springs the work for which he is most celebrated, the Italian poems (*Rime*), which he affected to despise as mere trifles in the vulgar tongue but which he collected and revised throughout his life.

Meeting with Laura

Classical studies and career (1330–40). He spent the summer of 1330 at Lombez, a little Gascon town, the bishop of which was an old friend from Bologna, Giacomo Colonna. In 1335 he received a canonry there but continued to reside at Avignon in the service of the Cardinal, with whom he stayed until 1337. Quite apart from his love for Laura, this period was an important one for Petrarch. These were years of ambition and unremitting study (notably in the field of classical Latin). They were also years of travel. In 1333 his journeying took him through France, Flanders, Brabant, and the Rhineland, where he visited men of learning and searched monastic libraries for "lost" classical manuscripts (in Liège he discovered copies of two speeches by Cicero). In Paris he was given a copy of the *Confessions* of St. Augustine by a friend and spiritual confidant, the Augustinian monk Dionigi of Sansepolcro, and he was to use this more and more as the breviary of his spiritual life.

Petrarch's synthesis of the classical tradition and Christianity

These experiences bring Petrarch's mission as a stubborn advocate of the continuity between classical culture and the Christian message more sharply into focus. By making a synthesis of the two seemingly conflicting ideals —regarding the one as the rich promise and the other as its divine fulfillment—he can claim to be the founder and great representative of the movement known as European Humanism. He rejected the sterile argumentation and

endless dialectical subtleties to which medieval Scholasticism had become prey and turned back for values and illumination to the moral weight of the classical world. The list of his books (1333) reveals that Cicero, Virgil, and Horace kept company with St. Augustine and St. Jerome; no Scholastic work is mentioned. In 1337 he visited Rome for the first time, to be stirred among its ruins by the evident grandeur of its past. On returning to Avignon he sought a refuge from its corrupt life—the papacy at this time was wholly absorbed in secular matters—and a few miles to the east found his "fair transalpine solitude" of Vaucluse, which was afterward to become a much-loved place of retreat.

The chronology of Petrach's writings is somewhat complicated by his habit of revising, often very extensively, everything he wrote. (This also means that it is very difficult to perceive the real movement of Petrarch's mind: attitudes properly belonging to his later work often seem to have been anticipated in his earlier writings.) By the time he discovered Vaucluse, however, he had written a good many of the individual poems that he was to include in the *Epistolae metricae* ("66 letters" in Latin hexameter verses) and some of the vernacular *Rime* inspired by his love for Laura. At Vaucluse he began to work on *Africa,* an epic poem on the subject of the Second Punic War, intended to vie with Virgil's *Aeneid* but never completed. He also began work on *De viris illustribus,* intended as a series of biographies of heroes from Roman history (but later modified to include famous men of all time, beginning with Adam, as Petrarch's desire to emphasize the continuity among ideals of the Old Testament, of the classical world, and of Christianity increased).

Moral and literary evolution (1340–46). Meanwhile, his reputation as a scholar was spreading; in September 1340 he received invitations from Paris and Rome to be crowned as poet. He had perhaps sought out this honour, partly from ambition but mainly in order that the rebirth of the cult of poetry after more than 1,000 years might be fittingly celebrated. He had no hesitation in choosing Rome, and accordingly he was crowned on the Capitoline Hill on April 8, 1341, afterward placing his laurel wreath on the tomb of the Apostle in St. Peter's Basilica: again, the symbolic gesture linking the classical tradition with the Christian message.

Poet laureate

From Rome he went to Parma and the nearby solitude of Selvapiana, returning to Avignon in the autumn of 1343. It is generally believed that he went through some kind of moral crisis at this time, rooted in his inability to make his life conform to his religious faith and possibly heightened by his brother's decision to solve his own problems by entering a Carthusian monastery. At any rate, this is a common reading of the *Secretum meum,* written between the autumn of 1342 and the winter of 1343 (later revised). An autobiographical treatise, it consists of three dialogues between Petrarch and St. Augustine in the presence of Truth: he identifies his "secret" as an infirmity of the will that prevents him from achieving a noble life, even while aspiring to it; he accuses himself of all seven deadly sins except envy; he finally convinces himself that his love—even though chaste and directed toward a pure woman—and his desire for glory are sinful. But he declares himself unable to reform. Even so, the work's real conclusion is its most important aspect, one truly Humanistic: for he maintains hope that, even amidst worldly preoccupations and error, even while absorbed in himself and his own affairs, a man might still find a way to God. Thus, Petrarch's spiritual "problem" found a coherent solution, one that can be said to express the Petrarchan vision and the Humanist's religious and moral outlook. It reinforces his experience the year before, when, on climbing Mt. Ventoux in France, he had drawn from his pocket the *Confessions*; his eye fell upon the passage where St. Augustine speaks of men who ". . . go to admire the high mountains and the immensity of the ocean and the course of the heaven . . . and neglect themselves."

It was therefore an evolution—both moral and literary—rather than a "crisis" that made Petrarch decide his love for Laura was love for the creature rather than for the Creator and therefore wrong—proof of his attachment to the world. It was an evolution in his thinking that led him to break through the barriers of his too-exclusive admiration for antiquity and to admit other authoritative voices. It was now, for example, that *De viris* was enlarged to include material from sacred as well as secular history, while in the *De vita solitaria* (1346; later revised and enlarged) he developed the theoretical basis and description of the "solitary life" whereby man enjoys the consolations of nature and study together with those of prayer.

The Petrarchan myth of "solitude"

Break with his past (1346–53). The events of the next few years are fundamental to his biography, both as a man and as a writer. In the first place, he became enthusiastic for the efforts of Cola di Rienzo to revive the Roman republic and restore popular government in Rome—a sympathy that divided him still more sharply from the Avignon court and in 1346 even led to the loss of Cardinal Colonna's friendship. The Plague of 1348, known as the Black Death, saw many friends fall victim, including Laura, who died on April 6, the anniversary of Petrarch's first seeing her. Finally, in the jubilee year of 1350 he made a pilgrimage to Rome and later assigned to this year his renunciation of sensual pleasures.

These are the landmarks of Petrarch's career, but the time in between was filled with diplomatic missions, study, and immense literary activity. In Verona in 1345 he made his great discovery of the letters of Cicero to Atticus, Brutus, and Quintus, which allowed him to penetrate the surface of the great orator and see the man himself. The letters spurred him on to write epistles to the ancient authors whom he loved and to make a collection of his own letters that he had scattered among his friends. These great collections (*Familiarium rerum libri xxiv; Senilium rerum libri*) record not only Petrarch's genius for friendship but also all those shifts in attitude by which he left behind the Middle Ages and prepared for the Renaissance. Toward the end of 1345 he returned again to the peace of Vaucluse and spent two years there, chiefly revising *De vita solitaria* but also developing the theme of solitude in a specifically monastic context, in *De otio religioso*. Between November 1347 and his pilgrimage to Rome in 1350 he was also in Verona, Parma (where the news of Laura's death reached him; he noted it in his copy of Virgil, in order that he might constantly be reminded of the decay of all earthly goods), and Padova. Much of the time was spent in advancing his career in the church; the manoeuvring and animosities this involved resulted in an intense longing for the peace of Vaucluse; not even a visit from his lifelong friend the poet Boccaccio, who offered him a chair to be established under his guidance in the university of Florence, could deflect him. He left Rome in May 1351 for Vaucluse.

Rime. Here he worked on a new plan for the *Rime.* The project was divided into two parts: the *Rime in vita di Laura* ("Poems During Laura's Life") and the *Rime in morte di Laura* ("Poems after Laura's Death"), which he now selected and arranged to illustrate the story of his own spiritual growth. The choice of poems was further governed by an exquisite aesthetic taste and by a preference for an approximately chronological arrangement, from the description of his falling in love to his final invocation to the Virgin; from his "youthful errors" to his realization that "all worldly pleasure is a fleeting dream"; from his love for this world to his final trust in God. The theme of his *Canzoniere* (as the poems are usually known) therefore goes beyond the apparent subject matter, his love for Laura. For the first time in the history of the new poetry, lyrics are held together in a marvellous new tapestry, possessing its own unity (and similar to one that Petrarch was constructing, on the Ciceronian model, from his letters). By selecting all that was most polished and at the same time most vigorous in the lyric tradition of the preceding two centuries and filtering it through his new appreciation of the classics, he not only bequeathed to humanity the most limpid and yet passionate, precise yet suggestive, expression of love and grief, of the ecstasies and sorrows of man, but

The *Rime in vita* and *Rime in morte*

also created with his marvellous sensibility the form and language of the modern lyric, to provide a common stock for lyric poets of the whole of Europe.

He also continued work on the *Metricae,* begun in 1350; he embarked on a polemic against the conservative enemies of his new conception of education, which rejected the prevailing Aristotelianism of the schools and restored the spiritual worth of classical writers—the new studies to be called *litterae humanae,* "humane letters." He also began work on his poem *Trionfi,* a more generalized version of the story of the human soul in its progress from earthly passion toward fulfillment in God (in it, Love is conquered by Chastity, Chastity by Death, Death by Glory, Glory by Time, and Time by Eternity).

Later years (1353–74). But the death of his closest friends, dislike of the newly elected pope, Innocent VI, increasingly bitter relations with the Avignon court, all finally determined Petrarch to leave Provence and settle in Italy. He found rooms in Milan and stayed there for most of the next eight years (in spite of harsh criticisms for accepting hospitality from the Visconti family, enemies of Florence, for whom he carried out a number of important diplomatic missions). During these eight years he also completed the first proper edition of the *Rime,* continued assiduously with the *Familiares,* worked on the *Trionfi,* and set in order many of his earlier writings.

Early in 1361 he went to Padova, hoping to escape the Plague. He remained there until September 1362, when, again a fugitive from the Black Death, he sought shelter in Venice. He was given a house, and in return Petrarch promised to bequeath all his books to the republic. He was joined by his daughter Francesca, and the tranquil happiness of her little family gave him great pleasure. He was visited by his dearest and most famous friends (by the great chancellor Benintendi de' Ravegnani, by Padovan scholars, by Boccaccio, who in 1366 presented him with a long-desired Latin translation of Homer's poems); he was invited to play an honourable part in the life and politics of the city; he worked peacefully but with great concentration at the definitive versions of the *Rime,* the *Familiares,* the *Seniles,* and the *De remediis utriusque fortunae,* a collection of 254 treatises on how to bear with both good fortune and bad. Nevertheless, after receiving an insult from four young men who followed the Arab "naturalist" interpretation of Aristotle's work (they called Petrarch "a good man . . . but illiterate and in fact an idiot" because he believed in Christian doctrines rather than those attributed to Aristotle), Petrarch was induced to move back to Padova in 1367. He remained there until his death, dividing his time from 1370 between Padova and Arquà, in the neighbouring Colli Eugenei (Euganean hills), where he now had a little house. There he wrote the defense of his Humanism against the critical attack from Venice, *De sui ipsius et multorum ignorantia.* He was still in great demand as a diplomat; in 1370 he was called to Rome by Urban V, and he set off eager to see the fulfillment of his great dream of a new Roman papacy, but at Ferrara he was seized by a stroke. Yet he did not stop working; in addition to revision he composed more minor works and added new sections to his *Posteritati,* an autobiographical letter to posterity that was to have formed the conclusion to his *Seniles*; he also composed the final sections of the *Trionfi.*

Petrarch died during the night of July 18–19, 1374, while working in his study at Arquà, and was found the next morning, his head resting on a manuscript of Virgil.

Petrarch's achievement

The hallmark of Petrarch's thought was a deep consciousness of the past as the nutriment of the present. His abiding achievement was to recognize that, if there is a Providence that guides the world, then it has set man at the centre. Petrarch provided a theoretical basis for the enrichment of man's life. But, even more important, the Humanist attitudes of the Italian 15th century that led into the Renaissance would not have been possible without him.

MAJOR WORKS

The chronology of Petrarch's works is difficult to ascertain because the poet habitually and heavily revised everything he wrote. His collected works (*Opera*) were published in 1554 and 1581.

LATIN WORKS: *Epistolae metricae* (he worked on the collection in three phases: 1350, 1357, and 1363; ed. by D. Rossetti, *Poesie minori del Petrarca,* 1831); *Africa* (begun 1338 or 1339; published posthumously 1396; critical edition by N. Festa, 1926); *De viris illustribus* (begun 1338–39; revised 1341–43, and again 1351–53; later replanned with the title *Quorundam virorum illustrium epithoma,* preface by Petrarch, completed by his disciple Lombardo della Seta; ed. by L. Razzolini, 1874–79); *Secretum meum* (1342–43, revised 1353–58; ed. by E. Carrara in *Prose,* 1955); *De vita solitaria* (1346 and later—one addition of 1371; ed. by G. Martellotti in *Prose,* 1955); *De otio religioso* (1347, revised and enlarged 1357; ed. by G. Rotondi, 1958); *Familiarium rerum libri xxiv* (arrangement of the letters begun 1351–53, i-viii; 1359–60, ix-xix; 1363–64, xx-xxiii; 1366, xxiv; critical edition by V. Rossi and U. Bosco, 1933–42); *De remediis utriusque fortunae* (begun 1354; published 1366; Petrarch probably returned to it later); *Senilium rerum libri* (after 1361 to 1374); *Posteritati* (first draft before 1367; additions 1370–72; ed. by E. Carrara and P.G. Ricci in *Prose,* 1955); *De sui ipsius et multorum ignorantia* (written 1367, published 1371; ed. by L.M. Cappelli, 1906).

OTHER LATIN WORKS: Prayers (1335–38); *Epistolae sine nomine* (written 1342–58; collected 1359–60; ed. by P. Piur, 1925); *Rerum memorandarum libri* (1343–45; critical edition by G. Billanovich, 1945); *Bucolicum carmen* (1346–48; definitive version 1364; ed. by A. Avena, 1906); *Salmi penitenziali* (1347; ed. by H. Cochin, 1929); *Invectivarum contra medicum quendam libri* iv (1352–55; ed. by P.G. Ricci, 1950); *Invectiva contra quendam magni status hominem sed nullius scientiae aut virtutis* (1355; ed. by P.G. Ricci, 1949); *Itinerarium Syriacum* (*Itinerarium breve de Janua usque ad Jerusalem et Terram Sanctam*) (1358 and later); *Invectiva contra eum qui maledixit Italiae* (1373); *Epistolae variae* (65 letters not included by Petrarch in his collections).

ITALIAN WORKS: *Rerum vulgarium fragmenta* (*Rime in vita di Laura* and *Rime in morte di Laura,* or *Canzoniere*; first edition, 1360; second edition, 1373; definitive edition, preserved in Cod. Vat. Lat. 3195; ed. by G. Contini, 1949; commentaries by G. Carducci and S. Ferrari, 6th edition, 1957; ed. by E. Chiorboli, 1924); *Rime disperse* (or *Estravaganti,* poems not included by Petrarch in his main collection, many of doubtful authenticity, ed. by A. Solerti, 1909); *Trionfi* (begun 1351, revised between 1356–1374, unfinished, published posthumously; critical edition by K. Appel, 1901; to be read in conjunction with *Un inedito petrarchesco,* ed. by R. Weiss, 1950).

BIBLIOGRAPHY. M. FOWLER, *Catalogue of the Petrarch Collection Bequeathed by Willard Fiske* (1916), an exhaustive guide to all the early editions of Petrarch and works concerning him; articles and reviews appear regularly in the journal *Studi petrarcheschi* (1948–).

Editions: Edizione nazionale delle opere di Francesco Petrarca (1926–), includes so far: *Familiares, Africa, Rerum memorandarum libri,* and *De viris illustribus; Rime, Trionfi e Poesie latine,* ed. by F. NERI *et al.* (1951); *Prose,* ed. by G. MARTELLOTTI, E. CARRARA, and P.G. RICCI (1955), the most compendious modern editions of Petrarch.

Selections: J. VIANEY, *Le Pétrarquisme en France au XVI^e siècle* (1909); L. FORSTER, *The Icy Fire: Five Studies in European Petrarchism* (1969); T.G. BERGIN (comp.), *Rhymes: A Selection of Translations* (1955), introductions to the vast theme of imitation and translation of Petrarch in European love poetry; DAVID THOMPSON (ed.), *Petrarch: A Humanist Among Princes* (1971), an anthology of selections from Petrarch's works and letters.

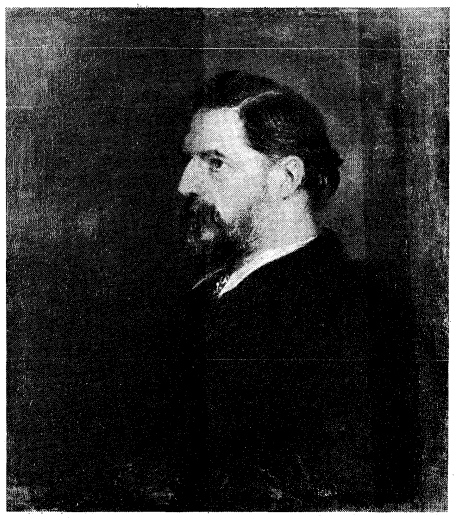
Biographical and critical studies: ABBE DE SADE, *Mémoires pour la vie de François Pétrarque,* 3 vol. (1764–67), the fundamental biography of Petrarch; E.H. WILKINS, *The Making of the Canzoniere and Other Petrarchan Studies* (1951), *Studies in the Life and Works of Petrarch* (1955), *Petrarch's Eight Years in Milan* (1958), *Petrarch's Later Years* (1959), and *Life of Petrarch* (1961), the major American contribution to Petrarch scholarship; PIERRE DE NOLHAC, *Pétrarque et l'humanisme,* 2nd ed., 2 vol. (1907), *Petrarch and the Ancient World* (1907), an abridged version of the great work by de Nolhac; J.H. WHITFIELD, *Petrarch and the Renascence* (1943, reprinted 1966); HANS BARON, *From Petrarch to Leonardo Bruni* (1968); G. BILLANOVICH, *Petrarca letterato* (1947); U. BOSCO, *Francesco Petrarca,* 2nd ed. (1961), the best general Italian presentation of Petrarch; N. SAPEGNO, articles on Petrarch in *Il Trecento,* 3rd ed. (1966); and in vol. 2 of the CECCHI-SAPEGNO *Storia della letteratura italiana* (1965), with copious bibliography.

(J.H.W.)

Petrie, Sir Flinders

Sir Flinders Petrie, a renowned English archaeologist and Egyptologist, made valuable contributions to the techniques and methods of field excavation and invented a method of sequence dating that made possible the reconstruction of history from the remains of ancient cultures.

By courtesy of the National Portrait Gallery, London



Petrie, oil painting by G.F. Watts (1817–1904). In the National Portrait Gallery, London.

Born at Charlton, near Greenwich, London, on June 3, 1853, William Matthew Flinders Petrie was named for his maternal grandfather, Matthew Flinders, British navigator, pioneer hydrographer, and explorer of Australia and Tasmania. A frail child, Petrie was privately educated, early developing archaeological and ethnological interests, particularly in the area of ancient weights and measures, and in Egyptology.

At the age of 24, Petrie wrote *Inductive Metrology, or the Recovery of Ancient Measures from the Monuments*, a work that represented a new approach to archaeological study. Fieldwork done at various locations in Britain, including Stonehenge, enabled him to determine by mathematical computations the unit of measurement for the construction of the monument. His *Stonehenge: Plans, Description, and Theories* was published in 1880, and in that same year he began the surveys and excavation of the Great Pyramid at Giza, which initiated his four decades of exploration in the Middle East.

Excavations and discoveries

During the 1884 excavation of the Temple of Tanis, he found fragments of a colossal statue of Ramses II. In 1885 and 1886, at Naukratis and Daphnae in the Nile Delta, he uncovered painted pottery by which he proved that those sites had been trading colonies for the ancient Greeks. It was this discovery that caused him to believe that history could be reconstructed by a comparison of potsherds (pottery fragments) at various levels of an excavation.

He first applied his principle of sequence dating in Palestine, at Tel Ḥasi, south of Jerusalem. In 1890, in a period of only six weeks, the indefatigable excavator found a series of occupations for which he supplied tentative dates of habitation. His work at the hill site marked the second stratographic study in archaeological history—the first was carried out at Troy by Heinrich Schliemann. The excavations of these two men marked the beginning of the examination of successive levels of a site, rather than the previously practiced method of haphazard digging, which had produced only a jumble of unrelated artifacts. Most of Petrie's contemporaries in archaeology questioned his hypothesis that chronology could be established by potsherds, whether painted or undecorated. But, with the progressive sophistication of the science of archaeology, examination and classification of broken pottery became, as it is now, routine procedure.

He made other important discoveries in the Fayyūm region of Egypt. At Gurob, he found numerous papyri and Aegean pottery that substantiated dates of ancient Greek civilizations, including the Mycenaean. At the Pyramid of Hawara he searched through the tomb of Pharaoh Amenemhet III to discover how grave robbers could have found the tomb's opening and made their way through the labyrinth surrounding the two sarcophagi (stone coffins) that they emptied. He concluded that they must have been given the master plan by an informer. At al-Fayyūm also he made a rich find of 12th-dynasty jewelry (housed at the Metropolitan Museum of New York since 1919). He was delighted by his discovery of the earliest known Egyptian reference to Israel on the stele (a stone slab monument) of Merneptah, pharaoh of ancient Egypt from *c.* 1236–1223 BC.

In 1892 Petrie was made Edwards professor of Egyptology at University College, London, and held the position until 1933, when he became professor emeritus. In 1894, he founded the Egyptian Research Account, which in 1905 became the British School of Archaeology.

Petrie added to the knowledge of the pyramid builders during his exploration of the necropolis of Abydos, holy city of the cult of Osiris, god of the dead. At Tell el-Amarna, he excavated the city of Akhenaten, or Amenhotep IV, ruler of Egypt from 1379–1362 BC, revealing the now-famous painted pavement and other artistic wonders of the Amarna age (14th century BC). Three thousand graves found by Petrie at Naqādah, northeast of Thebes, were identified as those of primitive ancient Egyptians.

Methods and Aims in Archaeology

In 1904, Petrie published *Methods and Aims in Archaeology*, the definitive work of his time, in which he lucidly defined the goals and methodology of his profession along with the more practical aspects of archaeology —details of excavation, including the use of cameras in the field. With uncommon insight, he noted that research results were dependent on the personality of the archaeologist, who, in addition to broad knowledge, had to have insatiable curiosity. His own abundance of that characteristic was never questioned.

Inscriptions Petrie found on the Sinai Peninsula represented an intermediate stage (not later than 1500 BC) of written communication between Egyptian hieroglyphics and the Semitic alphabet. Although he wrote *The Formation of the Alphabet* (1912), language was not Petrie's forte, and he depended on a sixth sense for free translation of inscriptions and for establishing dates through the study of the forms of hieroglyphs.

Petrie was knighted in 1923. Under the auspices of the American School of Research, he excavated in Palestine from 1927 until 1938, when he was 85. In those years, again at Tel Ḥasi, he uncovered the ruins of ten cities. His scientific methods provided the guidelines for all subsequent Palestinian excavations. He died at Jerusalem, July 28, 1942, at the age of 89.

BIBLIOGRAPHY. W.F. ALBRIGHT, *History, Archaeology and Christian Humanism* (1964), in the course of discussing specific contributions to archaeology, evaluates Petrie's work in the development of archaeological method. MARGARET ALICE MURRAY, *The Splendor That Was Egypt*, rev. ed. (1963), is dedicated to her longtime colleague at University College, London, with tribute to the man and the archaeologist. He is given similar recognition in C.L. WOOLLEY, *Dead Towns and Living Men*, rev. ed. (1954). J. LIEBOVITCH, *Ancient Egypt* (1938), a history of the antiquities in the Egyptian Museum, Cairo, illuminates the finds of Petrie and other Egyptologists and is profusely illustrated, including alphabetic signs. KURT W. MAREK (C.W. CERAM, pseud.), *Götter, Gräber, und Gelehrte: Roman der Archäologie* (1963; Eng. trans., *Gods, Graves and Scholars*, 2nd ed. rev., 1967), considers Petrie the outstanding interpreter of discoveries; also contains a chronological table of Egyptian dynasties and a separate bibliography for the section on the pyramids.

(G.J.P.)

Petroleum

Petroleum can be applied as an inclusive term to the complex mixture of hydrocarbons that occurs in the earth in liquid, gaseous, or solid forms. As generally used, however, petroleum is a liquid (crude oil) that is recovered

from some depth within the Earth through boreholes. The liquid and gaseous phases of petroleum occur naturally underground, primarily within the pore spaces of sedimentary rocks, and under normal conditions they migrate freely until stopped by a subsurface obstruction or trap. In its liquid and gaseous forms, petroleum is the most important of the primary fossil fuels that supply the bulk of the world's energy requirements.

History of petroleum usage. The word petroleum (literally "rock oil," from the Greek and Latin words *petra*, "rock" or "stone," and *oleum*, "oil") was first used by the German mineralogist Agricola who described its recovery and refining in a treatise published in 1556. Man's use of petroleum, however, goes back into prehistory. Beginning more than 5,000 years ago, the Sumerians, the Assyrians, and the Babylonians each used the large seepages at Hit on the Euphrates, and similar seepages were known and used in various ways at many places in Mesopotamia and adjacent areas bordering the eastern Mediterranean. The Dead Sea was known in ancient times as Lake Asphaltites (from which is derived the term asphaltum) because of the semisolid petroleum washed up on its shores from underwater seepages.

Because there was a scarcity of natural stone and timber in the lower reaches of the Tigris and Euphrates rivers, the inhabitants of that region used bitumen, or asphalt, for many purposes. Excavations at Susa in Iran and Ur in Iraq reveal that bitumen was mixed with sand and fibrous material for use in construction of watercourses and levees of the extensive irrigation systems of the area and as a mortar for bricks. Bitumen was widely used in the caulking of ships and the building of roads. It was employed to fasten pieces of flint into bone in making sickles and was used in paints and waterproofing baskets and mats, as a cement for mosaic and inlaid work, and in the construction of tools of warfare. It had wide usage in magic and medicine, both as a liniment and a laxative, as well as in fumigation. Artistic and religious objects were carved from bitumen-impregnated sands, and the mining of rock asphalt was an important industry.

The ancient Greeks were familiar with many of the uses of bitumen, but they did not transmit this knowledge to the conquering Romans. The latter had developed excellent mortar from other sources for their building construction. The tars and pitches derived from the cedars of Lebanon, as well as from woods found in many other places in their empire, provided adequate substitutes for the mineral bitumens of the East. Many ancient writers described the natural occurrences of oil and gas seepages in the Middle East, especially in the Baku region of what is now the Soviet Union. Oil and gas from seepages, probably first ignited by lightning, provided the fuel for the "eternal fires" that formed the main centre of the fire-worshiping religion of the ancient Persians.

In China and Burma petroleum was encountered in drilling brine wells. The Chinese are reported to have drilled for salt to a depth of 450 feet (140 metres) about 200 BC and to have ignited and used the associated natural gas to evaporate the brine from these salt wells by the 4th century AD.

Use as an illuminant

Early in the Christian Era the Arabs and Persians developed an interest in crude petroleum and its distillation into illuminants, and, probably as a result of the Arab invasion of Spain, the industrial art of distillation was available in western Europe by the 12th century. The Spanish discovered seepage deposits in Cuba, eastern Mexico, Bolivia, Ecuador, and Peru; Sir Walter Raleigh found the famous pitch seepages of Trinidad; and the Dutch made similar discoveries in Java and Sumatra. All of these explorers used the pitch to caulk their ships. The first explorers of the Great Lakes region found oil seepages near Cuba, New York, and Oil Creek, Pennsylvania, where they reported that the Indians used it for medicinal purposes.

Until the beginning of the 19th century illumination in the United States was little improved over that known by the ancient Greeks and Romans. The increasing development of urban centres, with an accompanying need for better illumination and the growing demand for lubricants, proved only the beginning of new needs for additional sources of oil. Whales were becoming increasingly hard to find, and by the middle of the 19th century kerosine, or coal oil, derived by the distillation of coal, was common in both Europe and America. The rapidly approaching Industrial Revolution required a much cheaper and more convenient source of both lubricants and illuminating oil, and thus the stage was set for completion of the first well specifically drilled for oil. This well, drilled by E.L. Drake in western Pennsylvania and completed in August of 1859 at a depth of 69½ feet (21 metres), opened up the petroleum industry and the closely associated modern industrial age. Within a short time, cheap oil from underground reservoirs was being refined at already existing coal-oil refineries, and the spread of oil field discoveries before the 20th century covered 14 states from New York to California and from Wyoming to Texas. During the same period oil fields were discovered in Europe and the Far East.

World energy consumption. At the outset of the 20th century the Industrial Revolution, characterized by the advent of the automobile, had proceeded so far that the use of refined petroleum for illuminants ceased to be of primary importance, and the oil industry instead became the major supplier of energy.

Although petroleum is an important raw product in the petrochemical industry, its greatest use today is in the manufacture of energy upon which mankind's present and future economy depends. Figure 1 shows that the

By courtesy of Standard Oil of California

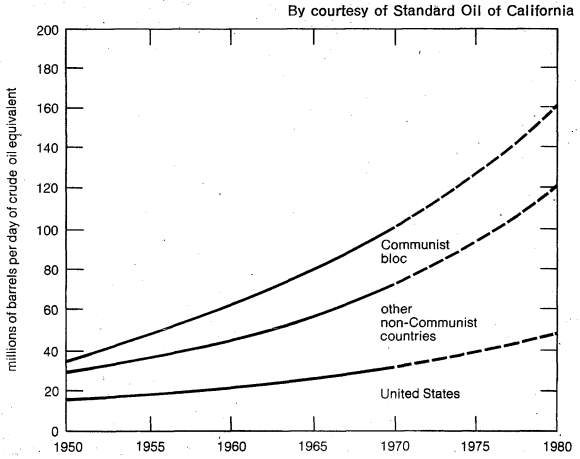


Figure 1: World energy consumption in crude oil equivalent. (Data for China unavailable.)

United States currently uses about 32 percent of the total world energy, the remaining non-Communist countries about 41 percent, and the Communist countries (with the exception of China, for which data are not available) about 27 percent. In 1970 the peoples of the world used the energy equivalent of about 100,000,000 barrels of oil per day, equalling the energy contained in a major or giant oil field, each day of the year. The United States is increasing its energy consumption by about 4 percent per year, other non-Communist countries at almost 6 percent annually, and the Communist countries (again excluding China) are averaging about 5 percent, equal to the worldwide average. Using the 5 percent figure, energy consumption by 1980 will rise to the equivalent of 162,000,000 barrels of oil per day.

Figure 2 shows the world energy consumed each day in terms of hydroelectric power, coal, oil, natural gas, and nuclear power; the energy equivalent is given as millions of barrels of oil. Oil has met the increased demands in world energy best. Use has increased from 11,000,000 barrels daily in 1950 to 46,000,000 barrels in 1970, with a projected use of 80,000,000 barrels per day by 1980. Oil and natural gas together will supply two-thirds of the world energy demand by 1980. Coal consumption shows a small steady increase, but it has failed to hold its relative position in the world energy market. Prior to 1940 coal supplied about 80 percent of Europe's energy demands; today its share is 28 percent. Although coal is still

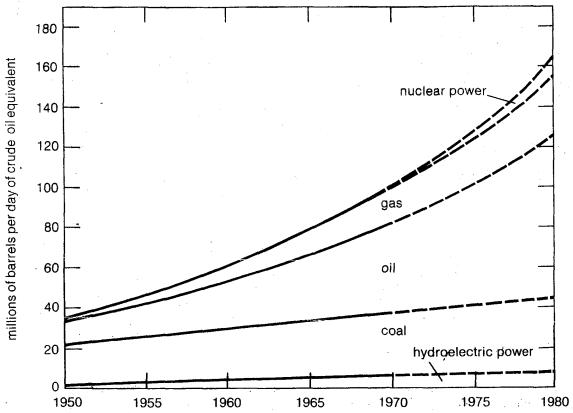


Figure 2: World energy consumption. (Data for China unavailable.)
By courtesy of Standard Oil of California

the predominant fuel for the generation of electricity, the modern demand for reduction of air pollution has forced many users to turn to low-sulfur oil and gas. Hydroelectric power can never become an important competitor in the world energy market; most of the possible generation dam sites are already power committed. Although the development of nuclear power generation is currently behind schedule, by 1980 it may supply about 9 percent of the world's energy.

The growth of consumption of crude oil is shown in Figure 3: the United States, Canada, western Europe, and

By courtesy of Standard Oil of California

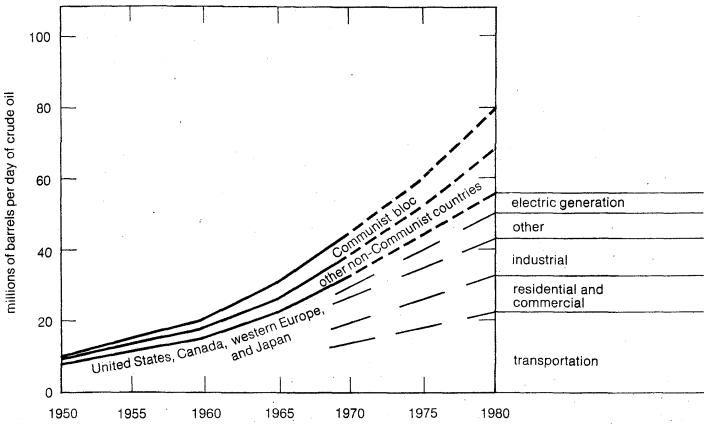


Figure 3: World oil consumption and use. (Data for China unavailable.)

Japan are considered in one group; other non-Communist countries are in another; and the Communist countries as a third. The "other non-Communist countries" category represents the comparatively small amounts of oil used in South and Central America, Africa, the Middle East, the Pacific Islands, and non-Communist Asia. In many of these countries the present amounts of energy uses are extremely limited, but the potential energy needs are enormous.

Projected consumption

The projected total consumption of almost 80,000,000 barrels of oil a day by 1980 represents an increase of 5.7 percent per year between 1970 and 1980; this is less than the growth rate for the period between 1960 and 1970. The tremendous increases in demand for energy throughout the world have caused a very critical supply situation. By the beginning of 1970 a total of about 225,000,000,000 barrels of crude oil had been produced by the petroleum industry since the drilling of the Drake well in 1859. From 1970 to 1980, however, the petroleum industry will be called upon to produce almost as much oil as it produced in the previous 110 years. In 1980 oil probably will furnish almost 50 percent of the world's energy, and natural gas will furnish approximately 18 percent.

This article treats the characteristics, origin, and accumulation of petroleum, and its world occurrence and distribution. Type examples of major oil fields are described, and the relation of proved petroleum reserves and probable future discoveries to ever-increasing energy demands is discussed. Further information on allied energy sources is presented in the articles COALS; NATURAL GAS; and OIL SHALES. See also SEDIMENTARY ROCKS; SEDIMENTARY FACIES; and SALT DOMES for additional information on petroleum environments and source rocks; PETROLEUM AND GAS EXTRACTION for methods of exploration and production; and ROCK DEFORMATION for a summary of the geological structures that are involved in many petroleum traps.

CHARACTERISTICS OF PETROLEUM

Physical properties. Petroleum consists of a closely related series of complex hydrocarbon compounds that range from a light gas, methane, to a series of heavy solids. The various mixtures that comprise liquid petroleum or crude oil are separated by distillation under increasing temperatures. The components of the mixture, classified from light to heavy, are the gases, gasolines, kerosine, gas oil, lubricating oil, residual fuel oil, and, finally, asphalt and paraffin.

Colour, consistency, and weight

Because crude oils vary greatly in their chemical compositions, their physical properties such as colour, gravity, and viscosity also vary widely. They are generally greasy or oily to the touch but may vary from a thin liquid with the consistency of gasoline to a thick, viscous black asphalt. In thin layers they range from colourless through all degrees of translucency to opaque, and in reflected light the colours vary from yellow, green, red, through brown, to black. Crude oil is immiscible with water and is lighter than the water; hence, it floats on associated water.

Specific gravity is the ratio of the weight of equal volumes of any material and pure water (considered to be 1). Because the volume of a liquid is affected by both temperature and pressure, the weight of oil is measured at standard conditions of 60° F (16° C) and one atmosphere of pressure. The widely used American Petroleum Institute (API) gravity scale is based on pure water with an arbitrarily assigned API gravity of 10°; oils of progressively lower density have higher values of API gravity. Examples of the relationship of the API gravity scale to the specific gravity of oils through their normal ranges are given below:

specific gravity at 60° F	API degrees
1.0000 (pure water)	10
0.9340	20
0.8762	30
0.8251	40
0.7796	50
0.7389	60

Crude oils that range within 5° of 30° API gravity are the most common and constitute the major portion of the world's production. The specific gravity of crude oil generally tends to decrease with depth of burial, with a corresponding rise in the API gravity. The same phenomenon also may be observed in individual fields with a number of separate reservoirs at varying depths, although there are notable exceptions, as in the great Burgan Field in Kuwait and in a number of fields in the Baku oil province of Russia.

The price of crude oil is ordinarily based on its gravity. The lighter oils command a higher price because they contain a larger amount of the more profitable and easily extracted gasoline hydrocarbons. The colourless or light-coloured crude oils have the highest API gravities, those of medium gravity are often green, and the heavier varieties vary through the darker colours to black.

Because petroleum is always at a temperature above the boiling point of some of its compounds (methane has a boiling point of −164° C [−263° F]), the more volatile constituents are constantly escaping into the atmosphere unless confined. On the other hand, the boiling points of some of the compounds are too high to be normally measured, and it is therefore impossible to determine or refer to a common boiling point for crude oil. For the same reasons, it is impossible to speak of a freezing point

for any particular petroleum, because the individual compounds solidify at different temperatures. The pour point, or the temperature below which crude oil becomes plastic and will not flow, is important and always is determined, however; it ranges from +90° to below −70° F (32° to −57° C).

Viscosity of petroleum is the measurement of the resistance to flow of a certain quantity of oil through a fixed opening at a given temperature and pressure; the greater the viscosity, the longer the time required for flow. The viscosity of petroleum is mainly determined by the amount of dissolved gas it contains and the temperature of the fluid; the more gas in solution and the higher the temperature, the lower the viscosity. The temperature of the earth rises with depth of burial at a rate that varies substantially from place to place; it ranges roughly between 0.5° to 2.2° F per 100 feet, averaging about 1.6° F (0.9° C) per 100 feet. As the oil viscosity decreases with a rise in temperature, the viscosity of the crude oil in the reservoir rock below the surface decreases with depth of burial.

As the depth of burial of the reservoir rock in which petroleum is present increases, so does the pressure exerted on the gas or liquid in the reservoir. This pressure, depending on the depth of burial, may vary from a few hundred pounds to many thousands of pounds per square inch. The amount of gas that can be dissolved in the oil therefore increases with depth of burial, and the volume of the solution (oil and dissolved gas) increases as well, until the oil is saturated with gas (the bubble point). In deep, high-pressure reservoirs, oil may contain 150 times its volume of surface-equivalent dissolved gas. As the oil is brought to the surface from the reservoir, the dissolved gas comes out of solution as the pressure is reduced, and the volume of the reservoir oil decreases to "stock-tank" oil. With the loss of the dissolved gas at the surface the crude oil becomes denser and more viscous.

Crude oil is measured in the United States in barrels of 42 gallons each; the weight per barrel of API 30° gravity oil would be about 306 pounds. In many countries it is measured by the metric ton; for the same gravity oil a metric ton is equal to about 252 imperial gallons, or approximately 7.2 U.S. barrels.

Chemical characteristics. Although petroleum consists basically of compounds of only two elements, carbon and hydrogen, these form a large variety of complex molecular structures. Regardless of physical or chemical variation, however, almost all crude oils will range between 82 and 87 percent carbon by weight and 12 to 15 percent hydrogen. The more viscous asphalts will vary between 80 and 85 percent carbon and 8 to 11 percent hydrogen.

Paraffin, naphthene, and the aromatic series

It is possible to group crude oils into three basic series, paraffin, naphthene, and aromatic. Most petroleums are mixtures of these three series in various and seemingly endless proportions, and no two crude oils from different sources are completely identical. The paraffin series of hydrocarbons, also called the methane (CH_4) series, form the most common hydrocarbons in both crude oils and natural gas. It is a saturated straight-chain series and has a general formula C_nH_{2n+2}, in which C is carbon, H is hydrogen, and n is an integer. The paraffin gases, starting with their most stable and common member—methane, or "swamp gas"—form most of the natural gases, whereas the members of the series that are liquid at normal temperatures but boil between 40° to 200° C (approximately 100° to 400° F) are the major constitutents of gasoline. The paraffin series tend toward the lower densities, or higher API gravities; the residues obtained by refining are paraffin waxes, both plastic and solid.

The naphthene series has the general formula C_nH_{2n} and is a saturated closed-ring series. It contributes an important part of all liquid refinery products but also forms most of the complex residues from the higher boiling-point ranges, and the series is generally heavier (with lower API gravity) for this reason. The residue of the refinery process is an asphalt, and crude oils in which this series predominates are called asphalt-base crudes.

The aromatic series has the general formula C_nH_{2n-6} and is an unsaturated closed-ring series. Its most common member, benzene (C_6H_6), is present in all petroleums, but the aromatics as a series generally comprise only a small percentage of most crude oils.

In addition to the practically infinite mixtures of hydrocarbon compounds that form crude oil, sulfur, nitrogen, and oxygen are generally present in small but often important quantities. Sulfur, which may occur as free sulfur, hydrogen sulfide gas, or organic sulfur compounds, usually varies between 0.1 and 5 percent by weight, but crude oils with less than 0.5 percent sulfur content are rated as "low sulfur," a highly desirable quality in view of the recent reforms in air quality standards. It is now a common practice to remove excess sulfur during refining. The sulfur compounds are heavy, so that the heavier crude oils tend to be higher in sulfur content. Oxygen, generally less than 2 percent by weight, may be present either as free oxygen or, more commonly, as part of heavier hydrocarbon compounds. For this reason, the heavier oils contain the most oxygen. Nitrogen, generally less than 0.1 percent by weight, is present in practically all crude oils, but it may be there as a common part of the dissolved natural gas.

Sodium chloride is also present in most crude oils, and when it exceeds 0.7 percent by weight, it generally is removed, like sulfur. Many metallic elements are found in the ash residue after complete combustion, including most of the elements also found in seawater. They probably are present because of the intimate association between seawater and the organic forms from which petroleum is generated. Among the most common metallic elements in petroleum are vanadium and nickel. These apparently occur in organic combinations in petroleum, just as they do in living organic plants and animals. It has been observed that the vanadium–porphyrin complexes are associated with high-sulfur oils and the nickel–porphyrin complexes with the low-sulfur crudes. Uranium is also found in petroleum, but its relationship to crude oil is still poorly understood.

Metallic elements in petroleum

Crude oil also can contain a small amount of various decay-resistant organic remains, such as siliceous skeletal fragments, wood, spores, resins, coal and lignite, and many other remnants of former life.

ORIGIN OF PETROLEUM

In spite of the common occurrence of both the liquid and gaseous forms of petroleum, and the great amount of scientific research on petroleum that has been carried out by many workers, there remain many unresolved questions regarding its origin. Although it is recognized that the original source of carbon and hydrogen was in the materials that made up the primordial earth, it is generally accepted that these two elements have to pass through an organic phase to be combined into the varying complex molecules recognized as petroleum. There are numerous geochemical and geological reasons for this belief, a few of which may be summarized here:

1. Petroleum commonly is associated with sedimentary rocks, principally those deposited under marine conditions but also including continental sediments; conversely, there is a complete absence of commercial deposits of petroleum where only igneous or metamorphic rocks (formed under great heat and pressure) are present.
2. The optical activity of petroleum (the ability to rotate the plane of polarized light) is almost completely confined to compounds of biogenic origin.
3. Most types of petroleum contain complex hydrocarbon compounds termed porphyrins, formed either from the green colouring matter of plants (chlorophyll) or from the red colouring matter of blood (hemin).
4. Carbon isotope ratios ($^{13}C/^{12}C$) indicate that petroleum may be derived in large part from the lipid (fats and waxes) fractions of organisms.
5. Many petroleum-like hydrocarbons have been found in recent marine sediments as well as in soils in many places throughout the world; these occurrences form a link between present living organisms and the petroleum found in sediments of older geological ages.

Organic source materials. The organic material that is the source of most petroleum has probably been derived

Derivation from plants and animals

from the single-celled planktonic (free-floating) plants, such as diatoms and blue-green algae, and single-celled planktonic animals, such as foraminifera, that live in the aquatic environments of marine water, brackish water, or fresh water. These simple forms were abundant in the seas long before the beginning of the Paleozoic Era, 570,000,000 years ago, and could have formed the source organisms of the petroleum found in the Precambrian and Early Paleozoic rocks; they also may have contributed to much of the petroleum found in younger rocks. In addition, land plants brought into the lakes and seas by rivers apparently have been the source of some crude oils. The larger, more complex forms of sea life—such as corals, mollusks, crustaceans, and shellfish—are neither abundant enough nor are their remains adequately preserved from the sea scavengers to constitute a source for crude oil.

All organic matter can be divided into the classes of proteins (amino acids), carbohydrates (sugars, cellulose), lignin, pigments (including porphyrins), and lipids (fats, fatty acids). All but lignin are present in both living plants and animals; the source material for petroleum is among these five building blocks of living organisms.

The proteins and their amino acids are relatively easily decomposed and probably contribute little to the petroleum source material. The carbohydrates make up the major portion of both plant and animal matter. They are subject to rapid and nearly total degradation but may provide a logical source material for some petroleum. The lignins must definitely be considered as a major contributor to organic deposits of land plants, and although these have contributed largely to lignite and coal deposits, many authorities believe that they may constitute a progenitor of petroleum hydrocarbons as well. The pigments, particularly the porphyrins, are known to form a minor but recognizable source of crude oil.

It is the lipids, however, that may form the chief and primary source of petroleum. Biochemically the lipids (fats and fatty acids) are insoluble in water but are soluble in ether, benzene, or chloroform. Some of the planktonic plants (phytoplanktons) produce and store fatty oils during photosynthesis. Moreover the $^{13}C/^{12}C$ ratios in petroleum closely resemble the ratios found in the lipid fractions of the various phytoplanktons. Because the lipid fraction, which contains the hydrocarbons most closely resembling petroleum, is much more stable than the water-soluble proteins and carbohydrates, it could well be the main building block from which petroleum is constructed.

Living organisms, however, do not generate or synthesize petroleum during their life; further evolution through diagenesis (physical and chemical changes after deposition) of the organic matter over some unknown length of time is required before true petroleum is created.

Importance of rapid burial

Creation of source beds. The preservation of organic materials after the death of their contributors is best obtained by rapid burial within fine-grained, clay-sized sediments along sea margins, particularly in the vicinity of river deltas. These conditions effectively preserve the protopetroleum hydrocarbons for later diagenesis into true petroleum. The same requirements for aquatic but nonmarine environments, as in terrestrial lakes, have also been adequately met in many places in the world in the past, as in the ancient inland-lake deposits of the Uinta Basin of Utah, the Vienna Basin of Europe, and in parts of the eastern and western Venezuela basins.

With the rapid burial of petroleum source material in fine-grained sediments and subsequent shielding of the material from the bottom scavengers of the sea floor, its preservation from immediate and total decay can only occur where anaerobic bacteria are present. This type of bacteria allows the preservation of the material under an oxygen-free (reducing) environment, in contrast to the aerobic bacteria, which require the presence of oxygen to live and which are responsible for complete destruction of organic matter. It is, however, quite doubtful that bacteria can produce petroleum, with the exception of the gas methane, and there is no positive evidence of bacterial action in sediments below depths of a few hundred feet. It appears that the main result of the bacteria is to help create the reducing environment that protects the protopetroleum already developed and allows for the further formation of true petroleum.

Throughout the past, just as today, there have been portions of the seas, as well as areas behind reefs or ridges on the sea floor, where the circulation of waters is essentially restricted and where reducing conditions may prevail. Where the reducing conditions most favourable to the preservation of the organic material are accompanied by the rapid accumulation of fine-grained sediments in an area of rapid subsidence, so that there is a buildup of a substantial thickness of source beds, the most likely conditions prerequisite to the formation of petroleum are encountered.

In addition to the deposition of fine-grained clay-size sediments that are necessary to bury and preserve the organic matter, coarser sands or carbonates also must be present to act as porous carriers or reservoir beds. If porous carrier beds are lacking in a thick shale sequence, any petroleum that may be generated will remain in the source beds and fail to accumulate in commercial quantities. The close association of good reservoir rocks in carbonate reef environments and the nearby source material areas provides the explanation of the importance of reefs in petroleum accumulation.

Evaporites and petroleum

The very common association of evaporite deposits of salt and anhydrite with a sequence of beds that contain oil and gas is probably related to the fact that both evaporites (*q.v.*) and petroleum tend to develop in closed basins, termed silled or barred, when a land or submarine barrier at least partially separates the deeper part of the basin from the open seas. The silled or barred basin is considered by most geologists to be a prerequisite for the deposition of evaporites, but it also forms an excellent sedimentary environment for the accumulation and preservation of the plankton that are concentrated near the surface of the waters and float or are carried across the barrier during life, to be preserved by an anaerobic bottom environment after death.

The total thickness of sediments deposited in individual areas may vary from only a few feet to well over 50,000 feet (15,000 metres), and petroleum occurs today in commercial quantities at depths varying from less than 100 feet (30 metres) to more than 25,000 feet (7,500 metres). Burial thus can create substantial changes in the environment of the original source material. The weight of the overlying sediments, or overburden, increases about one pound for each foot of depth of burial; this results primarily in the continued expulsion of interstitial water contained within the muds and clays and the continued compaction of those sediments with additional burial. The very fact, however, that crude oil of all kinds is found at all depths from just below the surface to below 25,000 feet strongly indicates that direct pressure has no particular effect on the creation of petroleum.

Temperatures of formation

There also are good reasons for believing that most crude oils have never been exposed to temperatures much higher than 212° F (100° C), although in isolated cases oil is commercially produced at temperatures up to 340° F (170° C). Earth temperatures of these higher magnitudes are attainable within reach of the drill only in areas of very high temperature gradients, greatly in excess of those normally encountered. There is no reason to believe that earth temperatures in most of the major oil-producing areas of the world have ever reached anything near the extremes that would be required for conversion of organic material into petroleum, temperatures in excess of 662° F (350° C). It has been suggested that, given the lengths of geologic time available, the formation of petroleum might be accomplished at much lower temperatures than those required in the laboratory to transform the protopetroleum of the recent sea-floor sediments or of kerogen shales into petroleum, although many workers do not accept this premise. It has also been suggested that the presence of catalysts such as acid silicates found in certain clays, some of which are used as catalysts in modern refining, may facilitate the creation of petroleum during geologic time. Opponents of this theory, however,

point to the presence of unconverted oil shales in the oldest rocks as indicating that low temperatures, even given geologic time, cannot create petroleum.

Association of natural gas. Inasmuch as all accumulations of liquid petroleum have gas associated with them, it is assumed by a number of workers that the origin of gas is the same as the origin of oil. A common exception to this assumption is made for methane, which is a product of bacterial decomposition of vegetable matter. Many individual accumulations of gas have no oil associated with them, however, and it may well be that the origin of some gas is different than the origin for the liquid hydrocarbons. For instance, in the great Groningen gas field of The Netherlands, with over 65,000,000,000,000 cubic feet of reserves in Permian rocks (225,000,000 to 280,000,000 years in age), the gas is attributed to the carbonization of the underlying Carboniferous (280,000,000 to 345,000,000 years in age) coal because of the increased temperatures resulting from deep burial. It has also been noted in some areas that hydrocarbon density decreases with increasing depth. At temperatures greater than 350° F (180° C) liquid hydrocarbons probably cannot exist. Depending on the thermal gradient, in some areas liquid hydrocarbons would not exist at depths presently reached by the drill but would be converted into dry gas. In other instances the presence of major gas reserves at shallower depths may be the result of conversion of hydrocarbon liquids to a gaseous phase upon being subjected to temperatures above 350° F, followed by migration of the more freely moving gas into shallower depths under conditions that still allowed it to maintain its gaseous composition.

MIGRATION AND ACCUMULATION

In addition to accepting the organic origin of petroleum, geologists in general accept in one form or another the following history of migration and accumulation of petroleum:

1. The organic matter undergoing diagenesis moves out of the compacting clays under the influence of the expelled water and into more porous and permeable reservoir rocks, such as sandstones or porous carbonates. This initial movement is called primary migration, and, although it certainly occurs, in part, shortly after burial, it may also continue over geologic time.
2. During the period of primary migration or during the subsequent period of retention in the reservoir rock, the change to petroleum occurs. This is followed by a period of collection into a continuous phase, in turn followed by a period of secondary migration, again under the influence of water, through a carrier bed or reservoir rock to a place where further movement is impossible (a trap) and where the petroleum collects into pools or fields.

Character of reservoir rocks. Crude oil in commercial quantities is generally found in reservoirs provided by the pore space in sedimentary rocks; less than 1 percent of the world's oil has been found in fractures in igneous or metamorphic rocks, about 59 percent has been found in the pores between the mineral grains of sandstones, and about 40 percent in the void space present in dolomites or limestones (carbonates).

Porosity and permeability

The two most important characteristics of a reservoir rock are its porosity and its permeability. Porosity is defined as the ratio of the volume of pore space to the total bulk volume of the material expressed in percent. Permeability is the capacity of the rock to transmit fluids through the interconnected pore spaces of a rock; the customary unit of measurement is the millidarcy. Although there often is an apparent close relationship between porosity and permeability, because a highly porous rock may be highly permeable, there is no real relationship between the two; a rock with a high percentage of porosity may be very impermeable because of lack of communication between the individual pores or because of capillary size of the pore spaces.

Reservoir rocks may be divided into two main types: those in which the porosity is primary, or inherent, and existed at the time the rock was first deposited; and secondary, or induced, developed subsequent to deposition by solution, recrystallization, weathering, fracturing, or some other process imposed on the original rock.

The porosity of sandstone reservoirs is mainly primary and varies from 10 to 40 percent, depending for the most part on the original sorting of the sand grains and the amount of cementation that has occurred since burial. Permeabilities in average oil pools generally range from five millidarcies to more than 1,000 millidarcies, and any average greater than 100 millidarcies is considered as good. Figure 4 presents examples of primary porosity in

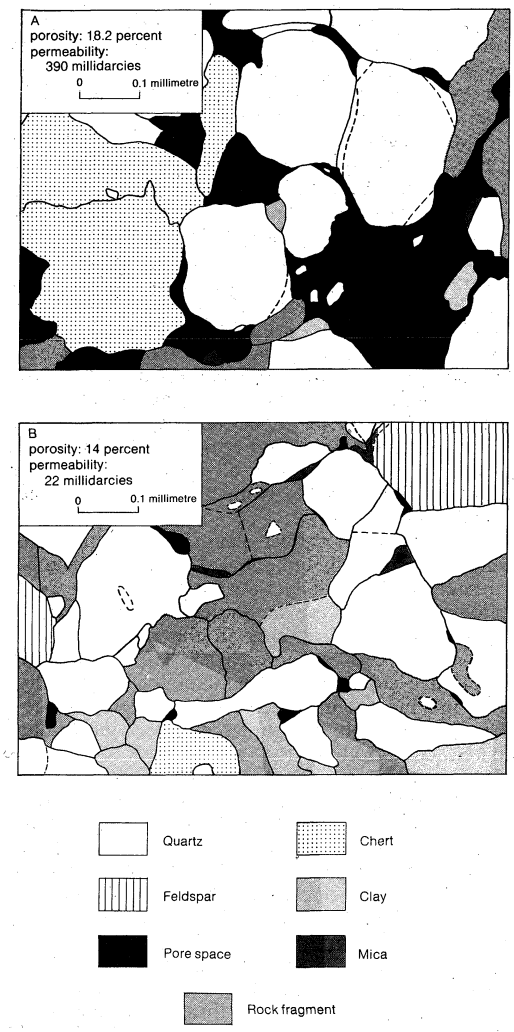


Figure 4: Sections of graywacke reservoir rocks in Alaska, showing (A) excellent reservoir rock and (B) inadequate rock.

a sandstone of an excellent reservoir rock and a poor reservoir rock. The pore spaces are shown in black. Some carbonate sediments may also possess very good primary porosity if the rock is a coquina (consisting principally of shells) or a reef or fragmental limestone.

Secondary, or induced, porosity may develop in a large number of ways after the original deposition of the sediments and is generally more common to carbonate reservoirs than to sandstones. The process of dolomitization of a calcium carbonate rock, a process that involves the substitution of magnesium for calcium during recrystallization, can create a porous and permeable reservoir of major importance in an otherwise dense limestone. Fracturing of carbonate rocks also is of significance. Fracture planes may combine with the initial rock porosity to provide excellent avenues of flow.

The development of secondary porosity by solution, generally involving surface waters that have become acid through the decay of organic material and the action of soil bacteria at the surface, has created many reservoirs,

both in limestones and dolomites, as the descending percolating waters follow existing pores, joints, and bedding planes through the rock.

Reservoir pressures

The connected pores in all reservoir rocks below the water table are filled with water that may be fresh or saline, except in the relatively small areas occupied by petroleum accumulations. This reservoir water generally contains dissolved hydrocarbon gases in measurable amounts. The water within the reservoir rock exerts a hydrostatic pressure equal to the weight of a column of overlying salt water, which normally is about 0.465 pound per square inch per foot of burial depth. Reservoir pressures may vary from much smaller values to almost that of the weight of the overburden, about one pound per square inch per foot of depth, however.

Primary and secondary migration. Freshly deposited muds have water-filled pore spaces equal to as much as 80 percent of their volume; the addition of only a few hundred feet of overburden will reduce this porosity to a little more than 30 percent, however, and by the time 10,000 feet of overburden is reached the porosity of the resulting shale may be less than 10 percent. The water retained in the muds immediately following deposition is in two forms; it exists in the pore space between the grains and also as an adsorbed film of water molecules around each clay particle. As the weight of the overburden increases, the interstitial water is expelled first, followed by the adsorbed water around the individual particles. The large amounts of water that are expelled provide an escape mechanism for hydrocarbons, which move toward the coarser-grained, more porous sediments in any direction—upward, laterally, or downward—because the pressure gradient is away from the compacting, fine-grained sediments.

The actual form taken by petroleum in its escape with the expelled water is still unknown. It has been held by various researchers that it is transported in water solution, colloidal solution, or as molecular films and is transformed into crude oil only in the pore spaces of supercapillary size in the reservoir rock. There is a strong possibility that a substantial amount of the original hydrocarbons in source beds never does undergo migration into carrier or reservoir beds but remains in its original disseminated condition in the source rock.

The movement of petroleum through reservoir rocks depends first upon its gathering into a continuous phase; the isolated globules must coalesce and form larger masses of oil on which the forces involving buoyancy difference between oil and water can act. Because of the increase in the amount of the gas that may be retained in solution by oil with increased depth and the constant rise in temperature with depth, both of which decrease its viscosity, oil during its migration can possess a mobility greatly exceeding its ability to move at the surface. The many active seeps of petroleum throughout the world attest to the fact that crude oil can migrate vertically; it moves upward through joints and fractures always in the direction of lesser reservoir pressure.

Direction of petroleum migration

In horizontal migration, the propulsive forces that cause the petroleum fluids to move through the water-saturated reservoir rocks are buoyancy, expansion of gas in the oil, and hydrodynamic forces exerted by the reservoir water. Any one of these three mechanisms or combinations operating in concert can move the oil during geological time intervals until it is entrapped or until it reaches the surface. In the case of entrapment any later change of the environment that forms the trap may allow the mobile petroleum to resume its migration. It should be emphasized that oil and gas may move through the reservoir rock in the same direction the water in the reservoir is moving or in a direction opposed to the movement of the water. A high rate of water movement downward toward a lower outlet may actually act as a seal, stagnating and offsetting the upward migration of oil and gas and allowing the accumulation of the petroleum in commercial quantities in areas where no true trap exists. There are a number of instances of oil accumulation where this concept of hydrodynamic entrapment best explains the local situation. It appears probable that a combination of the buoyancy and hydrodynamic forces involved in secondary migration best resolves the complex problem.

Formation of oil and gas pools and fields. The commercial accumulation of oil in a single reservoir is termed a pool. It is possible to have a series of pools in a common reservoir rock, each separated either by faulting or by zones of impermeable intervening rock, as well as a series of pools in a number of different reservoir rocks, all located on the same structure. In the latter case of a series of pools, as well as in the case of a single pool confined to a single reservoir, the area encompassed by the individual structure containing the oil pools is known as an oil field.

A group of oil fields having related origins and a more or less common geologic history constitute a petroleum province. The discovery of an oil province occurs with the initial discovery and development of the first oil pool in an area of regional size.

Petroleum accumulation has been taking place since the earliest part of the Paleozoic Era, about 500,000,000 to 600,000,000 years ago, and it is continuing today. Simply stated, accumulation is the collection of petroleum hydrocarbons that have been trapped and inhibited from further movement. Because both oil and gas have great mobility, however, this period of arrested movement may endure for a geologic period of any duration but can be interrupted any time the bounding forces that create the oil pool change. The change may be artificial, created by man by drilling wells, or natural, when folding or faulting changes the structure of the subsurface.

At one time many geologists believed that the source beds and the reservoir rock were the same, but this theory of accumulation *in situ*, except for very few and minor exceptions, has now generally been discarded. One of the reasons is that, in many oil fields throughout the world, the accumulation into structural traps has taken place many millions of years after the deposition of both the source and reservoir beds, a fact determined by dating the structural deformation that created the traps in the reservoir beds. In other cases the reservoir rock, either nonmarine sandstone or reef limestone, was deposited under conditions inimical to source material preservation; the oil must have accumulated in these reservoirs from areas of substantially different environment. The presence of commercial accumulation of oil in fractured igneous and metamorphic rocks is similarly excellent evidence of migration of oil from its place of origin to a favourable reservoir environment.

Role and types of traps

The accumulation of oil and gas in reservoir rocks is the final step in the formation of an oil pool. The movement of the petroleum into the trap has been along an interface between porous and permeable carrier rock and an overlying impermeable roof; the roof rock, therefore, must be concave at its apex as viewed from below in order to form a trap. When the reservoir water associated with the accumulated oil and gas is under hydrostatic conditions, the oil (and gas, if present) comes to rest in the highest portion of the enclosure, and the contact between the oil and water will be horizontal. A trap, therefore, is an area of low potential energy into which the oil has moved from an area of higher energy. Where there are hydrodynamic gradients the entrapped oil may be moved from the highest point on the structure toward the flank, displaced in the direction of water movement. In these cases, the oil–water contact will be tilted, again in the direction of the water movement. However, the contact between oil and free gas, where present, will be horizontal regardless of whether the water is or is not moving, because the gas–oil contact is determined by the difference in the density of the two materials, neither one of which at their interface is under the influence of moving water. The oil and gas accumulated in the trap will rapidly seek equilibrium determined by the surrounding physical conditions, but any change in those conditions, such as folding, regional tilting, fracturing of the cover rock, or the drilling of oil wells, may result in further movement of the entrapped petroleum to adjust to the areas of lower potential energy.

Structural traps. Structural traps caused by folding occur in an infinite variety of sizes and shapes (no two folds

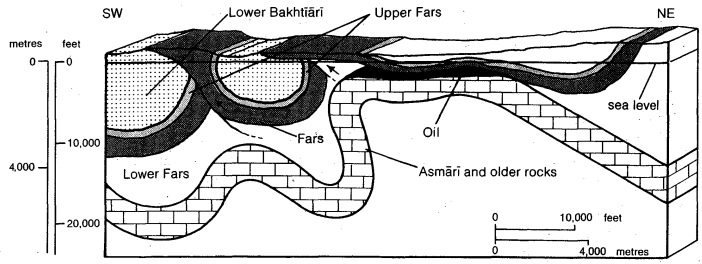


Figure 5: Section through the Masjed Soleymān oil field in Iran.

From *Geology of Petroleum*, 2nd ed., by A.I. Levorsen; W.H. Freeman and Company; copyright © 1967

are ever exactly alike). They may vary from essentially circular in plan view to long narrow folds or anticlines, with the flanks varying from symmetrical through asymmetrical to overturned. Folding can involve a number of reservoir rocks of the same or different types, and oil and gas accumulations frequently occur in a number of these separate but superimposed reservoirs. The folding may be due to compression or tension in the earth's crust, or the structure may result from the initial dip of beds deposited over and around buried hills, or from localized uplifts caused by intrusion of salt or shale plugs (see further SALT DOMES) or igneous intrusions. Many folds may be modified to various degrees by faulting, which may also affect the areas of accumulation of oil or gas. When rocks are folded, the volume of the rock itself is increased by jointing and fracturing; areas of lower potential energy result, toward which oil or gas will move. A good example of this is shown by the cross section through the Masjed Soleymān oil field, Iran (Figure 5). The Asmari limestone, a strong, competent rock unit in which the oil occurs, has been folded and the upper part of the fold contains the oil in a major fracture system.

Effects of folding and faulting

Faulting may form traps in which the fault or faults form one or more of the critical boundaries of the oil reservoir. Because faults may penetrate more than one reservoir rock, there can be a series of petroleum accumulations at various depths, with a single fault creating similar boundaries on different reservoirs. Where faults are curved or cut across low arches or folds, substantial quantities of oil may be entrapped. In certain areas such as the Gulf Coast of Texas and Louisiana, the Niger Delta, the Ganges-Brahmaputra, and the Vienna Basin, much of the seaward thickening takes place across faults that continued to grow during the deposition of the stratigraphic section, and this type of faulting is particularly effective in forming seals in these areas. Figure 6 shows

From *Geology of Petroleum*, 2nd ed., by A.I. Levorsen. W.H. Freeman and Company. Copyright © 1967

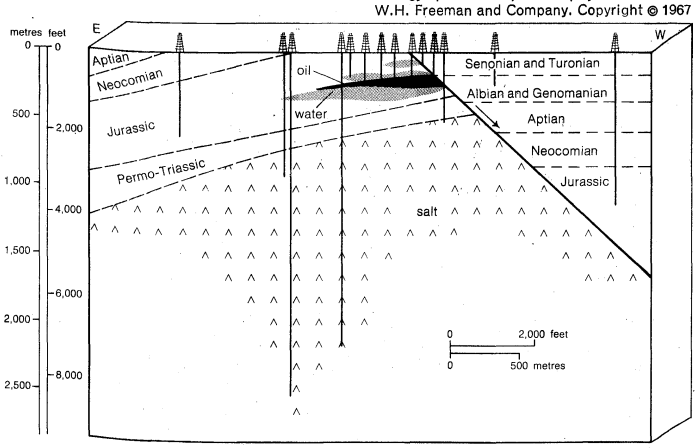


Figure 6: Dossor oil pool in the Emba Province of the Soviet Union. Oil is trapped in a faulted sandstone above the salt mass.

a cross section through the Dossor oil pool in the Emba Province of the U.S.S.R. Production comes from Jurassic sands (136,000,000 to 190,000,000 years in age) that butt against a west-dipping fault plane and younger impermeable beds on the downthrown side of the fault.

Stratigraphic traps. A stratigraphic trap results from a change of depositional environment during the formation of the reservoir or from a change or variation in the reservoir rock that occurred after deposition and diagenesis. The material that forms a porous and permeable reservoir rock changes laterally into impermeable and nonporous formation, so that the effective porosity of the reservoir is limited in areal extent. A lateral change from a porous sandstone into an impermeable shale may result in the creation of a reservoir covering a substantial area. Where a reef limestone grades into an impervious limestone or a shale section or where permeable dolomite changes to an impervious limestone, the stage is set for the development of oil pools. The fact that sand deposition commonly grades both basinward and shoreward into shales means that in many areas there is a concentration of oil pools along the limits of sand deposition, as in the eastern Lake Maracaibo area of Venezuela, the Maikop region along the northwestern Caucasus Mountains in the Soviet Union, and in the Gulf Coast region of Texas and Louisiana.

Shoestring-sand traps (literally, strings of sand that pinch out or lens out) may represent old offshore sandbars or fossil channel-filled stream beds. In either case the porous and permeable sand deposits form long, narrow lenses that may be filled with oil. This type of deposit is very important in many areas, such as eastern Kansas, southern Michigan, Pennsylvania, Oklahoma, and the Maikop oil region of the Soviet Union. A map of the Sallyards shoestring-sand trend in eastern Kansas is shown in Figure 7. On the right, the sand trend is super-

Adapted from N.W. Bass, "Origin of Bartlesville Shoestring Sands, Greenwood and Butler Counties, Kansas," *The American Association of Petroleum Geologists Bulletin*, vol. 18 (1934); American Association of Petroleum Geologists, Tulsa

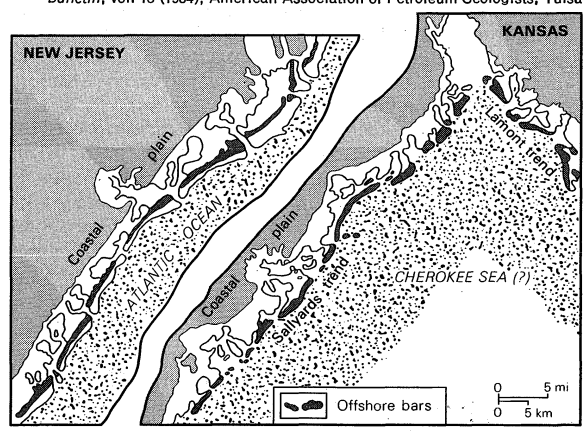


Figure 7: Sallyards shoestring-sand trend in eastern Kansas, superimposed on a hypothetical Cherokee Sea and compared with a portion of the present New Jersey coast.

imposed on a hypothetical map of the Pennsylvania Sea, 300,000,000 years ago, and, on the left, it can be compared with the offshore sandbars of the modern New Jersey coast.

Many great oil pools of the world occur in organic reefs. These reefs are present in all geologic ages from the Cambrian (500,000,000 to 570,000,000 years ago) to the Recent (the last 10,000 years, approximately), and are best referred to as organic reefs rather than coral reefs because corals are not the predominant life-form in many of them. All types of modern reefs are represented in ancient sediments. An ancient barrier reef very similar to present barrier reefs crops out in the Permian Basin of western Texas in the spectacular El Capitan Mountain. This reef is known to extend for at least 400 miles (640 kilometres) and has a thickness of more than 1200 feet (370 metres), with many oil and gas fields found along its buried portion. Most reefs of this type contain backreef, or lagoonal, sediments consisting of limestones, dolomites, shales, and evaporites, whereas the forereef, or basinward side, consists of bedded limestones or shales, which grade into the reef rock through a mixture of reef debris that has fallen from the main reef and accumulated as a talus slope.

Organic reef association

Oil fields associated with organic reefs are common in western Texas and New Mexico in Mississippian and Pennsylvanian rocks (about 280,000,000 to 345,000,000 years in age) and Permian rocks (225,000,000 to 280,000,000 years in age), the great reefs of western Canada in the Upper Devonian (about 350,000,000 years in age), the Middle Cretaceous (about 100,000,000 years in age) reefs of eastern Mexico, and in the Perm Basin reefs west of the Ural Mountains in the U.S.S.R.

An unconformity is a break in the stratigraphic sequence of deposition where the missing beds have been removed by erosion or were never deposited. If an unconformity represents a period of erosion, it is possible that considerable porosity has been developed by solution effected by percolating groundwater. Inasmuch as the reservoir rock may underlie or overlie the unconformity there are a number of ways as well as times in which the oil may migrate into the reservoir rock. In many cases there is good evidence that the source rock that contributed the oil lies above the unconformity and that the oil migrated downward to the porous reservoir rock, crossing the unconformity in its search for a porous and permeable site. The largest petroleum accumulation so far discovered in the contiguous 48 United States, the East Texas pool, is a stratigraphic trap associated with an unconformity where the Woodbine (Cretaceous) producing sand is truncated by a major unconformity and sealed by an impermeable cap rock. This field contained over 6,000,000,000 barrels of recoverable oil.

Other traps. A structural trap may be combined with a stratigraphic trap in any variety of ways, all of which find examples somewhere throughout oil and gas regions. Probably the most common combination is the updip pinch out of porosity, such as a change from porous dolomite to dense limestone or from sandstone to sandy shale on the flanks or across a portion of a subsurface structure, so that only a part of the structure contains commercial accumulation.

Salt domes and diapirs

The intrusion of deep-seated salt masses into overlying rocks may create a great variety of traps both in the intruded sediments and in those draped over the intrusion. Salt domes (*q.v.*) and structures associated with diapiric (piercement) salt masses are known from many places throughout the world. They have been studied most thoroughly in the Hannover Basin of Germany and the Gulf Coast of Louisiana and Texas, where important oil and gas fields occur.

The salt plugs of the Texas and Louisiana coast have come up through sediments from depths of more than 40,000 feet (12,000 metres). There is evidence that they had been moving through most of the Cretaceous (65,000,000 to 136,000,000 years ago) and all of the Tertiary (2,500,000 to 65,000,000 years ago) and that movement of some of them near the coast and under the waters of the Gulf is going on today. The Reitbrook salt dome, Germany, illustrates the prevalent faulting that is associated with salt masses when they breach the overlying sediments (Figure 8).

From A. Levorsen, *Geology of Petroleum* (1956); W.H. Freeman and Co.

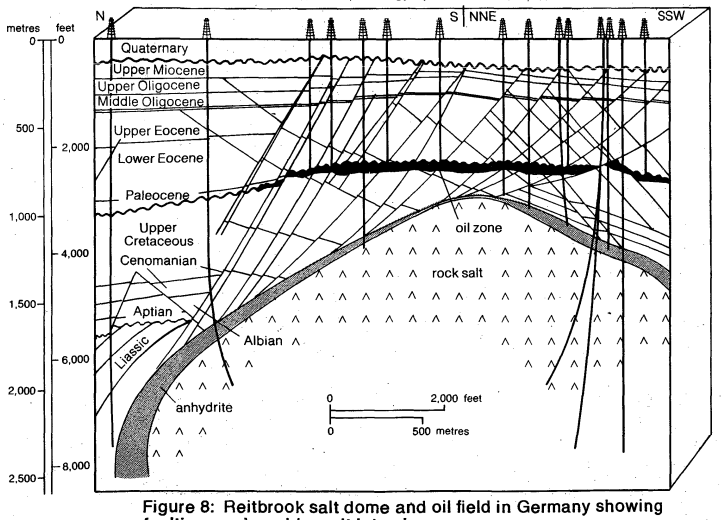


Figure 8: Reitbrook salt dome and oil field in Germany showing faulting produced by salt intrusion.

A hydrodynamic trap may occur where water is in movement through the reservoir rock and where the upward oil and gas migration is stabilized against the movement of the water. Under these conditions, it is possible to find oil and gas accumulations on structures where nothing but movement of the water would inhibit further migration of the oil and gas.

PETROLEUM OCCURRENCE AND DISTRIBUTION

Surface occurrences. Semisolid and solid occurrences of essentially discrete petroleum at the earth's surface are relatively rare, although veins and dikes of asphaltic or paraffinic petroleum residues are known from a number of widely scattered places, such as the Uinta Basin in Utah; South Africa; Trinidad; the Arctic Islands of Canada; Peru; Argentina; and many places in the Middle East. All of these occurrences result from the plastic nature of the incompetent heavy bitumen exuded under the pressure of the overburden along open joints and cracks in the surface rocks. There are also a number of known occurrences of discrete petroleum within small sealed cavities such as those within concretions and geodes, fossil casts, and the natural chambers within corals and ammonites, although the surrounding host rock may be completely barren of hydrocarbons. Liquid petroleum has been found filling small inclusions in both quartz and salt crystals. The origin of the oil found in these unusual occurrences constitutes an unresolved problem.

Lakes, tar pits, and mud volcanoes

Also at the earth's surface are seepages and lakes of bitumen, to a lesser or a greater degree thickened by evaporation. Some of these deposits consist of almost pure heavy oil, but many have substantial amounts of intermixed sand, clay, and other extraneous material. The heavy petroleum in these seepages has risen to the surface from underlying beds along fractures, joints, or faults connected with shallow subsurface folds, fault closures, or diapiric structures that contain residual oil pools. These petroleum lakes or surface seepages may vary in size from only a few yards in diameter through areas covering acres or even square miles, with depths of more than 100 feet. Among the most famous occurrences of this type are the Pitch Lake of Trinidad, the Bermudez pitch lake of eastern Venezuela, the Rancho La Brea Tar Pits of Los Angeles and other seepages in California, the seepages at Hit in Iraq and elsewhere throughout Iran and Saudi Arabia, and in the Baku area on the western shore of the Caspian Sea in the Soviet Union. In the early days of petroleum exploration, many oil fields were found by drilling in the vicinity of oil and gas seepages, although most of today's oil fields have no surface expression of oil or gas leakage nor are oil seepages necessarily related to commercial deposits of petroleum. In addition, many examples of active marine seepages are known throughout the world, among which are those near Santa Barbara, California; Trinidad; the Caspian Sea; and the Louisiana, Texas, and Yucatan portions of the Gulf of Mexico.

In a relatively few areas of the world, mud volcanoes are associated with subsurface accumulations of oil and gas. These surface expressions have no relationship to true volcanism but are mudflows, at times accompanied by oil and gas, that originate from shallow buried beds of unconsolidated clay and shale. The local points of eruption are generally along the crests of buried anticlines, other folds, or fault structures, and the amount of ejected material may be sufficient to build up a group of cones covering several square miles to heights of more than 1,000 feet. The gas escaping from the apex of the subsurface structure no doubt aids the diapiric upwelling of the clay. Commercial oil fields associated with both ancient or active mud volcanoes are found in the Baku region of the Caspian Sea, the island of Trinidad, and Burma, among other places.

Oil sands and tar sands

Sediments containing petroleum in the form of asphalts or thick heavy oil are generally called bituminous sands or bituminous limestones. These occurrences represent

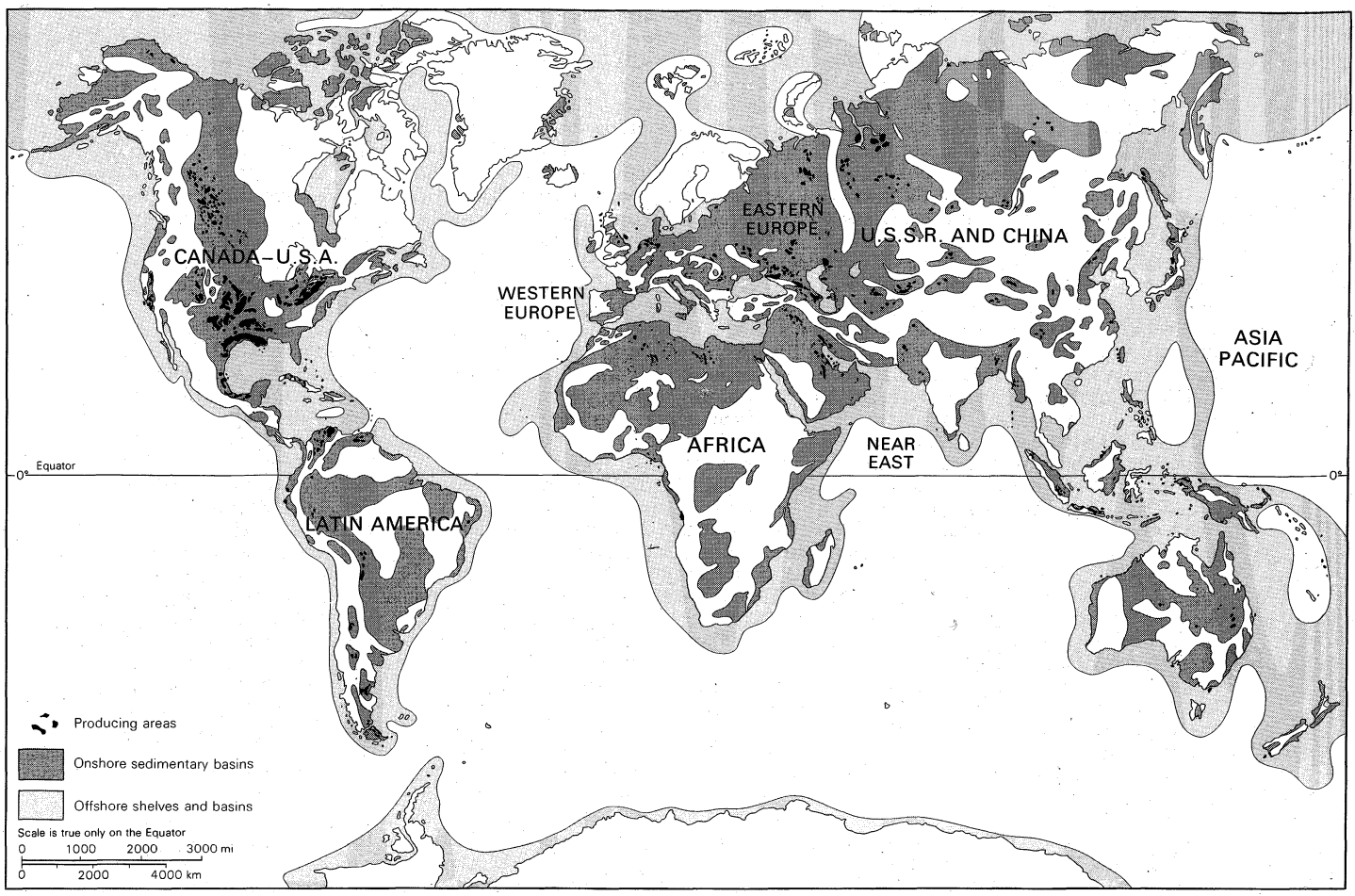


Figure 9: Offshore and onshore sedimentary basins and petroleum-producing areas of the world.
Adapted from "Worldwide offshore activity and sedimentary basins," *Oil and Gas Journal* (November 1970); Petroleum Publishing Co., Tulsa

inspissated, or thickened, deposits that have lost the lighter fractions through evaporation. In southern Oklahoma around the Ouachita and Arbuckle mountains there are many deposits of viscous or semisolid asphalt occurring in rocks ranging from Early Paleozoic to Late Mesozoic (about 100,000,000 to 500,000,000 years ago) in age. The deposits in this area probably represent inspissated oil pools that have been exposed by erosion. One of the largest known deposits of oil in the world, some 300,000,000,000 barrels, is represented by the Athabaska oil sands of Cretaceous age (65,000,000 to 136,000,000 years ago) in Alberta, Canada. Another great area of known tar sands lies just north of the Orinoco River in Venezuela on the southern flank of the Materin Basin, extending 375 miles (600 kilometres) in an east–west direction and from 25 to 50 miles (40 to 80 kilometres) wide. It is estimated that about 700,000,000,000 barrels of oil are locked up in this deposit of Lower Tertiary age (about 60,000,000 years old). A great deal of research has gone into mapping of the potential mining areas and building of pilot plants for the recovery of oil by distillation of the tar sands.

Subsurface occurrences. Until recently exploration for oil involved the search for sedimentary sections or sequences of rocks located on land areas of the continents, and as experience and technology has progressed, this search has probed ever deeper into the earth. It has now included essentially all known accessible geologic basins and prospective areas and has developed about four-fifths of the proved world petroleum reserves. The types of structures in which these reserves have been found are many and varied; no two fields anywhere in the world are exactly alike. The search for new oil and gas fields on the land areas will continue, particularly in the more inaccessible areas such as the Arctic north, the desert areas of North Africa and the Near East, and the jungle areas of the Far East. Fields that are more difficult to find and more expensive to produce will be added to the world's available supply of crude oil, but there is a noticeable downward trend in the land exploration rate throughout the world.

Offshore drilling and exploration

The petroleum industry started developing oil and gas fields located under the waters of the continental shelf off the coast of Louisiana about 20 years ago. This offshore drilling was a logical step from the sea-marsh drilling of the adjoining delta area, where everything pointed toward a continuation of the same sedimentary environments and types of structures that had proved so rewarding on land. It became apparent that the only difference between land and sea drilling was the greatly increased cost of the latter, and this directed attention to similar areas throughout the world. It also became apparent that many of the deep sedimentary basins of the world were located along the margins of existing continents and that in many cases the conditions for the accumulation of oil and gas improved seaward. Major offshore discoveries are currently being developed in the North Sea between England and Norway, in the Bass Straits between Australia and Tasmania, in the Nigerian offshore delta, near the California coast, in the Persian Gulf, and in other areas throughout the world. As the industry enters this virgin field of offshore water bottoms, it can be expected that the early discoveries will be located on the larger, easier to find structures.

For a period of years, water depths up to a few hundred feet were considered to represent the outer limit of oil exploration because of technological problems and cost. When favourable structures were reported to occur out to depths beyond the continental shelf (normally considered about 600 feet [180 metres]) on the much deeper conti-

nental slope and continental rise, the industry began concentrating on the ability to drill in ever deeper water. This has been spurred by testing of the Sigsbee Knolls in the deepest portion of the Gulf of Mexico, in water depths of over 12,000 feet (5,500 metres), where oil was found almost 500 feet (150 metres) below the sea floor in the calcite–gypsum cap rock of a salt dome. Similar diapiric features have been mapped in the Mediterranean Sea, and in the southern Atlantic off the coasts of Africa and Brazil salt domes have been mapped and tested in preliminary fashion.

A map indicating the locations of sedimentary basins and the oil- and gas-producing areas of the world is shown in Figure 9. Also shown on this map are the continental-shelf areas, in many instances the scene of current exploration. More than 60 percent of the world's known oil reserves occur in the small area at the head of and bordering the Persian Gulf; a large proportion of this is stored in a few giant reservoirs, such as Burgan in Kuwait and Ghawar in Saudi Arabia.

Considering the giant oil fields of the world, 8 percent of the oil is found in rocks of the Paleozoic Era (from 225,000,000 to 570,000,000 years in age), 63 percent in rocks of Mesozoic age (from 65,000,000 to 225,000,000 years), and 29 percent in Tertiary and Pleistocene age (from 10,000 to 65,000,000 years ago). If the great Middle East oil reserves were to be excluded from these figures, the youngest geologic period would have the greatest reserve, because most of the Persian Gulf Basin reserves are Mesozoic in age.

Eighty-nine percent of the giant fields of the world are located on anticlines, and only 9 percent are in stratigraphic or partly stratigraphic traps. These figures partly reflect the fact that the more easily found fields are discovered first. As the industry becomes more sophisticated in its search, more of the obscure reservoirs will be found and tested.

Type examples of major fields. *Bay Marchand–Timbalier Bay–Caillou Island.* This salt complex forms a single salt ridge that extends 28 miles (45 kilometres) along the Louisiana coast. The salt of the Late Triassic–Early Jurassic Period (about 190,000,000 years ago) has risen from depths of more than 40,000 feet (12,000 metres) to within 2,000 feet (600 metres) of the surface. The reserves of about 1,700,000,000 barrels are trapped in a multitude of sands ranging in age from Late Miocene to Pleistocene (from about 2,500,000 to 10,000,000 years ago) and in depth from 1,000 to more than 20,000 feet (300 to 6,000 metres). The form of traps associated with the salt uplift include superdomal arching above the salt, truncation of sand reservoirs against the diapiric salt and shale, and traps formed by faulting associated with the uplift.

Prudhoe Bay. This structure in the North Slope Province of Alaska covers an area about 45 by 18 miles (72 by 29 kilometres) and is a broad, anticlinal nose that plunges gently to the northwest. The reservoirs are terminated to the east by truncation against an eastward dipping unconformity. The major oil accumulations occur in very porous quartz sandstone members of the Prudhoe Bay section, mainly Triassic in age (190,000,000 to 225,000,000 years in age). The sands vary in thickness up to 600 feet (180 metres) and lie mainly in the 8,000- to 9,000-foot- (2,400- to 2,700-metre-) depth range. At the present time the field is estimated to contain 9,600,000,000 barrels of recoverable oil and 26,000,000,000,000 cubic feet of gas, making it by far the largest field in North America.

Hassi Messaoud. This oil field, in the eastern portion of the Algerian Sahara Desert and 350 miles (560 kilometres) south of the Mediterranean, produces from a Cambrian (500,000,000 to 570,000,000 years in age) sandstone reservoir that forms a large dome. The depth to the reservoir is about 11,000 feet (3,400 metres), the oil column is about 900 feet (270 metres) thick, and the productive area covers about 500 square miles (1,300 square kilometres). The structure was formed shortly after deposition, but later periods of uplift and erosion exposed the top of the ancient Hassi Messaoud structure, after which 10,000 feet (3,000 metres) of Mesozoic sediments were deposited. Late Silurian (395,000,000 to 430,000,000 years ago) shale zones located on the flanks of the structure are considered to be the primary source beds; if this is so, then the Silurian-generated oil was lost during the pre-Triassic erosion. After the deposition of Tertiary rocks (2,500,000 to 65,000,000 years of age), the Silurian shales would have to yield another generation of hydrocarbons, which migrated downward into the Cambrian reservoir rock. The porosity of the reservoir averages about 8 percent and the permeability ranges from zero to 1,000 millidarcies. There is no gas cap, and the oil is undersaturated. The ultimate recoverable oil is estimated at 5,000,000,000 barrels.

Ghawar. Located about 50 miles (80 kilometres) inland from the western shores of the Persian Gulf in Saudi Arabia, the Ghawar field is a structural accumulation along 140 miles (225 kilometres) of a north–south trending anticline. The productive area covers approximately 900 square miles (2,300 square kilometres) and the vertical oil column is about 1,300 feet (400 metres). The oil–water contact occurs at a depth of about 6,500 feet (2,000 metres) near the middle of the structure, and it dips from the southwest to the northeast. The reservoir, consisting of shallow-water carbonate sediments of Upper Jurassic age (about 150,000,000 years), is sealed by a cap rock of anhydrite evaporites. The reservoir rock varies from sand size to coarser carbonate sediments in which recrystallization and dolomitization is common. The API gravity varies from 32° to 36°, the oil is undersaturated, and there was no initial gas cap. The structure began to grow by Lower Cretaceous time (about 130,000,000 years ago), following the deposition of the reservoir beds, and its growth was probably completed by Miocene time (7,000,000 to 26,000,000 years ago).

It is believed that the source beds for the oil were the same formations that now form the reservoir rock, but that secondary migration into the Ghawar structural area took place during the Upper Cretaceous and Early Tertiary. The Ghawar field is generally recognized as the largest in the world, having a recoverable reserve of about 75,000,000,000 barrels of oil.

Burgan. The structure of this oil field, about 15 miles (24 kilometres) west of the Persian Gulf in the small country of Kuwait, is a large, broad anticlinal fold with a north–south trend. Production is from thick Middle Cretaceous sandstones (about 100,000,000 years old) that were first encountered at about 3,700 feet (1,100 metres); the productive sands comprise 800 feet (240 metres) out of a total of 1,100 feet (340 metres) of section. The development of sandstone reservoirs at Burgan is in striking contrast to most of the oil fields of Saudi Arabia, Iraq, and Iran, in which the reservoir rocks are limestone. The oil is asphaltic in type with a gravity of API 30° and has a high sulfur content of 2 to 2.5 percent. The recoverable reserves of the Burgan oil complex are estimated at about 66,000,000,000 barrels, making it the second largest oil field in the world.

Golden Lane. The Golden Lane fields of Mexico lie just south of Tampico and form a north–south-oriented oval bisected by the coastline and measuring 100 by 40 miles (160 by 64 kilometres). The oil fields, together with some undrilled offshore structures, trace in continuous fashion a Middle Cretaceous (about 100,000,000 years in age) atoll rim that represents an ancient buried reef more than 5,000 feet (1,500 metres) thick. Within the atoll rim, evaporites, calcareous sands, and dolomites were deposited; elsewhere the reservoir rocks are fossiliferous limestones. One of the wells drilled on the western, or land, portion of the atoll produced more than 95,000,000 barrels of oil, probably the world's record from a single well. The depth to the producing section in this area varied from 1,600 to 2,600 feet (500 to 800 metres), and porosities and permeabilities were so high that these fields were rapidly depleted. The marine portion of the atoll was only recently discovered. The whole reef structure was sharply tilted to the southeast during the Early and Middle Tertiary (from 2,500,000 to 65,000,000 years ago), and the hydrocarbons of the reef

probably migrated from southeast to northwest during this period. The group of fields in the western portion of the ancient atoll have produced more than 1,250,000,000 barrels of oil.

Proved world reserves and future provinces. Tabulated below are the estimated proved reserves of crude oil for the world, by country, together with the production during 1970 in barrels per day.

Estimated Proved World Reserves of Crude Oil (1970)

continent and country	crude reserves at year end (000 barrels)	reserves (percent of world)	annual production (000 barrels)
North America	53,128,315	9.7	4,131,155
Canada	8,558,980	...	438,023
Mexico	5,508,000	...	117,599
U.S.	39,001,335	...	3,515,533
West Indies	605,716	0.1	52,722
Trinidad and Tobago	605,000	...	51,043
Others	716	...	1,679
South America	24,970,352	4.6	1,678,830
Argentina	1,572,500	...	143,410
Bolivia	300,000	...	8,820
Brazil	857,000	...	60,003
Chile	120,000	...	12,410
Colombia	1,580,000	...	73,000
Ecuador	6,000,000	...	1,500
Peru	500,000	...	26,269
Venezuela	14,040,852	...	1,353,418
Western Europe	6,489,200	1.2	117,714
Austria	189,700	...	18,994
Denmark	500,000	...	...
France	120,000	...	17,225
Germany, West	544,500	...	54,974
Italy	305,000	...	11,603
The Netherlands	230,000	...	13,108
Norway	3,500,000	...	...
Spain	100,000	...	1,168
U.K.	1,000,000	...	642
Eastern Europe	59,460,000	10.9	2,719,031
Albania	‡	...	10,950
Bulgaria	278,000	...	2,190
Czechoslovakia	‡	...	1,460
Germany, East	‡	...	1,825
Hungary	‡	...	14,600
Poland	‡	...	3,066
Romania	867,000	...	98,915
U.S.S.R.	58,000,000	...	2,565,950
Yugoslavia	315,000	...	20,075
Northern Africa	39,477,400	7.2	1,736,916
Algeria	8,098,300	...	373,090
Egypt	990,100	...	119,229
Libya	30,000,000	...	1,212,048
Morocco	9,000	...	323
Tunisia	380,000	...	32,226
Central and southern Africa	6,840,000	1.2	471,486
Angola§	750,000	...	35,770
Congo (Brazzaville)	5,000	...	146
Gabon	485,000	...	39,729
Nigeria	5,600,000	...	395,841
Asia, Middle East	342,133,500	62.6	5,118,424
Abu Dhabi	14,850,000	...	253,231
Bahrain	394,500	...	27,973
Dubai	1,300,000	...	30,710
Iran	60,000,000	...	1,397,220
Iraq	29,000,000	...	561,640
Israel	2,000	...	548
Kuwait	75,500,000	...	998,110
Neutral Zone	12,800,000	...	183,435
Oman	4,000,000	...	121,310
Qatar	4,100,000	...	132,357
Saudi Arabia	138,667,000	...	1,386,659
Syria	1,320,000	...	...
Turkey	200,000	...	25,231
Asia, Far East	872,152	0.2	64,238
Burma	50,000	...	6,570
India	742,700	...	49,275
Japan	25,452	...	5,656
Pakistan	35,000	...	2,555
Taiwan	19,000	...	182
Oceania	12,387,019	2.3	445,688
Australia	1,661,019	...	65,148
Brunei-Malaysia	700,000	...	58,400
Indonesia	10,000,000	...	321,629
New Zealand	26,000	...	511
Total world	546,363,654	100.0	16,536,204

*Estimate. †Incomplete. ‡Withheld by government.
§Including Cabinda.
Source: Figures compiled by *World Oil*, with aid of government agencies, industry associations, oil companies, and other sources.

It has been estimated that an additional 800,000,000,000 to 900,000,000,000 barrels of oil remain to be found in the world, of which about 40 percent will occur within the Communist countries. Although most of this oil and the oil still to be found in the Middle East will be found onshore, most of the potential oil estimated for the rest of the world will be found on the offshore continental shelf and slope, out to water depths of about 2,000 feet (600 metres). Exploratory activity around the world affirms the industry's belief in this statement. The continental shelves range in width from a few to more than 600 miles (950 kilometres), and form between 7 and 8 percent of the total ocean floor. Among the non-Communist countries' largest continental shelf areas are those of the Bering Sea off Alaska, the South China Sea, the Arafura Sea north of Australia, the Sunda shelf of Indonesia, the Atlantic shelf off Argentina, the shelf off northern and eastern Canada and the eastern United States, and the northwestern European shelf.

Of particular interest to explorationists are the marginal semi-enclosed seas, such as the North Sea, the Persian Gulf, the Irish Sea, Hudson Bay, the Gulf of St. Lawrence, the Black Sea, the Caspian Sea, the Red Sea, the Adriatic, and the Balearic Sea, each of which has an adequate sedimentary section and relatively shallow water depth. Their restricted nature favours the accumulation and preservation of the organic source material of petroleum, and favourable reservoirs are to be expected. Their environment is also favourable for evaporite deposition; diapiric salt and shale structures have already been mapped in some of them, and other marginal basins show the continuation of folding trends from the adjacent land areas.

In summary, with the exception of the Soviet Union, China, and the Canadian Arctic Islands, relatively few new petroleum provinces of major importance can be expected to be found on the earth's land areas. The submerged continental margins and inland seas, however, are already the scene of active exploration, and this search will certainly provide a number of major new productive areas during the next 20 years. Whether or not the oil industry can continue to supply the rising world demand for crude oil remains to be seen, however.

BIBLIOGRAPHY. ROBERT H. DOTT and MERRILL J. REYNOLDS (comps.), *Sourcebook for Petroleum Geology* (1969), is an excellent collection of petroleum geology scientific concepts. A.I. LEVORSEN, *Geology of Petroleum* (1954), is the outstanding textbook in its field. WALLACE E. PRATT and DOROTHY GOOD (eds.), *World Geography of Petroleum* (1950); and HAROLD F. WILLIAMSON and ARNOLD R. DAUM, *The American Petroleum Industry: The Age of Illumination, 1859–1899* (1959), contain historical data on oil occurrences and on the development of the petroleum industry. KENNETH K. LANDES, *Petroleum Geology*, 2nd ed. (1959); GRAHAM B. MOODY (ed.), *Petroleum Exploration Handbook* (1961); and JOHN D. HAUN and L.W. LEROY (eds.), *Subsurface Geology in Petroleum Exploration* (1958), are sources of information on theories of origin and accumulation of petroleum as well as practical applications of scientific knowledge to petroleum problems. The AMERICAN ASSOCIATION OF PETROLEUM GEOLOGISTS symposia *Diapirism and Diapirs* (1968), *Fluids in Subsurface Environments* (1965), *Geology of Giant Petroleum Fields* (1970), and *Habitat of Oil* (1958) contain collections of scientific papers in specific areas of petroleum technology. The AMERICAN PETROLEUM INSTITUTE, *Petroleum Facts and Figures*, and the UNITED STATES BUREAU OF MINES, *Minerals Yearbook*, include annual statistical reviews of the petroleum industry. The *American Association of Petroleum Geologists Bulletin* is a monthly scientific journal. *Oil and Gas Journal* (weekly), *World Oil* (14/yr.), and *Offshore* (monthly) are regularly published trade magazines of the petroleum industry.

(G.I.A.)

Petroleum and Gas Extraction

Though petroleum and gas extraction normally involves prospecting, drilling, and production techniques used in the exploitation of petroleum deposits, the transportation and storage of petroleum and gas are also included within the scope of this article.

Nature and occurrence. Petroleum is a naturally occurring hydrocarbon material of animal and vegetable

Table 1: World Crude Oil Production and Producing Wells

	producing oil wells*				annual production (000 barrels)		daily average production (barrels)		
	January 1, 1971			beginning of 1970 total	1970	1969	1970	1969	percent difference
	flowing	artificial lift	total						
North America	52,307	487,833	540,140	560,003	4,131,155	3,933,104	11,318,233	10,775,657	+ 5.0
Canada	3,938	15,750	19,688	19,268	438,023	393,403	1,200,063	1,077,816	+ 11.3
Mexico	1,441	1,834	3,275	3,095	177,599	167,831	486,573	459,811	+ 5.8
U.S.	46,928	470,249	517,177	537,640	3,515,533	3,371,870	9,631,597	9,238,000	+ 4.3
West Indies	831	2,772	3,603	3,582	52,722	59,254	144,443	162,339	− 11.0
Trinidad and Tobago	807	2,698	3,505	3,454	51,043	57,429	139,843	157,339	− 11.1
Others	24	74	98	128	1,679	1,825	4,600	5,000	− 8.0
South America	3,627	15,886	19,513	19,206	1,678,830	1,632,396	4,599,535	4,472,230	+ 2.8
Argentina	500	4,860	5,360	5,591	143,410	130,085	392,904	356,398	+ 10.2
Bolivia	...	...	...	...	8,820	14,750	24,164	40,411	− 40.2
Brazil	312	727	1,039	996	60,003	57,030	164,393	156,247	+ 5.2
Chile	273	88	361	351	12,410	13,350	34,000	36,575	− 7.0
Colombia	...	...	...	...	73,000	77,126	200,000	211,303	− 5.3
Ecuador	19	606	625	605	1,500	1,607	4,110	4,404	− 6.7
Peru	...	...	...	...	26,269	26,618	71,970	72,927	− 1.3
Venezuela	2,523	9,605	12,128	11,863	1,353,418	1,311,830	3,707,994	3,594,055	+ 3.2
Western Europe	236	912	2,148	2,196	117,715	123,896	322,505	339,442	− 5.0
Austria	151	1,100	1,252	1,263	18,994	18,723	52,039	51,295	+ 1.4
France	30	251	281	294	17,225	18,316	47,193	50,182	− 6.0
Germany, West	...	...	...	...	54,975	57,729	150,616	158,462	− 4.8
Italy	54	161	215	225	11,603	11,663	31,789	31,954	− 0.5
Netherlands, The	0	346	346	352	13,108	13,786	35,911	37,769	− 4.9
Spain	0	19	19	20	1,168	3,087	3,199	8,458	− 62.2
U.K.	1	35	36	42	642	592	1,758	1,622	+ 8.4
Eastern Europe†	...	...	...	...	2,719,250	2,538,222	7,450,000	6,954,031	+ 7.1
Albania	...	...	...	...	10,950	7,673	30,000	21,021	+ 42.7
Bulgaria	...	...	...	...	2,190	3,442	6,000	9,429	− 36.4
Czechoslovakia	...	...	...	...	1,460	1,352	4,000	3,705	+ 8.0
Germany, East	...	...	...	...	2,044	2,117	5,600	5,800	− 3.4
Hungary	...	...	...	...	14,600	13,385	40,000	36,671	+ 9.1
Poland	...	...	...	...	3,066	4,074	8,400	11,161	− 24.7
Romania	...	...	...	...	98,915	98,100	271,000	268,767	+ 0.8
U.S.S.R.	...	...	...	...	2,565,950	2,389,061	7,030,000	6,545,372	+ 7.4
Yugoslavia	...	...	...	...	20,075	19,018	55,000	52,105	+ 5.6
Africa, North	1,478	612	2,090	2,082	1,736,916	1,601,936	4,758,676	4,388,867	+ 8.4
Algeria	785	41	826	768	373,090	350,064	1,022,165	959,080	+ 6.6
Egypt	87	115	202	256	119,229	88,480	326,656	242,411	+ 34.7
Libya	527	443	970	952	1,212,048	1,134,838	3,320,680	3,109,146	+ 6.8
Morocco	39	10	49	59	323	422	885	1,157	− 23.5
Tunisia	40	3	43	47	32,226	28,132	88,290	77,073	+ 14.6
Africa, central and southern	511	133	684	555	614,430	434,112	1,291,742	688,751	+ 87.5
Angola and Cabinda	55	82	137	127	35,770	17,228	98,600	47,200	+107.6
Congo (Brazzaville)	...	...	...	...	143,090	182,900	400	500	− 20.0
Gabon	88	51	139	130	39,729	36,283	108,847	99,405	+ 9.5
Nigeria	408	0	408	298	395,841	197,701	1,084,495	541,646	+100.2
Asia, Middle East	2,117	929	3,164	3,038	5,117,328	4,592,369	14,020,081	12,581,834	+ 11.4
Abu Dhabi	115	0	115	102	253,231	219,404	693,783	601,106	+ 15.4
Bahrain	183	44	227	220	27,973	27,774	76,638	76,093	+ 0.7
Dubai	12	0	12	10	30,710	3,800	84,138	10,411	+708.2
Iran	324	0	324	256	1,397,220	1,231,828	3,828,000	3,374,872	+ 13.4
Iraq	...	...	118	113	560,545	552,000	1,535,741	1,512,329	+ 1.5
Israel	0	33	33	34	547	657	1,500	1,800	− 16.7
Kuwait	691	0	691	649	998,110	940,041	2,734,547	2,575,455	+ 6.2
Neutral Zone	51	328	379	341	183,435	156,056	502,563	427,550	+ 17.5
Oman	80	0	80	57	121,310	131,400	332,357	360,000	− 7.7
Qatar	91	0	91	88	132,357	129,742	362,622	355,458	+ 2.0
Saudi Arabia	545	265	810	888	1,386,659	1,173,898	3,799,065	3,216,160	+ 18.1
Turkey	25	259	284	280	25,231	25,769	69,127	70,600	− 2.1
Asia, Far East	1,197	1,206	2,430	2,535	64,238	64,819	175,995	177,591	− 0.9
Burma	...	...	...	...	6,570	6,898	18,000	18,900	− 4.8
India	355	36	391	340	49,275	49,079	135,000	134,464	+ 0.4
Japan	811	1,170	1,981	2,131	5,656	5,502	15,495	15,075	+ 2.8
Pakistan	...	...	27	27	2,555	2,759	7,000	7,560	− 7.4
Taiwan	31	0	31	37	182	581	500	1,592	+ 31.4
Oceania	548	2,411	2,959	2,885	445,177	336,518	1,219,663	921,967	+ 32.3
Australia	97	283	380	359	65,148	15,805	178,480	43,301	+312.2
Brunei-Malaysia	91	329	420	410	58,400	49,786	160,000	136,400	+ 17.3
Indonesia	360	1,799	2,159	2,116	321,629	270,927	881,174	742,266	+ 18.7
Total world†	62,892	543,694	576,731	596,082	16,677,761	15,316,626	45,300,873	41,462,799	+ 9.2

*Estimated. †Eastern Europe producing oil wells not included in world total.

origin. First formed in deep sedimentary beds, petroleum later migrated to its present position in porous underground rocks. Petroleum occurs as either a liquid, a gas, or a combination of both. Occasional deposits, elevated during the formation of mountain ranges, have been exposed by erosion to form tar (sometimes called pitch) deposits, and some of these have been known and used throughout recorded history. Other near-surface deposits of liquid petroleum seep slowly to the surface through natural fractures and fissures in the overlying rock. Accumulations from these oil seeps, called rock oil, were used commercially in the early 19th century. The vast majority of petroleum deposits, however, lie in natural rock at depths from 500 to 25,000 feet (about 150 to 7,600 metres) below the surface of the ground. The deposits are prevented from migrating upward by intervening layers of dense rock. As a general rule, the deeper deposits have higher internal pressures and contain greater quantities of the gaseous hydrocarbons.

In order to create a petroleum trap or reservoir, a porous rock (one capable of containing liquids), such as sandstone or limestone, must be surrounded on the top and sides by nonporous rock, such as shale. In such an event, either gaseous or liquid petroleum, which is less

dense than the always-present water, may be trapped in the top portion of the porous formation.

Since petroleum reservoirs are the result of an accumulation of liquid or gas in rocks originally containing water, there will always be some water left in the reservoir. Water production often accompanies the production of oil and gas.

History. When it was discovered in the 19th century that rock oil could yield a distilled product (kerosene) suitable for lanterns, sources of this oil were eagerly sought. It is now generally agreed that the first well drilled specifically to find petroleum was that of one Col. Edwin Drake in Titusville, Pennsylvania, in 1859. The success of this well, drilled close to an oil seep, led to further drilling in the same vicinity and soon to exploration elsewhere. The growing demand for petroleum products led quickly to oil wells not only in the United States but in other countries. At the end of 1969, there were almost 600,000 producing oil wells in over 60 different countries (see Table 1). Oil is produced in every continent except Antarctica.

First drilled well

Because of the widespread distribution of petroleum reservoirs, it was possible for new discoveries to be made at a rapid rate through the first half of the 20th century. As one result, there was little incentive for the development of efficient production techniques. Because little use had been found for the natural gas produced along with the crude oil, the gas often was burned at the wells. Such wasteful practices were gradually halted, partly by the increasing difficulty of finding new petroleum reservoirs and partly by the increased uses found for natural gas. New engineering methods led to an economically optimum recovery from each reservoir. These methods often were based on the injection of some less expensive fluid, such as water, to force additional petroleum into the producing wells.

As the value of the petroleum produced reached impressive proportions, governmental agencies in every producing area sought to prevent wasteful exploitation.

PROSPECTING AND EXPLORATION

Drake's original well was drilled close to a known surface seepage of crude oil, and for many years such seepages were the only reliable indicators of the presence of oil. As demand grew, however, new methods were devised for exploring the petroleum potential of underground rock formations. There are now three major exploration methods: (1) surface feature mapping, (2) seismograph observations, and (3) earth gravity surveys. But there is still no accurate way to predict the actual presence of underground petroleum deposits. Exploration methods carried out on the surface can only indicate the presence or absence of underground formations that are favourable for petroleum accumulation.

Surface methods. A seepage of crude oil sometimes appears as a tarlike deposit in a low area such as the oil springs at Baku, on the Caspian Sea, described by Marco Polo. More often, it can be seen as a thin skim of oil on small creeks that pass through the seepage area. This latter phenomenon was responsible for the naming of Oil Creek (a place in Pennsylvania), where Drake's well was drilled. Such seepages of natural gas cannot usually be seen, though instruments can detect natural gas concentrations in air down to less than one part in 100,000. Exploration methods have been used in which air samples from areas immediately over or under the ground surface are tested. Such samples taken from a wide area can yield a map of gas concentration points. A similar method has been used to test for traces of petroleum gases dissolved in seawater in the hope that undersea seepages could be thus detected. These geochemical prospecting methods are not applicable to the large majority of underground reservoirs that do not have leakage to the surface.

Oil Creek

A second class of surface prospecting methods is based on surface indications of underground rock formations. In some cases, subsurface folds and faults in rock formations are repeated in the surface features. Also, underground salt domes (*q.v.*) may announce their presence by a low bulge in an otherwise flat ground surface. Since these features usually produce possible petroleum traps, drilling in the area may be productive.

Seismographic methods. Most folds and faults in subsurface rocks are not apparent at the ground surface. Seismic methods that have been useful for mapping such underground formations involve the transmission and echo-producing properties of rock in a manner similar to underwater sonar methods. Sound waves travel at differing velocities depending on the types of rock formations and are reflected at interfaces between different types of rock. The sound wave source is usually a small explosion in a shallow drilled hole. Microphones are placed at various distances and in various directions from the source to record the arrival times of the echoes. This procedure is repeated at intervals over a large area. An experienced seismologist can then interpret the records to map the underground formation contours. Possible locations for petroleum accumulations may then be located for drilling.

Gravity meter surveys. The gravity meter has become a primary exploration tool. Although gravity at the earth's surface is very nearly a constant, it is slightly greater where large rock formations lie closest to the surface. The gravity field, therefore, increases over the tops of anticlinal (downward-inclined) folds and other similar subsurface anomalies. A salt dome results in lowered gravity because salt is less dense than other rocks. Gravity surveys are usually carried out from the air. A sensitive instrument, the gravity meter, is suspended on a cable below the aircraft, while the recording devices are operated from inside. As the aircraft is flown over an area on a precise grid pattern, a map showing the gravity strengths is prepared from the recordings. These maps can show the presence of such possible accumulation spots as anticlinal folds, fault blocks, and salt domes, even though there are no surface indications of such. The gravity meter survey is faster and less expensive than the seismic method. It can locate only large rock features that exist relatively close to the surface, however, and is generally used as a preliminary survey tool to locate areas in which seismic work may profitably be concentrated.

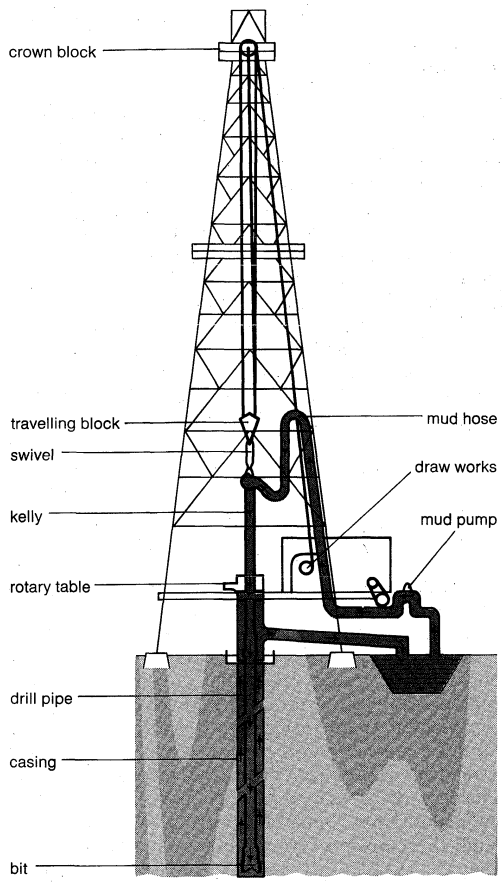
DRILLING AND COMPLETING WELLS

Early wells were drilled with impact-type tools in a method called cable-tool drilling, in which a chisel-like bit was suspended from a cable to a lever at the surface, and an up-and-down motion of the lever caused the bit to pound the bottom of the hole and chip away the rock. Such drilling had to be interrupted periodically to allow loose rock chips to be removed with a basket-like device on the cable. Also, the well had to be free of liquids during the drilling process so that the bit could effectively remove rock. These necessarily dry holes allowed oil and gas to flow to the surface in large quantities as soon as a producing formation was penetrated. This created the image of a "gusher" as a successful well. Often, a large amount of oil was wasted before the well could be capped and brought under control.

Cable-tool drilling

Rotary drilling. During the last 40 years, rotary drilling has become the preferred method for penetrating underground formations (see illustration). In this method, a special tool—the drill bit—rotates while bearing down on the bottom of the well, thus gouging and chipping its way downward. Probably the greatest advantage of rotary drilling over cable tooling is that the well bore is kept full of liquid during drilling. A weighted fluid (called drilling mud) in the bore hole serves two important purposes: by its hydrostatic pressure, it prevents the entry of formation fluids into the well, thus preventing "blowouts" and "gushers." In addition, the drilling mud carries the crushed rock to the surface, so that the drilling is continuous until the bit wears out.

The drill bit is connected to the surface equipment through a drill pipe, a heavy-walled tubing through which the drilling mud is fed to the bottom of the bore hole. In most cases, the drill pipe also transmits the rotary motion from a turntable at the surface to the bit at the bottom of the hole. The top piece of the drill pipe is



Rotary drilling rig.
By courtesy of the American Petroleum Institute

a tube of square or octagonal cross section called the kelly, which passes through a square or octagonal hole in the turntable. At the bottom end of the drill pipe are extra-heavy sections called drill collars that serve to concentrate the weight on the rotating bit. The drilling mud leaves the drill pipe in such a way that it washes the loose rock from the bottom and carries it to the surface. Drilling mud is carefully formulated to assure correct weight and viscosity characteristics for its required tasks. After screening to remove the rock chips, the returning mud is held in open pits to be recirculated through the well. The mud is picked up by piston pumps and forced through a swivel joint into the top of the drill pipe.

The characteristic tall derrick contains the hoisting equipment that is used to raise and lower the drill pipe into the well. The drill bit wears quickly and requires frequent replacement, making it necessary to pull the entire drill string from the well and stack it at one side of the derrick. Joints of drill pipe are usually 30 feet (nine metres) long. Sections of two or three such joints are separated and stacked vertically while the bit is being changed. Drilling mud is left in the bore hole during this time to prevent excessive flow of fluids into the well.

Such rotary drilling techniques as those described have made it possible for wells to be drilled to depths of more than 25,000 feet (7,600 metres). Wells have been drilled in both coastal waters and in the open sea. In relatively shallow waters (less than 200 feet or 60 metres), the wells are usually drilled from platforms that rest on the bottom; the drilling platform is constructed above the highest expected waves. In deeper waters and for exploratory work, drilling is done from a floating vessel that is kept in place by a combination of anchors and motors. Since the fixed platforms are quite expensive, many (20 or more) wells are often drilled from the same platform. Yet for most efficient coverage, the wells must enter the producing formation with more than 1,000 feet (300 metres) between them. Thus, each well except one has to be drilled in a precise slantwise direction. This directional drilling is achieved with a tool called a whipstock, which causes the drill bit to veer slightly in the desired direction. The whipstock is used at intervals until the total desired angular deviation is reached. Sharp bends in the bore hole must be avoided because the rather rigid drill pipe has to bend through the deviation while rotating.

Use of a floating vessel

A recent variation in rotary drilling has involved introducing a fluid-powered turbine at the bottom of the bore hole to produce the rotary motion of the bit. In this method, the drill pipe does not rotate but is used to regulate the weight on the bit and to carry the drilling mud that turns the turbine. In the early 1960s the Turbodrill was used extensively in eastern Europe.

Well logging and drill-stem testing. After the bore hole has penetrated the possibly productive formations, these formations must be tested to determine if expensive completion procedures should be used. The first evaluation is usually made by well-logging methods, in which the logging tool is lowered past the formations while response signals are relayed to operators on the surface. Often these tools make use of the difference in electrical conductivities of rock, water, and petroleum to detect possible oil or gas accumulations. Other logging tools depend on difference in absorption of atomic particles. With well-logging tools, the exact depths of possibly productive formations are determined. Only a production test, however, can establish the potential productivity of a well.

The production test normally used is the so-called drill-stem test, in which the bore hole is sealed above and below the formation to be tested, with only the drill pipe open to the formation. The drill pipe is then emptied of the drilling mud so that the formation fluid may enter. When enough time has passed, the openings into the drill pipe are closed, and the pipe containing the formation fluid is brought to the surface for analysis. If there is gas present, the gas may flow from the top of the drill pipe during the test.

Well completion. If the preliminary tests show that one or more of the formations in the bore hole will be commercially productive, the well must be prepared for the continuous production of the oil or gas. First, a large outside pipe, or casing, slightly smaller in diameter than the drill hole, is inserted to the full depth of the well. A cement slurry is forced between the outside of the casing and the inside surface of the drill hole. When set, this cement forms a seal so that fluids cannot pass from one portion of the well to the other through the bore hole. The casing is usually about nine inches (23 centimetres) in diameter. It creates a permanent well through which the productive formations may be reached. After the casing is in place, a production string of smaller (about three inches, or eight centimetres, in diameter) tubing is extended from the surface to the productive formation with a packing device to seal the productive interval from the rest of the well. If multiple productive formations are found, as many as four production strings of tubing may be hung in the same cased well. If a pump is needed to lift oil to the surface, it is placed on the bottom of the production string.

Since the casing is sealed against the productive formation, openings must be made to allow the oil or gas to enter the well. A down-hole perforator uses an explosive to shoot holes through the casing and cement and into the formation. The perforator tool is lowered through the tubing on a wire line. When it is in the correct position, the charges are fired electrically from the surface. Such perforating will be sufficient if the formation is quite productive. If not, an inert fluid may be injected into the formation at pressures high enough to fracture the rock around the well and thus open more flow passages for the petroleum. In early times, nitroglycerin was exploded in the well bore for the same purpose. An acid that dissolves rock may be used in place of the inert fluid.

Christmas tree valves

When the subsurface equipment is in place, a network of valves (called a Christmas tree) is installed on the surface, arranged so that flow from the well may be regulated and measured and so that tools to perform subsurface work may be introduced through the tubing. Christmas

trees may be very simple, such as those found on a low-pressure well that must be pumped, or they may be very complex, as in the case of a high-pressure well with multiple producing strings.

Artificial lift. Many oil wells do not have a formation pressure high enough to push the oil (which weighs about six pounds per gallon, or 0.7 kilogram per litre) to the surface. In these cases, some artificial method for lifting the oil to the surface must be installed. The most common installation involves a motor and "walking beam" (like a seesaw) on the surface that operates the pump on the bottom of the production string. A chain of solid metal rods connects the beam and the pump. Another method, called gas lift, uses the buoyancy of a bubble of gas in the tubing to push the oil to the surface. A third type of artificial lift forces some of the produced oil down the well at high pressure to operate a pump at the bottom of the well. Even though an oilfield may have enough pressure to produce naturally at first, artificial lift will usually be required in the later stages of production. Gas wells that produce little or no liquid do not need artificial lift devices.

Surface equipment. Water often flows into the well along with oil and gas. Surface equipment separates the water from the petroleum and separates the oil and the gas for storage and transportation. Also, the production from a number of wells must be gathered at a central point for distribution. A settling tank is usually used to separate the water, although either heat or chemical treatment may be necessary to break down emulsions. Gas is separated from the oil by allowing the produced stream to pass through a series of pressure-reducing valves and removing the liberated gas to storage. The remaining crude oil is kept at normal pressures in cylindrical tanks after being brought through pipelines from the individual wells. Gas is usually piped directly to a central processing plant (called a gasoline plant) where remaining liquids are removed before it is fed to the consumer pipeline.

RECOVERY OF OIL AND GAS

As indicated above, both oil and gas reservoirs usually start with a formation pressure that forces fluid into the well and sometimes to the surface. With artificial lift, reservoir fluids may be produced as long as enough reservoir pressure remains to cause the fluids to flow into the well. Since a decline in reservoir pressure invariably accompanies fluid production, a point is reached at which commercial quantities of oil or gas no longer flow into the well. In natural gas reservoirs, this does not usually occur until most of the gas is recovered. In nearly all crude-oil fields, however, less than one-third of the oil originally present can be produced using only the reservoir pressure.

Secondary recovery of oil. When a large part of the crude oil cannot be recovered simply by allowing the original reservoir pressure to furnish the driving energy, a method for supplying extra energy must be found. Usually, an injected fluid such as water or gas replaces the produced oil and thus maintains the reservoir pressure. When this procedure substantially increases the potential production of oil, it is called secondary recovery. When gas alone is used as the injected fluid, it is usually put into the top of the reservoir where gas normally collects to form a gas cap. The effective use of this type of secondary recovery is restricted to reservoirs that have a high flow potential so that gravity allows the oil to drain to the lower part of the formation where it may be produced. In such a situation, gas injection can be a very effective recovery method. When gravity segregation of the gas and oil cannot occur, however, or when little vertical relief exists in the formation, only a small increase over the natural pressure decline production is realized.

Water flooding

Another even more widely used secondary recovery method is water flooding. Water, after treatment to remove certain impurities, is injected through some of the wells. In moving through the formation, it pushes oil toward the remaining production wells. The wells to be used for water injection are usually spaced evenly among the producing wells so that the oil has the shortest possible paths to follow. Water has been an effective injection fluid and is not dependent upon gravity segregation. A water-flooding project often increases the oil recovery to at least twice that obtained by pressure decline methods. The major exception is oil reservoirs that are connected to an active subsurface water source and are thus subjected to water flooding without injection.

Although water flooding greatly increases the recovery of oil from a particular reservoir, it leaves one-fourth to one-third of the oil behind, and several other methods have been tried in efforts to achieve even better oil recovery. One such method uses a quantity of fluid, such as alcohol or liquefied carbon dioxide, ahead of the injected water to decrease the quantity of oil bypassed by the water. On occasion the injected water is made more viscous by the addition of jellying agents, which help to increase the reservoir volume invaded by water. Still another method requires that a small amount of liquid hydrocarbon, such as propane, be placed ahead of a gas injection to help push the oil ahead of the gas.

Several recent methods for secondary recovery involve the application of heat to the reservoir to make the oil more flowable. These methods are best suited for reservoirs that contain viscous crude oil. In one of them (in situ combustion), a portion of the oil in the formation is burned by igniting the oil around the well and then forcing air through the formation to push the burning region out into the reservoir. The heat thus generated makes the oil more flowable. In a successful project, about ten times as much oil is recovered as is burned. In another heat method, steam is injected into the reservoir; the heat liberated by the condensation decreases the viscosity of the crude oil. The heated oil is often produced from the same well into which the steam was injected. Several cycles of such injection and production may be used on a single well.

Gas cycling. Natural gas reservoirs often contain appreciable quantities of liquid hydrocarbons held in the gaseous state. Since these liquids are liquefied petroleum gases (LPG) such as propane and butane, they are important sources of both fuel and chemical feedstocks. If the pressure in such a reservoir is allowed to decline during gas production, these liquids will condense in the reservoir and become nonrecoverable. To prevent this decline in pressure, a quantity of the produced gas from which the liquids have been removed is put back into the reservoir. This process, called gas cycling, is continued until the optimum quantity of liquids has been removed. The reservoir pressure is then allowed to decline during the production of the nearly dry gas. In effect, this process defers the use of the reinjected natural gas until after the liquids have been produced.

Overall, the production of oil and gas has become an exacting engineering job. Since new oilfields are not being discovered at a rate high enough to supply the steadily increasing demand for products (see Table 2), fields already discovered must be used more efficiently.

TRANSPORTATION AND STORAGE

Petroleum is transported by every conceivable means. Among the major transportation methods are pipelines for overland movement and barge or tanker for water shipment. When a pipeline can be constructed, it is nearly always the most economical means of transportation.

Pipeline sizes

Pipeline diameters vary from the two-inch (five-centimetre) lines that are used in well-gathering systems to 48-inch (120-centimetre) lines for carrying petroleum from producing areas to large consumer centres. The pipelines, normally placed underground, are welded into continuous lengths. Since a major problem is corrosion of the metal pipe, the lines are coated with protective materials such as asphalted felt. Pumping stations are located periodically along the pipeline to help move the petroleum over long distances.

Although barges are used in inland waters, the overseas shipment of petroleum and products is made in specially constructed ships (tankers). For crude-oil transportation, these tankers vary in capacity from 100,000 barrels to 1,000,000 barrels or more. Several tanker ships have been

Table 2: World Petroleum Demand and Supply
(000 barrels per day)

	domestic demand	domestic supply			excess supply over demand	excess demand over supply	percent of world	
		crude oil	natural gasoline, etc.	total			domestic demand	domestic supply
1970								
North America	16,711	11,319	1,923	13,242	...	3,469	35.7	27.7
United States	14,716	9,632	1,680	11,312	...	3,404	31.4	23.7
Canada	1,485	1,200	177	1,377	...	108	3.2	2.9
Other North America	510	487	66	553	43	...	1.1	1.1
Latin America	2,440	4,743	95	4,838	2,398	...	5.2	10.2
Venezuela/West Indies	460	3,852	45	3,897	3,437	...	1.0	8.2
Other Latin America	1,980	891	50	941	...	1,139	4.2	2.0
Non-Communist Europe	12,241	323	224	547	...	11,694	26.1	1.2
Africa	1,183	6,203	13	6,216	5,033	...	2.5	13.0
Middle East	799	14,020	36	14,056	13,257	...	1.7	29.4
Far East, Oceania	6,369	1,396	2	1,398	...	4,971	13.6	2.9
Communist area	7,142	7,450	...	7,450	308	...	15.2	15.6
Total world	46,885	45,454	2,293	47,747	862	...	100.0	100.0
1969*								
North America	15,896	12,305	1,820	14,125	...	1,771	35.2	31.7
United States	14,137	9,238	1,590	10,828	...	3,309	31.1	28.0
Canada	1,237	1,078	168	1,246	9	...	2.9	2.7
Other North America	522	410	62	472	...	50	1.2	1.0
Latin America	2,280	4,818	90	4,908	2,628	...	5.4	10.7
Venezuela/West Indies	415	3,973	43	4,016	3,601	...	1.0	8.8
Other Latin America	1,865	845	47	892	...	973	4.4	1.9
Non-Communist Europe	11,057	339	204	543	...	10,514	26.2	1.2
Africa	1,087	4,424	12	4,436	3,349	...	2.6	9.7
Middle East	742	12,630	34	12,664	11,922	...	1.8	27.8
Far East, Oceania	5,621	1,672	1	1,673	...	3,948	13.3	3.7
Communist area	6,552	6,938	...	6,938	386	...	15.5	15.2
Total World	43,235	43,126	2,161	45,287	2,052	...	100.0	100.0

*Revised.

converted for the shipment of liquefied natural gas. This type of gas transport will undoubtedly open new markets to the natural gas industry.

Nonvolatile crude oil and its products are most often stored in cylindrical tanks set above ground. These storage tanks may be as large as 100 feet (30 metres) in diameter and 30 feet (nine metres) tall. Smaller versions are used to store crude oil at the reservoirs. Highly volatile hydrocarbons such as the liquefied petroleum gases are stored in spherical or nearly spherical tanks. Natural gas is the most difficult product to store. Liquefication requires either high pressures or extremely low temperatures. Probably the most successful storage technique is the use of an underground reservoir much like the original reservoir. The storage reservoir must have nonporous rock over and around it to prohibit the escape of the gas into nearby formations. Another relatively recent method is the storage of liquefied natural gas in underground caverns. It is believed that the low temperatures will freeze the moisture in the surrounding soil and thus prevent the escape of the gas.

BIBLIOGRAPHY. T.C. FRICK (ed.), *Petroleum Production Handbook*, 2 vol. (1962), comprehensive coverage of all phases of petroleum production technology, with diagrams, charts, and tables; F.W. COLE, *Reservoir Engineering Manual*, 2nd ed. (1969), a general treatise, including descriptions of petroleum reservoir evaluation and methods of increasing the production of oil and gas from a reservoir; B.C. CRAFT and M.F. HAWKINS, *Applied Petroleum Reservoir Engineering* (1959), and S.J. PIRSON, *Oil Reservoir Engineering*, 2nd ed. (1958), standard textbooks covering the technical side of petroleum reservoir analysis and production predictions; C.R. SMITH, *Mechanics of Secondary Oil Recovery* (1966), a textbook on the methods by which oil to be recovered may be increased by the injection of other fluids into the reservoir.

(B.H.C.)

Petroleum Refining

Crude oil as it is found in nature consists of complex mixtures of compounds containing hydrogen and carbon (hydrocarbons). In the petroleum refining process, useful products, such as gasoline, motor oil, and petrochemicals, are manufactured from crude oil which in general is unsuitable for direct use.

History. Petroleum refining began in the United States and Russia in the second half of the 19th century, following the discovery in 1859 of "rock oil" in Pennsylvania. In the earliest refineries, simple stills separated crude oil into impure gasoline, kerosene, lubricating oil, and fuel oil fractions. Kerosene, or lamp oil, was the principal marketable product. To improve its odour and appearance the kerosene was treated chemically with caustic soda or sulfuric acid.

The earliest automobile fuel was composed of those fractions of crude oil that were too light to be included in kerosene. Before the invention of the automobile this fraction had been virtually unsalable; as demand for it rose, refiners increased production and improved quality. Methods for the continuous distillation of crude oil were introduced.

After World War I a major improvement in refining came with development of the cracking process, consisting of heating surplus heavier oils under pressure and thereby cracking, or splitting, their large molecules into the smaller ones that form the lighter, more valuable fractions. After appropriate chemical treatment, gasoline manufactured by cracking performed better in automobile engines than gasoline derived from straight distillation.

During the 1930s and World War II, sophisticated refining processes involving the use of catalysts led to further improvements in the quality of fuels and increased their supply. These improved processes, including catalytic cracking, polymerization, alkylation, and isomerization (see below), permitted the petroleum industry to meet the high-performance demands of combat aircraft and, after the war, to supply the increasing quantities required by commercial aviation.

The 1950s and 1960s brought a large-scale demand for jet fuel and for high-quality lubricating oils. Catalytic reforming was established as the leading process for upgrading automotive motor gasolines for use in the higher compression engines. Hydrocracking, accomplished by the addition of hydrogen during refining, also improved the crude-oil fractions.

Classification and types of crude oils. The various hydrocarbon compounds that are mixed together in crude oil have different boiling points, but apart from the lightest, the differences between the boiling points of neighbouring members in the rising scale of molecular weight are so small (only fractions of a degree) that they cannot be separated by ordinary distillation. Fortunately, sep-

arating is not usually necessary; most common petroleum products consist of mixtures of compounds whose boiling points fall within a specified range.

Typical boiling ranges are: motor gasoline 75°–300° F (25°–150° C), kerosene 300°–450° F (150°–230° C), diesel oils 450°–650° F (230°–340° C), and fuel oils above 650° F. In practice, however, there are many different grades of each product with different boiling-point specifications.

Crude oil composition

Crude oils vary widely in boiling range and composition. Figure 1 shows the proportions of the products obtained by distillation of six typical crude oils, ranging

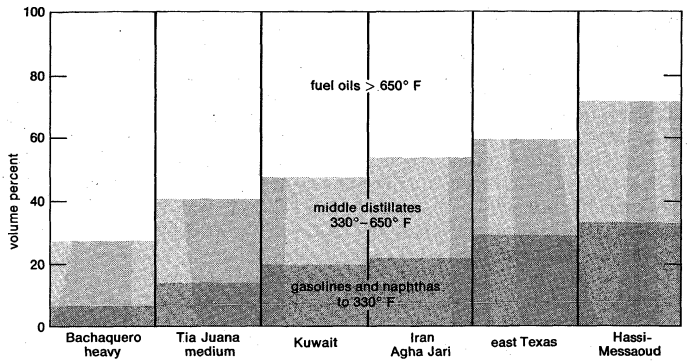


Figure 1: Proportions of the products obtained by distillation of six crude oils.

from West Venezuelan Bachaquero Heavy crude to the light Algerian crude Hassi-Messaoud. The percentage of lighter products (gasoline and naphtha) increases to the right of the chart. Middle distillates (which include kerosene and gas oil) also increase, while fuel-oil content diminishes.

Four main types of hydrocarbons are present in crude oil: normal paraffins, isoparaffins, cycloparaffins (also called naphthenes), and aromatics. Some crude oils, such as Pennsylvanian, consist mainly of paraffins. Others, such as the heavy Mexican and Venezuelan crudes, are predominantly naphthenic and are rich in asphalt (a high boiling semisolid material). Wax is usually, but not always, associated with paraffin-base crudes. The U.S. Bureau of Mines classifies crude oils as paraffin base, naphthene base, or mixed base by determining the properties of key fractions distilled from the oil.

In addition to the hydrocarbons, compounds of sulfur, nitrogen, and oxygen are present in small amounts in crude oil. These are harmful unless removed. Also, there are usually traces of vanadium, nickel, chlorine, and arsenic.

SEPARATION INTO COMPONENTS

The primary refinery process is fractional distillation, which may be followed by other physical separation methods, such as solvent extraction, in which superior lubricating stocks are extracted by means of a solvent, usually furfural obtained from oat hulls.

Fractional distillation. Modern petroleum distillation units operate continuously over long periods of time and are large compared with those that carry out the same process in other industries. Units with 100,000-barrel-per-day capacities are commonplace, and units of over 200,000 barrels per day are now in operation. The American barrel is the most widely used unit in the oil industry and comprises 42 U.S. gallons, or 35 imperial gallons (about 160 litres).

Figure 2 shows the principles of operation of a fractional distillation unit. Crude oil is pumped at a constant rate through steel alloy tubes in a furnace, fired by gas or oil, and heated to a temperature between 600° and 700° F (315°–370° C), depending on the type of crude oil and the end products desired. A mixture of vapour and unvaporized oil passes from the furnace into the fractionating column, a vertical cylindrical tower as much as 150 feet (45 metres) high, provided with 30 or 40 perforated fractionating trays spaced at regular intervals. The bubble-cap tray is the commonest type used, but the sieve tray (less expensive, but with a narrower operating range) is sometimes employed. It consists of a simple perforated plate with small holes about $\frac{3}{16}$ to $\frac{1}{4}$ inch in diameter.

The oil vapours rise up through the column and are condensed to a liquid in a water-cooled condenser at the top. A small amount of gas remains uncondensed; this is piped away to the refinery fuel-gas system. A pressure control valve on this line maintains fractionating-column pressure at the required figure, usually near atmospheric pressure.

Reflux

Part of the condensed liquid, called reflux, is pumped back into the top of the column and runs down from tray to tray, contacting the rising vapours as they pass through the slots in the bubble caps. The liquid progressively absorbs heavier constituents from the vapours and in turn loses its lighter components. Condensation and re-evaporation take place on each tray. Eventually, an equilibrium is reached in which there is a continuous gradation of temperature and of oil properties throughout the column, with the lightest constituent on the top tray and the heaviest at the bottom. The use of reflux and a column of this type distinguishes fractional distillation from simple distillation.

In the column shown in Figure 2, fractions called sidestreams are withdrawn at several points. These products have properties intermediate from the top and base of the column. Typical boiling ranges for various products are as follows: light gasoline (overhead) 75°–200° F (25°–95° C); naphtha (no. 1 sidestream) 200°–300° F (95°–150° C); kerosene (no. 2 sidestream) 300°–450° F (150°–230° C); and gas oil (no. 3 sidestream) 450°–650° F (230°–340° C).

In practice, the boiling ranges of these products can be varied within wide limits according to requirements. This is accomplished either by selecting different drawoff points in the column or by varying the quantity of oil withdrawn and has the effect of changing the equilibrium concentration of the liquid on the tray concerned.

The degree of fractionation, or sharpness of separation between hydrocarbons, depends on the number of trays and their efficiency in achieving equilibrium between vapour and liquid. It also depends on the reflux ratio; *i.e.*, the volume of liquid pumped back divided by the volume of overhead product. Reflux ratios in crude oil distillation columns are usually between 1:1 and 3:1.

Unvaporized oil entering the column (Figure 2) flows downward over another set of trays in the lower part of

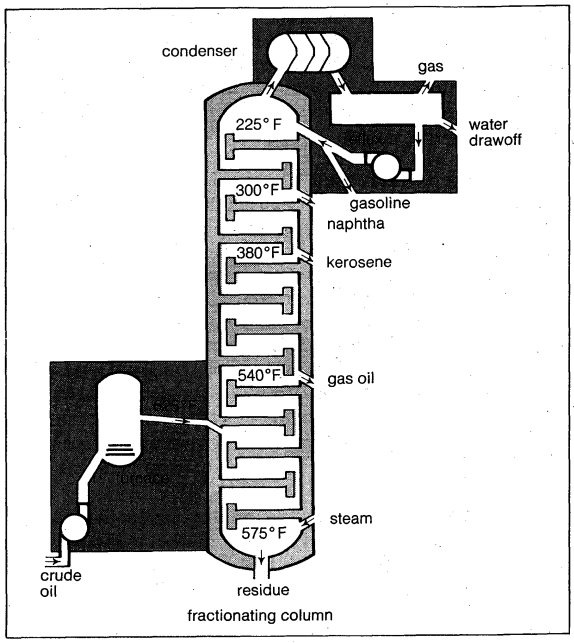


Figure 2: General operating principles of a fractional distillation unit (see text).

Treatment of unvaporized oil

the column called stripping trays, which remove any light constituents remaining in the liquid. To assist in this, steam is injected at the bottom of the column. The residue that passes from the base of the fractionating column is suitable for blending into fuel oils. Alternatively, it may be distilled a second time under vacuum conditions and further quantities of distillate recovered for use as a starting material for manufacturing lubricating oil or as feedstock for catalytic cracking.

Vacuum distillation. The principles of vacuum distillation resemble those of fractional distillation, and, except that larger diameter columns are used to maintain comparable vapour velocities at the reduced pressures, the equipment is also similar. The vacuum is produced by steam ejectors in vacuum distillation. Components that are less volatile can be distilled without raising the temperature to the range at which cracking occurs, as it would at atmospheric pressure. Firing conditions in the furnace are adjusted so that the oil temperature does not exceed about 750° F (400° C). The residue after vacuum distillation is asphalt, or bitumen.

Superfractionation. Superfractionation is an extension of fractional distillation employing columns with a much larger number of trays (*e.g.*, 100) and reflux ratios exceeding 5:1. With such equipment it is possible to obtain fractions containing only a few hydrocarbons or even to separate pure compounds. By this method, isopentane of over 90 percent purity is produced for aviation gasoline; isohexane and isoheptane concentrates are also prepared for the same purpose. These isoparaffins have much higher octane numbers than the corresponding normal paraffins.

Other methods. *Absorption and stripping.* Absorption and stripping are processes used to obtain valuable light products such as propane/propylene and butane/butylene from the gasoline vapours that pass out of the top of the fractionating tower. In the absorption process, gasoline vapours are bubbled through an absorption oil such as kerosene or heavy naphtha in equipment resembling a fractionating column. The light products dissolve in the oil while dry gases such as hydrogen, methane, ethane, and ethylene pass through undissolved. Absorption is more effective under pressures of about 100–150 pounds per square inch than it is at atmospheric pressure.

The light products are separated from the absorption oil in the stripping process. The solution of absorption oil and light products is boiled by steam and passes to a stripping column, where the light product vapours pass upward and are recovered by condensation by water-cooling under pressure. The unvaporized oil passes from the base of the column for reuse.

Solvent extraction. Solvent-extraction processes are used primarily for the removal of constituents that would have an adverse effect on the performance of the product in use. The quality of kerosene is improved by the extraction of aromatic compounds that burn with a smoky flame (these compounds can be used to advantage in tractor fuels).

Another important operation is the removal of heavy aromatic compounds from lubricating oils. Removal improves the viscosity–temperature relationship of the oil, extending the temperature range over which satisfactory lubrication is obtained. The usual solvents for extraction of lubricating oil are furfural and phenol. Other solvents are dichloroethyl ether, nitrobenzene, and a mixture of liquid propane and cresylic acid.

Adsorption. Certain highly porous, solid materials have the ability to select and adsorb specific types of molecules, thus separating them from other types. Silica gel is used in this way to separate aromatics from other hydrocarbons, and activated charcoal is used to remove liquid components from gases.

Adsorption is thus somewhat analogous to the process of absorption with an oil, although the principles are different. Layers of the absorbed material only a few molecules thick are formed on the extensive interior surface of the adsorbent; this interior surface may amount to several acres per pound of material.

Recent years have brought new developments in the use of adsorbents of a very selective nature called molecular sieves. Molecular sieves are produced by the dehydration of naturally occurring or synthetic zeolites (crystalline alkali-metal aluminosilicates). The dehydration leaves intercrystalline cavities that have pore openings of definite size, depending upon the alkali metal of the zeolite. Under adsorptive conditions normal paraffin molecules can enter the crystalline lattice and be selectively retained, whereas the other molecules are excluded This principle is used commercially for the removal of normal paraffins from gasoline fuels, thus increasing their octane number. The use of molecular sieves has also been extended to the separation of hydrocarbons of higher molecular weight.

Molecular sieves

Thermal diffusion. When a mixture of hydrocarbons is passed through a narrow gap, of the order of $\frac{1}{100}$ inch, between hot and cold surfaces, some of the constituents concentrate near the hot surface and others near the cold. The phenomenon, known as thermal diffusion, is not clearly understood, but it is believed that separation occurs as a result of differences in the shapes of the molecules. Though this process has been applied in the laboratory as an analytical tool, it is unlikely to find much use in industry as the thermodynamic efficiency is low.

Crystallization. The crystallization of wax from lubricating oil fractions is essential to make the oils suitable for use. A solvent, (for example, a mixture of benzene and methyl ethyl ketone) is first added to the oil and the solution chilled to about −5° F (−20° C). The function of the benzene is to keep the oil in solution and maintain its fluidity at low temperatures, whereas the methyl ethyl ketone acts as a wax precipitant. Rotary filters are used to filter off the wax crystals on a specially woven canvas cloth stretched over a perforated cylindrical drum. A vacuum, maintained within the drum, sucks the oil into it. The wax crystals are removed from the cloth by metal scrapers, after washing with solvent to remove traces of oil. The solvents are later distilled from the oil and reused.

ALTERATIONS OF MOLECULAR STRUCTURE

The separation processes described above are based on differences in physical properties of the components of crude oil. By chemically changing their molecular structure, it is possible to convert less valuable hydrocarbon compounds into those in demand.

The first of these conversion processes is cracking, or thermal decomposition of long-chain hydrocarbon molecules into shorter molecules with lower boiling points. For example, a paraffin molecule such as dodecane ($C_{12}H_{26}$) has such poor "antiknock" properties that it cannot be used in a modern automobile engine, but under intense heat it breaks down into shorter molecules such as paraffins (C_6H_{14};C_7H_{16}) or olefins (C_6H_{12};C_5H_{10}), which are suitable for motor fuels. The chemical reactions that take place in a cracking operation are complex.

The product derived from a cracking process is, in effect, a synthetic crude oil. Desired gasoline fractions must be separated by distillation.

Thermal cracking and reforming. The earliest cracking techniques are typified by processes in which kerosene or gas oil materials were converted by heating to temperatures between 850° and 1000° F (450°–540° C) at pressures of 250 to 500 pounds per square inch. This process produced gasolines of about 70 octane number (low by modern standards).

Visbreaking, another thermal cracking process, reduces viscosity of heavy crude oil residues to make them more suitable for inclusion in fuel oils.

The steam cracking process, by which ethylene and other olefins are made from naphtha, differs from thermal cracking in that it is carried out at low pressures and higher temperatures.

Thermal reforming, a modification of the thermal cracking process, reforms or alters the properties of low-grade components such as naphthas by converting the molecules into those of higher octane number. Pressures used are somewhat higher than in cracking. At temperatures from 950° to 1,050° F, it is possible to obtain gasolines with octane numbers of 70–80 from components of less than 40.

Catalytic cracking. The development of catalytic cracking was one of the major achievements of the petroleum industry. Enormous catalytic cracking plants dominate the skylines of modern oil refineries and operate continuously for years on end. The catalytic cracking of gas oil is one of the most important processes for manufacture of gasoline.

Use of a catalyst (a material that assists a chemical reaction but does not take part in it) in the cracking reaction increases the yield of improved quality products under much less severe operating conditions than in thermal cracking. Typical temperatures used are from 850° to 950° F at much lower pressures of 10 to 20 pounds per square inch. At first natural or synthetic clays were used as catalysts. A typical synthetic clay contains 12.5 percent alumina and 87.5 percent silica. In the last few years zeolitic or molecular sieve-base catalysts have been introduced. These give greater gasoline yields while reducing the formation of gas and coke.

Cracking may be accomplished either with a "fixed bed" of catalyst, with a "moving bed" of granular catalyst, or with a "fluid" catalyst, which is a finely divided solid having properties analogous to a liquid when agitated by air or oil vapours.

Fluid catalytic cracking process

The fluid catalytic cracking process may be examined as an example. Figure 3 shows the principles of operation of a fluid catalytic cracking unit.

From *Hydrocarbon Processing* (September, 1968)

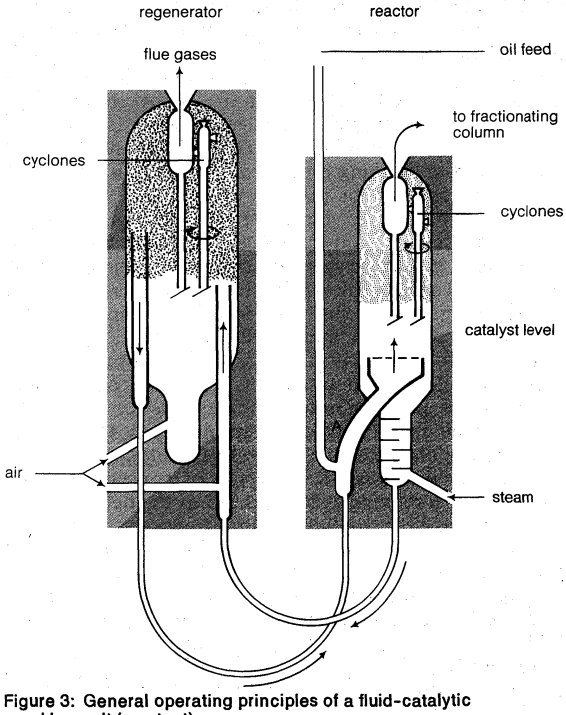


Figure 3: General operating principles of a fluid-catalytic cracking unit (see text).

In this arrangement a reactor and a regenerator vessel are located side by side. The oil is vaporized when it meets the hot catalyst at point A. The vapours force the catalyst up into the reactor, which (because its cross section is larger in area) allows the catalyst to settle into a bed whose depth can be varied to regulate the reaction time. The oil vapours maintain the bed of catalyst in a turbulent condition, thus effecting the close contact between vapours and catalyst necessary for cracking.

As the cracking reactions proceed, carbon is deposited on the catalyst particles. Since these deposits impair its efficiency, catalyst is continuously withdrawn from the base of the reactor and regenerated in the other vessel by burning off the carbon with a current of air. This also heats up the catalyst to the required reaction temperature.

Product vapours pass out of the top of the reactor through separators using centrifugal force, called cyclone separators, where the catalyst is removed and dropped back into the reactor bed. Similar cyclones recover catalyst from the flue gases that emerge from the top of the regenerator. A small amount of fresh catalyst is added to the system from time to time to maintain the activity and a similar amount withdrawn.

The products are fractionated in a distillation column. The yield of gasoline is in the region of 40–60 percent of the gas oil feed.

Catalytic reforming. This process uses catalysts to mold molecular structure into desirable forms without the formation of unwanted heavy products and coke. As in thermal reforming, a naphtha-type material is used as the feedstock, but the reactions are carried out in the presence of hydrogen, which inhibits the formation of unstable unsaturates that polymerize into high-boiling compounds.

Use of platinum, chlorine, and fluorine

Most catalytic reforming processes use platinum as the active catalyst. It is distributed on the surface of aluminum oxide as a carrier. Small amounts of chlorine and fluorine act as catalyst promoters. In spite of the high cost of platinum, the process is feasible because of the long life of the catalyst and the high quality and yield of the products obtained. This process has almost entirely replaced thermal reforming.

Other important reactions include the breaking down of long-chain hydrocarbon molecules into smaller saturated chains and the formation of isoparaffins, made up of branched-chain molecules. Formation of ring compounds, or cyclization, of paraffins into naphthenes also takes place and naphthenes are in turn dehydrogenated into aromatic compounds. The desirable end products are isoparaffins and aromatics, both with high octane numbers.

Commercially available processes include platforming, powerforming, ultraforming, and catalytic reforming. In a typical platforming unit, the naphtha charge mixed with hydrogen (the mixture is about five molecules of hydrogen to one of hydrocarbon) is heated to a temperature normally in the range of 455°–535° C. This gaseous mixture passes downward through the pellets of catalyst contained in three reactors in series. Operating pressures vary from 200 to 800 pounds per square inch. Because heat is absorbed in the reforming reactions, the mixture is reheated in intermediate heaters between the reactors.

After leaving the third reactor, the liquid product is condensed, separated, and passed to the stabilizer column, where hydrocarbon gases produced in the reactions are removed. The catalytic reformate product is then available for blending into motor gasoline without further treatment. The hydrogen leaving the products separator is returned to the reactors by the compressor.

Operating conditions are set to obtain the required octane number, usually between 90 and 100. At the higher levels, reduced yields are obtained. Octane numbers can be raised by the addition of between 1.5 and 3 millilitres of tetraethyl lead per U.S. gallon (3.8 litres). In some countries, legislation is proposed to limit the amount of tetraethyl addition to motor gasolines as an antipollution measure. Such restrictions will tend to increase the use of catalytic reforming, an important process for production of high-octane components.

Polymerization. Gaseous hydrocarbons produced in thermal and catalytic cracking processes may be used for refinery fuel, or they may be converted into more valuable products by various methods such as polymerization. In polymerization, the hydrocarbon molecules are induced to combine, or polymerize, into molecules of two or more times the molecular weight, forming a material that performs well as a motor fuel. Early commercial processes used 65 percent sulfuric acid as catalyst but were soon superseded by processes using catalysts of phosphoric acid on kieselguhr or the phosphates of copper and cadmium. High pressures, of the order of 400 to 1,200 pounds per square inch are required at temperatures from 350° to 450° F.

The polymer gasolines derived from petroleum fractions containing propylene and butylene have octane numbers above 90, and with the addition of tetraethyl lead, above 100.

Alkylation. The alkylation reaction also achieves a longer chain molecule by the combination of two small-

er molecules, one being an olefin and the other an isoparaffin (usually isobutane). During World War II, alkylation became the main process for the manufacture of isooctane, which made possible the production of large quantities of aviation gasoline.

Sulfuric acid alkylation. Concentrated sulfuric acid (98 percent) is used as the catalyst, and the process is carried out in reaction vessels of mild steel, as corrosion of this material with concentrated acid is relatively low. The reaction must be carried out at low temperatures, 35° to 45° F (2° to 7° C). Refrigeration is necessary because heat is generated by the reaction.

The octane numbers of alkylates produced range from 89 to 95. For aviation fuel, butylene fractions are usually used as feedstocks. After the addition of tetraethyl lead, octane numbers of over 100 are attainable. The sulfuric acid alkylation process was announced in 1938 and is still extensively used for production of aviation gasoline and high-octane motor gasoline.

Hydrofluoric acid alkylation. Hydrofluoric acid is used as the catalyst in this variant of the alkylation process. The chemical reactions are similar, but it is possible to use higher temperatures (between 24° and 46° C), thus avoiding the need for refrigeration. Recovery of the hydrofluoric acid is simple and it can be accomplished by distillation. Stringent safety precautions must be exercised when using this highly corrosive and toxic substance.

Isomerization. The demand for aviation gasoline became so great during World War II and after that the quantities of isobutane available for alkylation feedstock were insufficient. The deficiency was remedied by isomerization (changing molecular structure) of the more abundant normal butane into isobutane. The isomerization catalyst is aluminum chloride supported on alumina, promoted by hydrogen chloride gas.

Commercial processes have also been developed for isomerization of normal pentane and normal hexane to the corresponding isoparaffins. Here, the catalyst is usually of the platinum type. The reactions are carried out in the presence of hydrogen, as in catalytic reforming. Hydrogen is neither made nor consumed but is used to inhibit undesirable side reactions. The objective of these processes is to convert low-octane-number pentane and hexane feedstocks into high-octane-number gasoline-blending components or into highly pure isopentane or neohexane (2,2, dimethylbutane). In the latter case the isomerization product is purified by molecular sieve extraction followed by distillation. Application of this process is increasing and will be more valuable if limitations are placed on the use of tetraethyl lead in gasolines.

Hydrogenation and hydrocracking. One of the most far-reaching developments since World War II has been the use of hydrogen in refining, in part as the result of the availability of hydrogen as a by-product of catalytic reforming. But refining with hydrogen is now becoming so popular that hydrogen-manufacturing plants are being built.

One of the earliest applications was the conversion of octenes (eight-carbon-atom olefins) produced by polymerization into isooctane for manufacture of aviation gasoline, but the alkylation process has since rendered this obsolete. Today, two main types of hydrogenation exist; hydrotreatment and hydrocracking. Hydrotreatment is used to remove impurities such as sulfur, nitrogen, and oxygen (see below *Purification processes*).

Advantages of hydrocracking

Although capital costs are high, the flexibility of hydrocracking is such that it promises to become the most effective process available to the refiner for extending the use of the barrel of crude oil. This process can be used for producing gasoline or middle distillates from heavy gas oils, for converting residues into lighter oils, or for producing liquefied petroleum gases from straight-run naphthas. By cracking these in the presence of hydrogen and a catalyst, valuable products are obtained without simultaneous formation of coke and large quantities of gas. The activity of the catalyst can be maintained for long periods. Continuous regeneration is not necessary, as it is in catalytic cracking.

In the United States, the main objective has been to increase gasoline production, particularly from heavy crudes. In the Eastern Hemisphere there has been a trend to install units designed for middle-distillate production, to augment the supplies of heating oils; but in Europe, gasoline-producing units are being erected, releasing naphtha (petrochemical feedstock) for catalytic reforming.

Hydrocracking is accomplished at lower temperatures than catalytic cracking—*e.g.*, at 500° to 800° F—but at much higher pressures—1,000 to 2,000 pounds per square inch. The design and manufacture of the large, thick-walled vessels for operation under these conditions has been a considerable engineering achievement.

The catalysts used vary widely. The cracking reactions are induced by materials of the silica–alumina type. Hydrogenation catalysts such as nickel, tungsten oxide, platinum, or palladium are incorporated. Molecular sieve-type catalysts with a hydrogenation constituent are also being used.

PURIFICATION PROCESSES

Before petroleum products are marketed, certain impurities must be removed or made less obnoxious. The most common impurities are sulfur compounds, which, apart from their evil odour, are technically undesirable. In motor and aviation gasolines these compounds reduce the effectiveness of the antiknock tetraethyl lead; in diesel oils, they cause corrosion in the engine. Finally, there is increasing need in many countries to limit sulfur content of fuel oils to reduce atmospheric pollution.

Common impurities

Most crude oils contain small amounts of hydrogen sulfide; this is increased by the breakdown of heavier sulfur compounds during distillation. Though the bulk of the hydrogen sulfide passes off in the overhead gases, some remains dissolved in the gasoline. It is easily removed by alkaline solutions. The light distillates and products from cracking processes also contain more complex sulfur compounds that are not removed so readily. The most common of these are the mercaptans, which have a particularly unpleasant smell. Other undesirable constituents include oxygenated compounds such as alkyl phenols and naphthenic acids, nitrogen bases, gums and gum-forming constituents, and unstable compounds that can lead to colour formation in the product.

Processes for the purification of petroleum products are described below. Product quality is assessed by standard test methods. In addition to chemical tests for impurities, they include many physical tests designed to measure performance characteristics against relevant specifications.

Sulfuric acid treatment. The use of concentrated sulfuric acid as a refining agent is as old as the petroleum industry. It has been applied to almost every type of product. Though still important, its use has declined in recent years, following the introduction of solvent extraction and catalytic processes using hydrogen. Sulfuric acid removes sulfur compounds by solution. It also polymerizes highly reactive hydrocarbons and neutralizes nitrogen bases. Asphaltic constituents and easily oxidizable compounds are also removed.

Sulfuric acid treatment improves the colour of cracked gasolines and inhibits the formation of gum in storage. Sulfuric acid is particularly valuable for refining special boiling-point solvents and white spirits when a very low sulfur content is required. In the past it was extensively used for treatment of lubricating oils. The introduction of more modern processes, such as solvent extraction, has displaced it except for special purposes.

A serious inconvenience associated with acid treatment is the need to dispose of the acid tars that are formed as waste. An acid recovery plant is required to regenerate the sulfuric acid for reuse, both to economize in acid consumption and to prevent pollution. Most modern refineries have adopted catalytic refining processes to avoid these problems. Another disadvantage is the slight loss of useful product in acid treatment, since there is some polymerization to high-boiling compounds. These compounds are eliminated by redistillation of the product.

Sweetening. Sweetening processes convert evil-smelling and corrosive mercaptans into relatively innocuous disulfides. Mercaptans are slightly acidic compounds. Their

acidity decreases as the number of carbon atoms increases. Examples are methyl mercaptan, ethyl mercaptan, and propyl mercaptan. The lighter ones are partially removed by alkali treatment, but in sweetening reactions they are oxidized. Catalysts assist oxidation. In what is known as the doctor process of oxidation, the catalyst is a solution of sodium plumbite, made by dissolving lead oxide in caustic soda. This inexpensive process was very popular until recently, but the necessity of adding elemental sulfur to make the reactions proceed causes an increase in the total sulfur content of the oil. The doctor process is consequently being superseded by processes in which there are no undesirable side reactions and no increase in total sulfur content. One such is the copper chloride process, in which the catalyst is a suspension, or slurry, of copper chloride and fuller's earth. It is applicable to both gasolines and kerosenes. The oil is heated and then brought into contact with the slurry while being agitated by a stream of air or oxygen that oxidizes the mercaptans to disulfides. The slurry is then allowed to settle and is separated for reuse. The heater raises the temperature to a point that will keep the water formed in the reaction dissolved in the oil, so that the catalyst slurry remains at the right degree of hydration. After sweetening, the oil is water washed to remove any traces of catalyst and dried by a salt filter.

The copper chloride process

Mercaptan extraction. Simple sweetening is adequate for many purposes and is widely used, but other methods must be adopted if actual removal of sulfur compounds is required. One of these methods is mercaptan extraction. When certain chemicals called solutizers, such as potassium isobutyrate and sodium cresylate, are added to caustic soda solution, the solubility of the higher mercaptans is increased and they can be extracted from the oil. The oil is first pretreated with caustic soda to remove traces of hydrogen sulfide and alkyl phenols and then brought into contact with the solutizer solution, either in a packed column or other mixing device. The mixture is allowed to settle and the product pumped away to storage, via a water washer.

Regeneration of the solution containing sodium mercaptides can be accomplished by heating or by oxidation with air in the presence of a catalyst. In the latter case the regenerated solution is washed with an oil to remove the disulfides that are formed before being recycled to the extraction stage.

Clay treatment. Some natural clays, activated by roasting or by treatment with steam or acids, have been used for many years in the industry for removing traces of impurities. The phenomenon is similar to that described under the adsorption separation process; the clays retain the larger long-chain molecules within their highly porous structure.

Clay treatment is occasionally used to remove gum and gum-forming materials from thermally cracked and thermally reformed gasolines in the vapour phase. A more economic procedure, however, is to add small quantities of inhibitors, synthetic anti-oxidants, to the gasoline. These prevent, or greatly retard, gum formation.

Final purification of lubricating oils by clay treatment is in general use. Resins and highly coloured bodies remaining after solvent extraction are removed, and a pleasing appearance is given to the oil. The treatment may be by contact—that is, the clay is added to the oil, the mixture heated, and the clay filtered off—or by percolation, in which the heated oil is allowed to flow through a large bed of active adsorbent.

The spent clay is usually discarded, although there are processes in which the clay is regenerated by roasting. Disposal of spent clay presents a problem to refineries. Newer hydrofinishing processes are available in which mild hydrogenation achieves results similar to treatment by clay.

Hydrogen treatment. A major application of the hydrogen treatment is in the removal of sulfur compounds of all types. Hydrogen treatment can be applied to a wide range of products. The processes are known as hydrofining or hydrodesulfurization. In them, the oil is mixed with hydrogen and vaporized before passing over a catalyst such as tungsten, nickel, or a mixture of cobalt and molybdenum oxides, supported on an alumina base. Mild operating temperatures are used—*e.g.*, between 500° and 800° F—and pressures between 200 and 1,000 pounds per square inch.

The sulfur in the oil is converted to hydrogen sulfide, which is eliminated from the circulating hydrogen by absorption in a solution such as diethanolamine, which can then be heated to remove the hydrogen sulfide and reused. The hydrogen sulfide recovered is useful for manufacturing elemental sulfur of high purity.

Hydrotreatment processes are used primarily for sulfur removal from gasolines, naphthas, kerosenes, and gas oils. There is no breakdown of hydrocarbons and so the oils are virtually unchanged, except for the purification.

Molecular sieves. Molecular sieves are also used to purify petroleum products since they have a strong affinity for polar compounds such as water, carbon dioxide, hydrogen sulfide, and mercaptans.

Usually sieves with a pore size of 13 Angstrom units (13 $\times 10^{-8}$ centimetres) are used. These are prepared by the dehydration of a sodium aluminosilicate (zeolite). Fixed beds of adsorbent are used. The petroleum product is passed through for a predetermined period, depending on the amount of impurity to be removed. The adsorbed contaminants may then be eliminated from the sieves by purging with a gas stream at temperatures between 400° and 600° F. The cycle is then repeated.

PETROLEUM PRODUCTS AND THEIR USES

Gases. Refinery gas has a heating value similar to natural gas and is used to enrich supplies. It is used as a starting material for petrochemical industries. It is also converted into liquefied petroleum gas (LPG), a convenient fuel for domestic and light industrial use.

Refinery gas

Gasolines. *Motor gasoline.* A good motor gasoline must possess three characteristics. (1) The fuel must burn smoothly in the engine without detonating, called knocking. A common performance characteristic is the octane number, an expression of the engine performance of a gasoline when compared with two standard test fuels, normal heptane and isooctane. Normal heptane, a fuel of poor quality (*i.e.*, very prone to "knocking"), is assigned a value of 0 and isooctane, of very high quality, 100. Thus, when a motor gasoline is given an octane number of 90, it means that its test performance in a standard engine is equivalent to that of a mixture of 90 parts by volume of isooctane and 10 parts of normal heptane. There are two methods for carrying out the test: the Research Method, under mild conditions of temperature and running speed (120° F and 600 revolutions per minute), and the Motor Method, under more severe conditions (300° F and 900 revolutions per minute). In commercial practice, Research Method ratings are usually quoted. Motor gasolines are blended to required octane number (90 to 100 for most modern cars) from the components whose manufacture is described above. A small amount of tetraethyl lead (TEL) is usually added. (2) A good motor gasoline must start the car easily in cold weather. (3) The fuel system must not become vapour locked when temperatures are high. Volatility and vapour pressure define these latter two requirements quantitatively.

Octane number

Aviation gasoline. Piston-type aircraft engines are extremely fastidious in their demands on fuel quality. Knocking cannot be tolerated without risk of serious mechanical failure; this is true whether a weak mixture is being used, as under cruising conditions, or whether an enriched mixture is being used, as during takeoff or under maximum power.

Antiknock ratings of aviation gasolines are therefore determined under both conditions. The methods used, known respectively as the F3 and F4 methods, can rate fuels with octane numbers above 100. Aviation fuels are designated by quoting both figures. Thus a 100/130-grade fuel has an F3 rating of 100 and an F4 rating of 130. Because of the low temperatures in flight at high altitudes, a maximum freezing point of −60° C (−76° F) is specified to ensure that the fuel remains fluid.

Kerosene. Though its use as an illuminant has greatly diminished, kerosene still finds considerable use throughout the world in cooking and space heating. Newer uses are as fuel for tractors and as a turbine fuel for jet and turboprop aircraft.

Burning oil. Kerosene used as a burning oil must not give a smoky flame and there must be a minimum of charring of the wick or formation of "bloom" on lamp glasses. Standard laboratory procedures test these properties by burning the oil in special lamps. All kerosene fuels must satisfy minimum flash-point specifications, 120° F (49° C), to limit fire hazard in storage and handling.

Jet fuels. Jet fuels must burn cleanly in the aircraft engine combustion chambers. The first fuel used was a kerosene, and this type fuel still finds considerable use in civil aircraft. For military purposes, however, and to increase availability, it was necessary to extend the boiling range to include both gasolines and kerosenes.

As with aviation gasolines, freezing point is important and the fuels are freed of suspended water that might cause blockage of the fuel system with ice particles.

Gas oils and diesel oils. The main use of gas oils is to power the diesel engines in trucks, buses, railway engines, and similar vehicles. In a diesel engine, combustion is induced by the heat of compression of the air in the cylinder caused by the moving piston. Detonation, which leads to knocking in a gasoline engine, is a necessity for diesel fuel. A good diesel fuel starts to burn at several places in the cylinder soon after the oil is injected as a mist. Once the flame is started, any further fuel entering the cylinder ignites at once (see also DIESEL ENGINE).

Straight-chain hydrocarbons make the best diesel fuels. In order to have a standard reference scale, the oil is matched against blends of cetane (normal hexadecane) and alpha methylnaphthalene, the latter of which gives a very poor engine performance. High-quality diesel fuels have cetane numbers around 50, equivalent to a blend containing 50 percent cetane and 50 percent alpha methylnaphthalene.

For the slower diesel engines used in ships and in stationary engines on land, it is possible to cut deeper into the crude oil than with high-speed diesel fuels. The heavier, more viscous grades of oil are heated to permit easy pumping and to give the correct viscosity at the fuel injectors for good combustion.

Fuel oils. Fuel oil, or furnace oil, consists largely of residues from crude oil distillation. These are blended with other suitable fractions to obtain the required viscosity, the basis on which fuel oils are sold. About 70 percent of these oils are used to generate steam in power stations, ships' boilers, and industrial boilers. Their calorific value ranges from about 19,000 British thermal units per pound (Btu/lb), for the lightest fuels, to about 18,300 Btu/lb for the heaviest. (The British thermal unit is the amount of heat required to raise the temperature of one pound of water one degree Fahrenheit.) Calorific value depends on the carbon:hydrogen ratio of the fuel and also on its sulfur content, which may vary from about 1 to 4 percent, according to the source of the crude oil. The combustion properties of fuel oils of high sulfur content are satisfactory if the flue gases do not impinge on cool surfaces, which could lead to corrosion by the condensation of acidic sulfur trioxide.

With increasing attention being given to air pollution, some highly industrialized countries propose to reduce the concentration of sulfur oxides in the atmosphere by limiting the sulfur content of fuels burned. In the early 1970s the oil industry was avoiding the cost of desulfurizing this product by processing mixtures of low and high sulfur crude oils.

Fuel oils derived from waxy crudes can solidify under storage conditions; it is usual, therefore, to specify a pour point for the oil. This point (the lowest temperature at which the oil still flows) varies from about 35° F for the lighter fuel oils to about 70° F for the heavier. The heavier oils are usually heated in storage and in use. A pumpability test, measuring the viscosity–temperature relationships of the oil, gives a good indication of its performance in actual use.

Pumpability test

Lubricating oils. At one time the suitability of petroleum fractions for use as lubricants depended entirely on the sources of the crude oils from which they were prepared. Those from Pennsylvanian crudes were recognized as having very superior properties, being largely paraffinic in nature. But with the advent of solvent extraction the choice of suitable crude oils has been considerably extended.

Viscosity is the basic property by which lubricating oils are classified. Requirements vary from the thin oil needed for the high-speed spindles of textile machinery to the viscous, tacky compositions applied to open gears or wire ropes. Between these extremes is a range of products, many highly refined, with special characteristics (see also LUBRICATION).

Automobile oils represent a large proportion of those manufactured. In the U.S., the Society of Automotive Engineers established a system of SAE numbers to indicate the viscosity at a particular temperature, 0° F (−18° C). Oils in common use have SAE numbers varying from 5 to 50.

When ordinary mineral oils having satisfactory fluidity at low temperatures are used over an extended temperature range, excessive thinning occurs at the high temperatures. To correct this, multigrade oils incorporating synthetic long-chain polymers have been developed. These have much more satisfactory viscosity–temperature relationships. Thus an oil designated SAE 10W/40 has the viscosity of an SAE 10W oil at 0° F (−18° C), and of an SAE 40 oil at 210° F (99° C). Such an oil will help start an engine in winter (hence the suffix W) and will lubricate well under running conditions in summer. Other additives that improve the performance of lubricating oils are anti-oxidants and detergents, which maintain engine cleanliness by keeping fine carbon particles in suspension in the oil.

Gear oils. In gear lubrication the oil separates metal surfaces, reducing friction and wear. Extreme pressures develop in some gears, notably those in the rear axles of cars; special additives are incorporated to prevent freezing together (seizing) of the metal surfaces. The oils contain sulfur compounds that form a resistant film on the surfaces, preventing actual metal-to-metal contact. Such compounds are known as extreme pressure (EP) additives.

Greases. Greases are lubricating oils to which thickeners, or gelling agents, are added. Soaps of aluminum, calcium, lithium, and sodium are commonly used, while more recently nonsoap thickeners (*e.g.*, carbon, silica, and polyethylene) have been employed.

Other petroleum products. Highly purified distillate fuels are used in pottery kilns, for glass manufacture, and many other purposes, including the baking of bread. Metallurgical operations employing petroleum products include the smelting of ores, forging, and alloy melting.

White spirits are used as thinners for paints and varnishes and for dry cleaning. They are highly purified to avoid unpleasant odours. Similar materials are used in the rubber industry and for the extraction of vegetable oils.

Waxes and petroleum jellies are familiar products. Asphaltic bitumen is the earliest known petroleum product, having been used by the pre-Babylonian inhabitants of the Euphrates valley before 2500 BC. Today it is widely used for the construction of roads and airfields. Specialized uses of bitumen include the manufacture of roofing felts, waterproof papers, pipeline coatings, and electrical insulation.

Asphaltic bitumen

Carbon black was previously made by the incomplete combustion of natural gas. Today it is manufactured by decomposing liquid hydrocarbon fractions; it is compounded with rubber in tire manufacture and is a constituent of printing inks and lacquers.

Petroleum chemicals. By definition, petroleum chemicals are simply chemicals that happen to be made from a starting material obtained from petroleum. They are, in almost every case, virtually identical with the same product from other sources, such as coal and coke or fermentation processes.

Olefins. The thermal cracking process in refineries, as

it developed from 1910 onwards, aimed at increasing the quantity and improving the quality of gasoline. As a by-product of this process, gases were produced that included a significant proportion of lower molecular weight olefins, particularly propylene and *n*-butylenes (olefins are unsaturated compounds of carbon and hydrogen and are chemically very reactive). Commercial production of chemicals from this source dates from 1919. Today catalytic cracking is a valuable source of propylene and *n*-butylenes but not a very good one of ethylene, a most important petroleum chemical.

It is important to differentiate between the current practices of the U.S. and most of the rest of the world in producing olefins. In the U.S., propylene and *n*-butylenes are produced from refinery catalytic cracking. Ethylene is largely produced in the U.S. by cracking ethane extracted from natural gas. In Europe, Japan, and other major centres of petroleum chemical production, catalytic cracking is a less important refinery operation, and appropriate supplies of natural gas are not readily available. As a result, Europeans (now also Japanese) generally crack a liquid hydrocarbon fraction in the gasoline range (light or heavy naphtha). In a single operation this yields a whole range of olefins, notably ethylene, propylene, the *n*-butylenes, and butadiene.

Cracking plants are usually situated adjacent to a refinery from which they derive the fractions used as feedstock and to which they return liquid cracking product.

Petroleum chemical plant complexes

Plants that consume these various olefins are clustered around the cracker. Among typical plant complexes in Europe are ICI at Wilton, England; Shell at Carrington, England; BP Chemicals at Grangemouth, Scotland; Shell and Naphtachimie near Marseilles, France; Shell, Dow, Gulf, ICI, and other companies near Rotterdam, The Netherlands; Shell/BASF at Wesseling, Germany; and BP/Bayer at Dormagen, Germany. Similar major industrial sites are characteristic of the Japanese petroleum chemical industry (for example at Yokkaichi and Tokuyama).

Petroleum has become the favoured source of organic chemicals. Over 80 percent of the world's organic chemicals are now made from petroleum; and the proportion is expected to rise to about 98 percent within the next 15 years.

Inorganic chemicals. Various inorganic chemicals are also made from petroleum, some, particularly ammonia and sulfur, in large quantities. Ammonia production requires hydrogen from a hydrocarbon source. Traditionally, this hydrogen was produced from a coke and steam reaction, but today most ammonia is synthesized from liquid petroleum fractions, natural gas, or refinery gases. It has been pointed out that sulfur is an unwanted impurity in many petroleum products. In the process of eliminating it from oil products, it is ultimately recoverable either as elemental sulfur or as sulfuric acid. Though various inorganic chemicals manufactured from petroleum represent a substantial tonnage, the total is still only a modest proportion of the world production of such chemicals.

Aromatic compounds. The aromatic compounds, prominent in the refinery process of catalytic reforming, are major sources of petroleum chemical products as well as the lower olefins mentioned above. In the traditional chemical industry, aromatics such as benzene, toluene, and the xylenes were made from coal during the course of coal carbonization in making coke or town gas. Today a much larger volume of these compounds is made in petroleum refineries by catalytic reforming. A further source of supply now increasingly used in Europe is the aromatic-rich liquid fraction produced in the cracking of naphtha to make ethylene and other olefins.

Polymers. A highly significant proportion of products made from these basic chemicals are the plastics, synthetic rubbers, and synthetic fibres known as high polymers because their molecules are the high-molecular-weight compounds made up of the repeated structural units that have combined chemically, or polymerized. In plastics the four major products are polyethylene, polyvinyl chloride, and polystyrene, all made from the simple structural unit, or monomer, ethylene, and polypropylene made from the monomer propylene. Major raw materials for synthetic rubbers include butadiene, ethylene, benzene, and propylene.

Among synthetic fibres, nylon still represents the largest tonnage; in it, the most important raw material is benzene. The second major synthetic fibre is represented by the polyesters, which are a combination of ethylene glycol (made from ethylene) and terephthalic acid (made from one of the xylenes). The third major synthetic fibre is represented by the acrylics, in which the major raw material is the propylene derivative acrylonitrile (see also FIBRES, MAN-MADE).

By-products. Petroleum chemicals have advanced from the original concept of by-product usage and represent a large investment of money and technology. An additional group of products, however, that represent a direct by-product of oil-refining operations, the naphthenic acids, can be abstracted for use in paint dryers and other derivatives. Naphthasulfonates recovered in sulfuric acid refining of oil products serve as emulsifying agents. Many cresylic acid products can be recovered as by-products.

Table 1: World Refining Capacities (1969)

country or area	barrels per calendar day (000)	percent share	percent average annual increase 1964–69
U.S.	11,960	25.6	3.8
Canada	1,210	2.6	3.5
Mexico	530	1.1	3.0
Total North America	13,700	29.3	3.8
Venezuela	1,280	2.7	3.5
Trinidad	420	0.9	4.3
Netherlands West Indies	830	1.8	5.0
Others (Caribbean area)	700	1.5	11.5
Total Caribbean area	3,230	6.9	5.5
Argentina	440	1.0	3.3
Brazil	460	1.0	9.3
Others (South America)	280	0.6	8.3
Total South America	1,180	2.6	6.5
Total Western Hemisphere	18,110	38.8	4.3
Belgium	640	1.4	15.8
France	2,180	4.7	12.0
West Germany	2,310	4.9	9.8
The Netherlands	1,150	2.5	15.8
Italy	2,670	5.7	13.8
U.K.	2,170	4.6	10.3
Others (Western Europe)	2,180	4.7	19.5
Total Western Europe	13,300	28.5	13.0
Aden	160	0.3	3.3
Bahrein	200	0.4	...
Iran	630	1.4	4.8
Saudi Arabia	380	0.8	12.5
Kuwait	440	0.9	6.3
Neutral zone	80	0.2	9.0
Others (Middle East)	330	0.7	8.0
Total Middle East	2,220	4.7	6.0
Africa	710	1.5	7.3
South Asia	630	1.3	17.5
Southeast Asia	1,080	2.3	11.8
Japan	2,870	6.2	14.3
Australasia	600	1.3	10.5
U.S.S.R., Eastern Europe, China	7,200	15.4	8.8
Total Eastern Hemisphere	28,610	61.2	11.3
World	46,720	100.0	8.3

Source: *BP Statistical Review of the World Oil Industry* (1969).

REFINERY PLANT AND FACILITIES

Refining facilities. Integration of the separate processing plants in a modern refinery has resulted in economies of services, intermediate tankage, and manpower. The simplest refinery contains a distillation unit, catalytic reformer, and various treatment units. Small refineries of this type supply local markets in many parts of the world. With such units, it is possible to produce motor gasolines

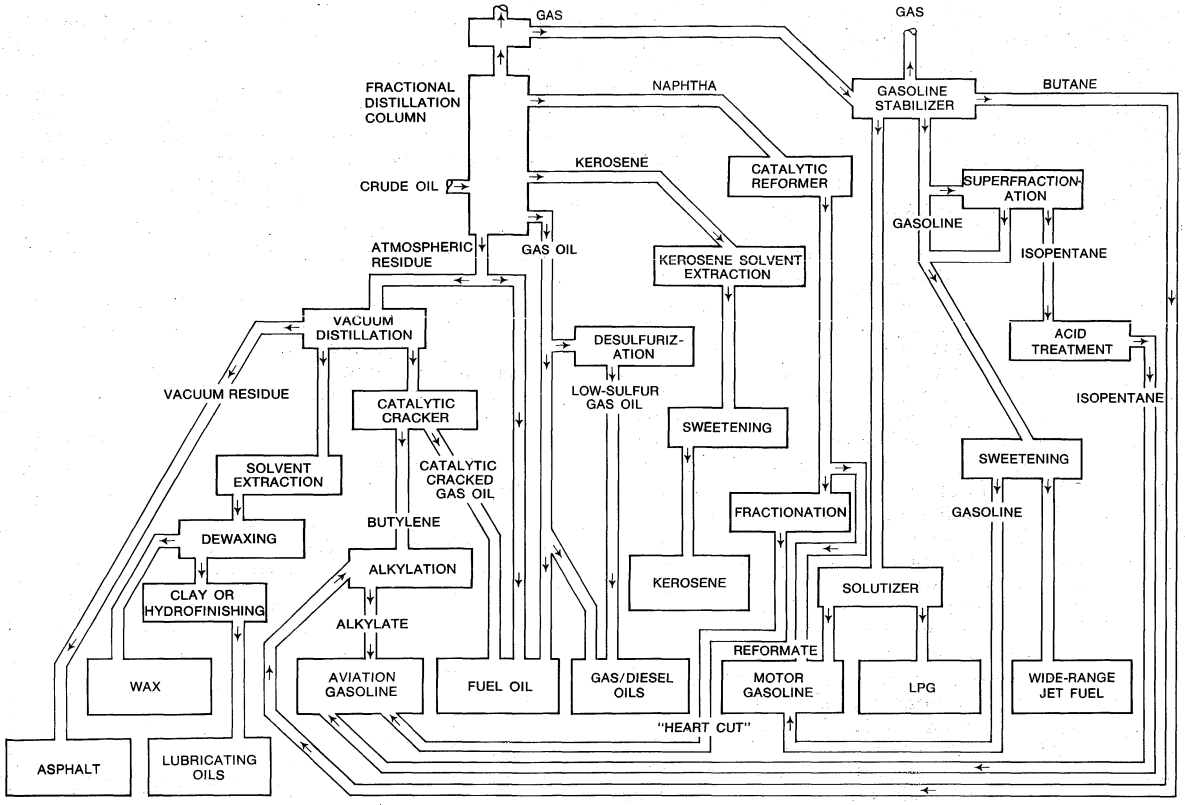


Figure 4: Generalized flow chart of the refinery process.

of different octane numbers, lamp kerosene, jet fuel, gas and diesel oil, and residual fuels. There may also be a plant for recovery of propane and butane.

In deciding which process to use, refinery managers analyze the market for residual fuel oil, which may amount to 40 or 50 percent of the crude oil processed. In some countries of the Eastern Hemisphere this fuel oil is easily marketed. But in the U.S. and Canada, where large quantities of gasoline are required, less fuel oil is used because natural gas is available within reach of consuming areas. Cracking plants, usually catalytic, together with vacuum-distillation units to produce fractions for use as feedstock, solve this problem. The production of gasolines can then be further augmented by polymerization or alkylation, superfractionation, and possibly by hydrocracking.

Effect of marketing research on processing

With further additions of plant it is possible to build a fully integrated, comprehensive refinery by stages, including solvent extraction (of kerosene), hydrodesulfurization, lubricating-oil plant (propane de-asphalting, solvent extraction, dewaxing, and clay treatment or hydrofining), and plant for the production of asphalt, wax, sulfur, and special products. The chart in Figure 4 shows how these plants are interrelated.

Tankage. The manufacturing units form only part of a refinery, though the most important one. A large proportion of the cost, and the area covered, is taken up by tankage, handling facilities, etc.

The storage capacity allocated to crude oil may be minimal if the refinery is connected to the oilfields by pipeline. When crude oil has to be imported, however, sufficient tankage is provided to enable continuous refinery operation while still allowing for the irregular arrival of tankers. Scheduling of tanker movements is important in integrated oil operations, particularly with the very large tankers (up to 250,000 tons and more) now in service.

Finished products require much more tankage in terms of storage capacity than crude oil. Provision is made for short-term variations in the sales of refinery products to avoid frequent alterations in the refinery process units. Nonvolatile products such as diesel oils and fuel oils are stored in large diameter cylindrical tanks with low-pitched conical roofs. Tanks with floating roofs are used to reduce evaporation losses in gasolines and other volatile products, including crude oils. The roof, which resembles a pontoon, moves up and down inside the tank, thus eliminating the air space that could contain a considerable proportion of petroleum vapour. Fire hazards are also reduced by the use of floating roof tanks.

For liquefied petroleum gases, pressure vessels (usually spherical) are used. Alternatively, these materials may be stored in suitable geological formations underground, such as cavities in salt domes. Large refrigerated tanks are a recent development. In these, the liquefied gas is stored at normal pressure but at reduced temperatures.

Bulk transportation. The supertankers have sharply cut the cost of transporting crude oil, making it practical to locate refineries near areas of consumption rather than adjacent to oilfields. Deepwater ports have been built to receive these large tankers, particularly in Europe, which obtains most of its supplies from the Middle East and Africa. Examples are Rotterdam, Gothenburg, Le Havre, Foss near Marseilles, and Milford Haven in the United Kingdom. Groups of refineries have been constructed near these ports or connected to them by pipeline.

If pipelines are not practical or economic, water transport is the first choice for petroleum products. Full-size tankers or smaller coastal vessels, with capacities of 15,000 barrels or less, distribute the different grades of oil to depots situated on the coastline.

Countries having networks of canals and navigable rivers afford many opportunities for using barges, a very cheap method of transportation. On the Mississippi, barges of over 5,000-tons (37,000-barrels) capacity are in use; the Rhine and the Seine are also suitable for large barges. Many canals, however, can only accommodate 200-ton barges (1,500 barrels). Each barge is divided into separate calibrated compartments so that more than one type of product can be carried on a voyage; pumps are fitted for the discharge of the cargo.

Transport by rail tank car is still widely used. The cars normally have capacities of up to 50 tons (390 barrels), but there is a trend to larger sizes.

The final stage of delivery to the majority of customers is by the familiar road tank trucks, whose carrying capacity is about 150 to 200 barrels.

Table 2: World Oil Consumption (1969)

country or area	barrels per day (000)	percent share	percent average annual increase 1964–69
U.S.	13,810	32	5.0
Canada	1,440	3	6.0
Mexico	420	1	6.3
Caribbean	980	2	6.0
Other Western Hemisphere	1,200	3	6.8
Total Western Hemisphere	17,850	41	5.3
Belgium and Luxembourg	490	1	10.8
The Netherlands	660	2	8.3
France	1,640	4	11.0
West Germany	2,430	6	11.3
Italy	1,540	4	10.3
U.K.	1,960	5	7.5
Scandinavia	980	2	9.8
Other Western Europe	1,640	4	12.5
Total Western Europe	11,340	28	10.3
Middle East	870	2	5.3
Africa	780	2	7.3
South Asia	550	1	10.5
South East Asia	1,030	2	16.0
Japan	3,290	8	17.3
Australasia	560	1	8.0
U.S.S.R., Eastern Europe, China	6,280	15	8.5
Total Eastern Hemisphere	24,700	59	10.5
World	42,550	100	8.0

Source: *BP Statistical Review of the World Oil Industry* (1969).

Pipelines and pumping stations. *Pipelines.* Development of crude-oil trunk lines began in the United States in the early years of the 20th century when refineries in the populous eastern area had insufficient supplies from convenient oilfields. Pipelines were used for the transportation of crude oil in the United States in the late 19th century, supplying refineries in the general areas of the known oilfields. After 1900 there was a rapid development of crude-oil trunk lines following the discoveries of oil in Texas, Kansas, and Oklahoma. These fields were used to supplement the declining production of the eastern oilfields. By 1940 more than 85 percent of the crude oil supplied to eastern refineries came from west of the Mississippi river. In 1970 there were more than 200,000 miles of pipeline in operation in North America, of which 70,000 miles are crude oil trunk lines and a similar amount smaller feeder lines in the oilfield areas. A major pipeline project was planned for the 1970s to bring crude oil from the newly discovered fields on the North Slope of Alaska. Europe also has several crude oil pipelines supplying inland refineries. In the Middle East, large pipelines have been built for transporting crude oil from the oilfields in Iraq and Saudi Arabia to deepwater export terminals in the eastern Mediterranean.

Crude oil and product pipelines

More and more large pipelines are coming into operation to pump batches of different products successively down the line, since it is possible to do this without intermingling them. The "Big Inch" line, completed in 1943 to transport crude oil from Texas and Louisiana to refineries in the eastern United States, was converted for carrying gas from Texas to New Jersey, a distance of 1,475 miles (2,375 kilometres). In 1964, it was replaced by the world's largest products line, which can transport 1,000,000 barrels per day of products from Houston, Texas, to New Jersey. In Europe, the first product line was the TRAPIL line from the lower Seine region to Paris, completed in 1953. This and other such lines on the Continent and in the United Kingdom enable inland refineries to dispose of products economically when the local market pattern does not match the proportions manufactured.

Centrifugal pumps are almost invariably used for mainline pumping, and with long lines the capacity can be increased by installing booster stations at intermediate points along the line. In recent years, many of these have been designed for unattended operation, the starting of pumps and the opening or closing of valves being carried out by automatic sequence control initiated by signals sent over telephone lines or by radio or by instruments responding to line pressure at the unattended pump station.

Service plant and facilities. Refineries today contain much equipment for servicing refinery operations. This includes the plant for steam generation and in some cases electric power, water treatment, effluent systems, and antipollution facilities. This equipment is essential for the operation of the industry within the community. One of the advantages of the self-contained refinery is its flexibility regarding surplus products; usually gas and heavy residue can be used for the generation of steam and power and for combustion in the oil-plant furnaces.

Table 3: Demand for Major Products (1969)
(000 barrels per day)

country	gasolines	middle distillates	fuel oils
U.S.	6,140	3,420	1,755
Canada	510	440	260
Belgium and Luxembourg	75	145	150
The Netherlands	100	155	155
France	310	655	340
West Germany	480	990	420
Italy	305	220	650
U.K.	570	425	620
Sweden	80	175	205
Japan	665	540	1,465
Australia	170	100	90

Source: *BP Statistical Review of the World Oil Industry* (1969).

BIBLIOGRAPHY. The literature on petroleum refining is extensive; each of the publications listed gives further references for study. W.L. NELSON, *Petroleum Refinery Engineering*, 4th ed. (1958), an authoritative textbook on all aspects of the industry; R.J. HENGSTEBECK, *Petroleum Processing* (1959), a highly readable account; A.E. DUNSTAN *et al.* (eds.), *The Science of Petroleum*, 6 vol. (1938–53), an encyclopaedic work prepared by worldwide authorities; W.A. GRUSE and D.R. STEVENS, *Chemical Technology of Petroleum*, 3rd ed. (1960); INSTITUTE OF PETROLEUM, *Modern Petroleum Technology*, 3rd ed. (1962); THE BRITISH PETROLEUM COMPANY LTD., *Our Industry Petroleum*, 4th ed. (1970); and SHELL INTERNATIONAL PETROLEUM COMPANY LTD., *The Petroleum Handbook*, 5th ed. (1966). A valuable reference work is the *Refining Processes Handbook* (pub. at two-yearly intervals by the journal *Hydrocarbon Processing*), containing current technical details and flow diagrams of commercial processes available to the petroleum refiner.

Books treating specific aspects of petroleum refining include: C.S. ROBINSON and E.R. GILLILAND, *Elements of Fractional Distillation*, 4th ed. rev. (1950); and V.A. KALICHEVSKY and K.A. KOBE, *Petroleum Refining with Chemicals* (1956). K.A. KOBE and J.J. MCKETTA, JR. (eds.), *Advances in Petroleum Chemistry and Refining*, 10 vol. (1958–65), contains articles useful for more detailed insight into specific topics.

The utilization of petroleum products is treated in the following works: H. ABRAHAM, *Asphalts and Allied Substances*, 5th ed., 2 vol. (1945); L.C. DENNY, L.L. LUXON, and B.E. HALL, *The Handbook of Butane-Propane Gases*, 4th ed. (1962); W. FRANCIS, *Fuels and Fuel Technology* (1965); V.B. GUTHRIE (ed.), *Petroleum Products Handbook* (1960); and H.H. ZUIDEMA, *The Performance of Lubricating Oils* (1952). The manufacture and utilization of chemicals derived from petroleum is extensively discussed in R.F. GOLDSTEIN and A.L. WADDAMS, *The Petroleum Chemicals Industry*, 3rd ed. (1967). The transportation of petroleum is treated in the publication of the ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Pipelines in the United States and Europe and their Legal and Regulatory Aspects* (1969).

Techniques in petroleum refining are advancing continuously; accounts of recent developments appear regularly in these technical journals: *Hydrocarbon Processing—Petroleum Refiner, Oil and Gas Journal, Industrial and Engineering Chemistry, Petroleum Engineer, Petroleum Press Service, The Petroleum Times,* and *World Petroleum Report.*

(J.E.C./A.L.W.)

Petronius Arbiter

The reputed author of the *Satyricon,* the first European novel, Petronius Arbiter was a close friend of the sybarit-

ic Roman emperor Nero. Described by the historian Tacitus as a complex man who, although possessing serious talents, devoted almost his entire life to the pursuit of pleasure, Petronius in the *Satyricon* presented a consummately hedonistic approach toward life that was common to his age.

The most complete and the most authentic account of Petronius' life appears in Tacitus' *Annals*, an account that may be supplemented, with caution, from other sources. It is probable that Petronius' correct name was Titus Petronius Niger. From his high position in Roman society, it may be assumed that he was wealthy; he belonged to a noble family and was therefore, by Roman standards, a man from whom solid achievements might have been expected. Tacitus' account, however, shows that he belonged to a class of pleasure-seekers attacked by the Stoic philosopher Seneca, men who "turned night into day"; where others won reputation by effort, Petronius did so by idleness. On the rare occasions, however, when he was appointed to official positions, he showed himself energetic and fully equal to public responsibilities. He is known to have served as governor of the Asian province of Bithynia and later in his career, probably in AD 62 or 63, to have held the high office of consul, or first magistrate of Rome.

Associate of Nero as "director of elegance"

After his term as consul, Petronius was received by Nero into his most intimate circle as his "director of elegance" (*arbiter elegantiae*), whose word on all matters of taste was law. It is from this title that the epithet "Arbiter" was attached to his name. Petronius' association with Nero fell within the Emperor's later years, when he had embarked on a career of reckless extravagance that shocked public opinion almost more than the actual crimes of which he was guilty. A chance remark by the biographer Plutarch suggests that Petronius in particular encouraged Nero's spendthrift schemes. That Nero, who may be discerned through the accounts of Tacitus and the historian Suetonius as a man with some artistic talent and independent pretensions to taste, should have chosen the stylish Petronius as his aesthetic adviser is some evidence of his discrimination. What Petronius thought of his patron may be indicated by his treatment of the rich vulgarian Trimalchio in the *Satyricon*. Trimalchio is a composite figure, but there are detailed correspondences between him and Nero that cannot, given the contemporary nature of the work, be accidental and that strongly suggest that Petronius was sneering at the Emperor.

Petronius' death

Tacitus records that Nero's friendship ultimately brought on Petronius the enmity of the commander of the Emperor's guard, Tigellinus, who in AD 66 denounced him as having been implicated in a conspiracy of the previous year to assassinate Nero and place a rival on the imperial throne. Petronius, though innocent, was arrested at Cumae in southern Italy; he did not wait for the inevitable sentence but made his own preparations for death. Slitting his veins and then bandaging them again in order to delay his death, he passed the remaining hours of his life conversing with his friends on trivial topics, listening to light music and poetry, rewarding or punishing his slaves, feasting, and finally sleeping "so that his death, though forced upon him, should seem natural." In a special codicil to his will, instead of the customary legacies to the Emperor and his favourites for the protection of the accused person's dependents, he composed a detailed catalog of Nero's sexual perversions, even going so far as to include a list of the Emperor's partners. When it was completed Petronius dispatched this scurrilous catalog, under his seal, to the Emperor, afterward destroying the seal he had used in order to ensure that it could not be used to incriminate others.

The *Satyricon*

The independence and the sometimes cynical opportunism of Petronius revealed in Tacitus' account are in keeping with the general character of the *Satyricon*. That Petronius did in fact write the book that is transmitted under his name is generally accepted by scholars, though the attribution rests mainly on circumstantial arguments. What insights it allows into the character and views of its author is another and a debatable question. The *Satyricon* is a comic, picaresque novel, related to several ancient literary genres. In style it ranges between the highly realistic and the self-consciously literary, and its form is episodic.

As with most works of creative literature, it is hazardous to attempt to identify any of the views expressed in it as those of the author, even when in his role of satirist he appears to address himself directly to the reader, especially since much of the work consists of fragments that have come down to us out of context. The verses in which the writer defends what he calls his "work of fresh candour" recall Tacitus' comment that Petronius' dissolute sayings and doings impressed his contemporaries "by their appearance of candour." Similarly the persistent hostility toward the more austere kinds of philosophy, which sometimes takes the form of parodies of Seneca or of commonplaces that he particularly favoured, reminds the reader of Tacitus' description of Petronius' "anti-philosophical" deathbed and of his rejection of disquisitions on the immortality of the soul in favour of light entertainment. The poem quoted above ends with an appeal to the authority of the philosopher Epicurus for pleasure to be the guiding principle of life; but in the *Satyricon* as a whole it is impossible to detect any consistent philosophical or moral position, Epicurean or other. Insofar as any moral attitude at all is perceptible, it is a trivial and debased brand of hedonism, which in this book is typical of the age and the culture from which the work sprang. The aim of the *Satyricon* was evidently above all to entertain by portraying certain aspects of contemporary society; modern readers who find the portrayal satirical in any deep and consistent sense are responding to the portrayal of corruption itself rather than to the author's views and intentions.

Considered as a portrait of society, the book is of immense value: superficial details of the speech, behaviour, appearance, and surroundings of the characters are exactly observed and vividly communicated. Of any real interest in human beings, comparable with that displayed by a great social satirist such as Charles Dickens, there is little or no sign. The wealth of specific allusions to persons and events of Nero's time shows that the work was aimed at a contemporary audience, and certain features suggest that the audience in fact consisted of Nero and his courtiers. The realistic descriptions of low life recall the Emperor's relish for slumming expeditions; and the combination of literary sophistication with polished obscenity is consistent with the wish to titillate the jaded palates of a debauched court.

If Petronius' book has a message, it is aesthetic rather than moral. The emphasis throughout the account of Trimalchio's dinner party, the longest and most elaborate episode in the work as it has survived, is on the contrast between taste and tastelessness. Stylistically, too, the *Satyricon* is what Tacitus' account of the author would lead one to expect. The language of the narrative and the educated speakers is pure, easy, and elegant, and the wit of the best comic passages is brilliant; but the general impression, even when allowance is made for the fragmentary state of the text, is that of a book written quickly and somewhat carelessly by a writer who would not take the necessary trouble to discipline his astonishing powers of invention. In his book, as in his life, Petronius achieved fame by indolence.

MAJOR WORKS

The *Satyricon* (1st century AD; *The Satyricon of Titus Petronius Arbiter*, "Loeb Series," 1913; "Penguin Classics Series," 1965), containing the episode of the "Cena Trimalchionis" (*Dinner at Trimalchio's*, 1925 and 1950) and the anecdote known as the *Widow of Ephesus*.

BIBLIOGRAPHY. The ancient sources for Petronius' life are TACITUS, *Annals xvi.* 18–19; PLINY THE ELDER, *Natural History* xxxvii. 20; and PLUTARCH, *Moralia* 60 d-e. On the identification with T. Petronius Niger, see K.F.C. ROSE, "The Author of the Satyricon," *Latomus*, 20:821–825; (1961); further discussion of the problems of identification and authorship, with bibliographies of the older literature, is in J.P. SULLIVAN, *The Satyricon of Petronius: A Literary Study* (1968); and P.G. WALSH, *The Roman Novel* (1970). There are biographical reconstructions by the Polish novelist,

H. SIENKIEWICZ in his *Quo Vadis* (1896), a highly idealized portrait; and by G. BAGNANI in his *Arbiter of Elegance: A Study of the Life and Works of C. Petronius* (1954), relying heavily on conjecture. The attempt of O. RAITH, in his *Petronius ein Epikureer* (1963), to demonstrate that Petronius was a professed Epicurean has met with general dissent.

(E.J.Ke.)

Pharmaceuticals, Production of

The World Health Organization (WHO) defines a drug or pharmaceutical preparation as

> any substance or mixture of substances manufactured, sold, offered for sale, or represented for use in . . . the diagnosis, treatment, mitigation, or prevention of disease, abnormal physical state or the symptoms thereof in man or animal; [and for use in] . . . restoring, correcting, or modifying organic functions in man or animal.

The same organization defines a pharmaceutical specialty as "a simple or compound drug ready for use, and placed on the market under a special name or in a characteristic form."

The modern pharmaceutical industry began in the 19th century with the discovery of highly active medicinal compounds that could most efficiently be manufactured on a large scale. As these compounds replaced herbal medicines of earlier times, the occurrence and severity of such diseases as pernicious anemia, rheumatic fever, typhoid fever, lobar pneumonia, poliomyelitis, syphilis, and tuberculosis were greatly reduced. Pharmaceutical industry research has greatly aided medical progress; of the 66 most valuable drugs introduced since aspirin in 1899, 57 were discovered and then produced in industrial laboratories.

History. The earliest records of medicinal plants and minerals are those of the ancient Chinese, Hindu, and Mediterranean civilizations. In 2735 BC, the Chinese emperor Shen Nung wrote an herbal in which he described the antifever capabilities of a substance known as Ch'ang Shang, since shown to contain antimalarial alkaloids. The school of alchemy that flourished in Alexandria in the 2nd century BC could prepare white lead (lead carbonate) from litharge, arsenic from realgar (arsenic disulfide), and mercury by roasting cinnabar (mercuric sulfide) in a current of air. It is recorded in *De materia medica*, a book of the 1st century BC, that verdigris (basic cupric acetate) and cupric sulfate were prescribed as medicinal agents. Remarkably, cupric sulfate is still used in medicine today.

Many crude drugs still used now, such as ipecacuanha root (ipecac), were known and employed by the ancients. The Egyptians treated constipation with senna pods and castor oil and employed caraway and peppermint to relieve indigestion. The Greek physician Galen (*c.* AD 130–*c.* 200) included hyoscyamus, opium, squill (a plant drug used as an expectorant, cardiac stimulant, and diuretic), and viper toxin among other drugs in his apothecary shop. He also insisted on purity in drugs; *i.e.*, the right variety and age of the botanical specimens.

Pharmaceutical practice improved markedly in the 16th and 17th centuries. In 1546 the first pharmacopoeia, or collected list of drugs and medicinal chemicals with directions for making preparations from them, appeared in Nuremberg (Nürnberg, West Germany). Previous to this time, medical preparations had varied in concentration and even in constituents. Other pharmacopoeias followed, in Basel (1561), Augsburg (1564), and London (1618). Despite its name, the *London Pharmacopoeia* was mandatory for the whole of England and thus was the first example of a national pharmacopoeia.

The first apothecaries

In 1617 the Society of Apothecaries, London, was founded, marking the emergence of pharmacy as a distinct profession. The separation of apothecaries from grocers was authorized by King James I; only a member of the society could keep an apothecary's shop and make or sell pharmaceutical preparations. In 1841 the Pharmaceutical Society of Great Britain, London, was founded, placing the education and training of the pharmacist upon a proper scientific basis. Today, pharmaceutical practice and education are carefully supervised throughout the world.

Known as "Hoffman's drops," ether was first employed as an anesthetic in 1842; chloroform followed soon afterward in 1847. Alkaloid compounds were also isolated from plant sources during this period. Narcotine was obtained from opium by a French pharmacist in 1803 and was followed by morphine in 1806, emetine and strychnine (around 1817), brucine and piperine (1819), colchicine and quinine (1820), nicotine (1828), atropine (1833), cocaine (1860), and physostigmine (1867). Isolation of these potent compounds was a milestone in pharmaceutical progress for three reasons. First, accurate doses could now be administered; this had been impossible previously with crude drugs of unknown and variable composition. Second, toxic effects due to impurities in crude drugs could now be eliminated if pure compounds were used; and third, knowledge of the chemical structure of these drugs led to attempts at laboratory synthesis, which led in turn to discovery of valuable related compounds.

Joseph Lister, in England, opened the modern era of antiseptic surgery in 1865 when he used phenol (carbolic acid) to prevent infections. In 1869 the soporific properties of chloral hydrate, the first synthetic hypnotic (sleep-producing drug), were discovered. In 1874 it was found that organic nitrites relax the blood vessels, and, in 1875, salts of salicylic acid were introduced as remedies for fever. The year 1879 witnessed the introduction of saccharin, still in use today as a sweetening agent for diabetic patients. The simple compound acetanilide, introduced in 1886, was one of the first analgesic–antipyretic drugs (*i.e.*, reducing both pain and fever) to be used but was later replaced by the less toxic phenacetin in 1887, by aspirin in 1899, and all of these to some extent by paracetamol in 1956. The hypnotic sulfonal (sulfonmethane) was discovered about 1888, followed a few years later by barbital; this latter led to a whole series of barbiturates of which phenobarbital is the best known.

Cocaine was the only known potent local anesthetic until about 1900, when the much simpler compound benzocaine was introduced. The closely related local anesthetic procaine followed early in the 20th century.

In 1909 arsphenamine, highly effective against syphilis, was introduced. Since arsphenamine is both insoluble in water and unstable, further work led to neoarsphenamine in 1912, a soluble derivative. Many other synthetic drugs followed, among which was the hormone progesterone, synthesized in 1934. Some thousands of similar compounds have since been prepared, some of which are used as oral contraceptives.

The sulfa drugs and antibiotics

In 1935 it was discovered that sulfanilamide (Prontosil) stopped the growth of bacteria. Over 6,000 derivatives of sulfanilamide—the sulfonamides, or sulfa drugs—were prepared and tested for their antibacterial properties. Today, the sulfonamides have partially been superseded by antibiotics, of which the first was penicillin, first isolated in 1941. In 1959, 6-aminopenicillanic acid was isolated for the first time; this led to production of many semisynthetic penicillins such as ampicillin, carbenicillin, cloxacillin, methicillin, oxacillin, and phenethicillin. Also used as antibiotics are compounds known as cephalosporins, the first of which was isolated in 1961.

Classification of pharmaceutical products. Drugs may be classified in one of three ways: by chemical group (*e.g.*, alkaloids, mentioned above); pharmacologically (*i.e.*, by the way they work in the body); and according to their therapeutic uses. Pharmacological and therapeutic classifications show considerable divergence, as drugs that act upon the body in different ways may bring about the same desired therapeutic result. Furthermore, classification by therapeutic usage is complicated by the fact that a drug may be used to combat more than one ailment. The antimalarial compound primaquine, for example, may also be employed to relieve arthritis. Some familiar drugs classified by therapeutic use include aspirin, an analgesic, or pain-killer; benzocaine, a local anesthetic; magnesium carbonate, an antacid; charcoal, an antiflatulent (preventing production of flatulent gases); penicillin, used against syphilis; calcium lactate, a calcium supplement; hexachlorophene, a deodorant; phenol-

phthalein, a laxative; levulose, a nutrient; cascara sagrada, a purgative; phenobarbital, a sedative; and thiamine hydrochloride, a vitamin. See also articles on major groups of drugs, such as ANESTHETIC; ANTISEPTIC AND GERMICIDE; CHEMOTHERAPEUTIC DRUGS; HORMONE; TRANQUILLIZER; and VITAMIN.

PRODUCTION OF PHARMACEUTICALS

Raw materials and their purity. Pharmaceutical raw materials may be plant, animal, or other biologic products; inorganic elements and compounds; or organic compounds. If the raw material is "official"—that is, if it is the subject of a monograph in a pharmacopoeia or national formulary—then the minimum acceptable degree of chemical purity is specified. Very often, however, because some raw materials at specified levels may begin to decompose after a time, purification far exceeds these minimum requirements. If extra purification is inconvenient, a preservative may be added. If the raw material is not the subject of an official monograph, then physical or chemical specifications or both are drawn up by the manufacturer in accordance with the pharmaceutical requirements of the finished product, on lines similar to the official monographs.

Crude drugs. The term crude drug is usually applied to plant or animal organs or whole organisms or exudations of these, either in the fresh state or dried and either ground or unground. Some crude drugs such as acacia, belladonna, and starch are official; for these, rigid specifications are available. Plant-derived crude drugs may come from cultivated sources or may be collected in the wild. Cultivation and collection are carefully planned to obtain drugs with the highest possible therapeutic activity or content of active constituents. The harvested drug plant must then be cleaned to remove matter such as sand and dirt. Next, unwanted plants or plant parts are carefully removed. Some crude drugs, such as belladonna, undergo curing, a process that consists either of slow drying or sweating, during which enzymes bring about chemical changes whereby the content of active ingredient is increased. Alternatively, as with cascara sagrada, the drug is carefully stored for a year, during which time unwanted constituents slowly decompose. The final stage in crude-drug production consists of drying, accomplished in air or with artificial heat. Drying aids preservation, stops various chemical reactions that might weaken or destroy the substance, facilitates subsequent grinding, and reduces weight and bulk. Drugs containing volatile constituents are dried at temperatures near freezing. After drying, some drugs (such as belladonna and ipecac) are reduced to a fine powder.

Drying the drug

Inorganic elements and compounds. Production methods for inorganic elements and compounds are not always the same as those for industrial chemicals, since purification is usually carried to a much higher level. Industrial zinc oxide, for example, may contain as much as 10 percent of impurities, including small quantities of lead. The pharmaceutical substance, by contrast, must contain at least 99 percent zinc oxide and no lead. For this reason, industrial zinc oxide is manufactured by combustion of zinc in air, whereas pharmacopoeial zinc oxide is produced from zinc sulfate.

Organic compounds. Organic compounds used as pharmaceuticals are either extracted from natural sources or prepared by chemical synthesis (various methods are described below). Physical specifications used in purity control of inorganic and organic pharmaceuticals include particle size, colour, odour, solubility, homogeneity, and freedom from particulate impurities. Some common chemical specifications are melting point, boiling point, density, viscosity, and freedom from impurities.

Preparation from natural sources. *The alkaloids.* As mentioned above, the first pure pharmaceuticals isolated from natural sources were the alkaloids. Though the methods adopted for their extraction from plants vary in detail, all are based on three general characteristics of these compounds. First, most alkaloids are only slightly soluble in water but readily soluble in certain organic solvents such as benzene, chloroform, ether, and light petroleum. Second, alkaloids combine with acids to form salts that are usually freely soluble in water but only slightly so in organic solvents. Third, alkaloids are liberated from their salts by alkalies.

Application of these general principles can be seen in the following generalized outline of extraction methods. The crude drug, ground to a suitable state of subdivision, is mixed with water, a water-soluble alkali such as lime, and some organic solvent that does not mix with water. The mixture separates into two layers: one contains water, lime, and impurities and is discarded; the other contains the alkaloid dissolved in the organic solvent. Fresh water and dilute acid are now added to the mixture. Again there are two layers, but the acid has caused the alkaloids to pass from the organic layer into the aqueous layer. The aqueous layer is now separated, and the alkaloidal salt or salts may be crystallized out by cooling or concentrating the solution. The process described above is used to obtain quinine sulfate from cinchona bark.

Extraction of alkaloids

Glycosides. Glycosides such as digoxin are another important group of drugs obtained from plants. Generally, the glycoside is extracted from the crude drug with alcohol. Adding basic lead acetate to the alcoholic extract causes the impurities to fall to the bottom. After being poured through a filter, the alcoholic extract is concentrated, and the glycosides may crystallize out. Often, however, concentration of the extract is not sufficient to cause crystallization, and more complicated procedures must be employed.

Volatile, or essential, oils. Volatile, or essential, oils, also obtained from plants, may be extracted by distillation, steam distillation, expression, or by extraction with fats or organic solvents. In steam distillation, the most common method, the crude drug—either fresh or dried—is used in powder form. Water is mixed with the powder, serving the double purpose of preventing decomposition of plant material by excessive heat and of facilitating volatilization of the essential oil. The powder–water mixture is usually placed in a basket through which the steam penetrates. The distillate consists of a water–oil mixture; the oil forms a separate layer, which is run off. Anise, cinnamon, clove, coriander, fennel, and peppermint oils are all obtained by this method, as is the pharmaceutical substance camphor from camphor wood.

Fixed or fatty oils cannot be obtained by distillation, but only by expression or extraction. Castor, olive, and sesame oil, for example, are all obtained by expression. Cod-liver and halibut-liver oils, rich sources of vitamins A and D, are both extracted by passing steam into tanks containing the livers suspended in water, until the temperature reaches 70–80° C (158°–176° F). The tissues disintegrate, and the oils float to the surface and are skimmed or centrifuged off.

Obtaining cod-liver oil

Insulin. Other drugs of animal origin primarily include glandular extracts containing hormones, such as thyroid and parathyroid, and purified hormones, such as insulin, obtained from the pancreas of a pig, sheep, or ox. For insulin production, the pancreas must either be absolutely fresh or frozen immediately after removal from the animal. Otherwise, enzymes will cause decomposition of the insulin. The glands are minced and immediately stirred with cold acidified alcohol, which both extracts the insulin and prevents its destruction by the enzyme trypsin. The gland residue is separated by centrifugation and further extracted with more acidified alcohol. The final alcoholic extract is passed through filter presses to remove the remaining fine particles of meat. The extract is now made sufficiently basic to cause precipitation of the inert proteins and allow their removal. Next, the extract is acidified again, and the alcohol is removed as quickly and with as little heat as possible. When the concentrated extract is cooled, fat separates and is filtered off. The crude insulin is then precipitated by adding common salt and is purified by repeated crystallization. For medical practice, insulin is almost always made up in the form of a solution for injection.

Antibiotics. Antibiotics are an extremely important group of drugs isolated from natural sources. Penicillin

was originally produced by growing the *Penicillium notatum* mold in small containers; the mycelium, or vegetative portion, formed a mat on the surface of the medium that contained the penicillin in solution. The process was difficult to operate and required a large labour force to inoculate and harvest the containers.

Deep culture of penicillin

Penicillin production was revolutionized by the discovery that a strain of *Penicillium chrysogenum* mold (isolated from an overripe cantaloupe) would produce high yields when grown in deep culture. The growth of a mold is aerobic (it requires air), and in surface culture the growth and consequent production of antibiotic is limited by the rate at which air can diffuse into the medium. In submerged or deep culture, the mold is grown in large tanks supplied with a continuous flow of sterile air.

Though the medium used in fermentation varies for the particular antibiotic being produced, all contain a source of carbon (which may be lactose or glucose); nitrogen (in the form of ammonium salts); and trace elements (such as phosphorus, sulfur, magnesium, iron, zinc, and copper). If phenylacetic acid is added, the mold will utilize it and produce benzylpenicillin (penicillin G), whereas, if phenoxyacetic acid is added, phenoxymethylpenicillin (penicillin V) is obtained. In addition, corn-steep liquor may be added for penicillin production, and soybean meal and dried distiller's residues for streptomycin; these help increase the yield of product. (Corn-steep liquor is prepared by steeping cleaned corn grain in water for about 40 hours at about 48° C [118° F]; the liquor is drawn off and evaporated at reduced pressure to a suitable concentration—about 55 percent solids. Corn-steep liquor stimulates penicillin formation due to certain amino acids, minerals, and precursors that it contains.) Before use the medium—as well as the fermenter and associated equipment—is steam sterilized, as bacterial contamination can destroy the antibiotic.

A large volume of concentrated, actively growing fungal suspension is required for the main fermenting tanks, to keep the fermentation time to a minimum. This is obtained in three stages. First, the selected culture is transferred from cold storage to a culture medium (agar) to produce an initial inoculum. This inoculum is then cultured in shake flasks to give a suspension. Finally, the suspension is grown in seed tanks in the plant for 24–28 hours to the desired volume and concentration before transfer to the main fermenters. Fermentation is continued for three to five days, during which the vessel is cooled—to keep the temperature between 23–27° C (73–81° F)—and stirred and aerated with sterilized air. The introduction of large volumes of air causes frothing, which is controlled by the addition of antifoams such as lard oil, octadecanol, or silicones. When fermentation is complete, the mycelium is removed on a rotary filter and the penicillin extracted into an organic solvent (such as butyl acetate or methyl isobutyl ketone), after acidification. The free acids of penicillins are generally unstable and are converted into a metal salt by extraction with alkali, followed by freeze-drying of the extract, or by addition of a concentrated solution of a metal salt such as potassium acetate, whereby the potassium salt of the penicillin is precipitated. All products that are to be administered by injection are sterilized by passage through a sterilizing filter, followed by freeze-drying, precipitation, or crystallization under sterile conditions.

Vaccines. Vaccines, viral or bacterial preparations used to induce active immunity, may be divided into three groups: toxoids, bacterial vaccines, and virus vaccines. Toxoids, such as diphtheria and tetanus toxoid, can only be prepared from toxin-producing bacteria. To produce diphtheria toxoid, for example, diphtheria bacilli are cultured under artificial conditions in a suitable broth to produce a fluid rich in toxin. The bacilli are then filtered off, and formaldehyde is added to the crude toxin filtrate, which is incubated for two to three weeks, with the result that the substance is no longer toxic and is called a toxoid. The toxoid is then purified by precipitation with ammonium sulfate and is filtered off. If alum or some other aluminum compound is used to precipitate the toxoid, the product is relatively insoluble and thus more slowly absorbed, thereby possessing increased antigenic (causing the production of antibodies) properties.

Preparation of bacterial vaccines

The first step in the preparation of bacterial vaccines, such as typhoid vaccine, is similar in that an artificial culture of the organism is first prepared. In this case, however, a solid medium is usually used, such as nutrient agar. If a mixed bacterial vaccine is being made, the individual species of organisms are cultured separately and later mixed. The culture is incubated at a suitable temperature for at least 24 hours until sufficient growth has taken place. The bacteria are then collected under aseptic conditions (free from pathogens) by pouring sterile salt solution on the culture and gently shaking or scraping until the micro-organisms have been suspended. If a liquid medium has been used for culture, it is centrifuged, and the bacterial sediment suspended in sterile isotonic saline. In either case, the crude product is shaken to produce a uniform suspension, and the bacterial concentration determined. The suspension is diluted with sterile isotonic saline, and the bacteria killed by a germicide or by heat; finally, a preservative is added.

In the preparation of virus vaccines, such as yellow-fever vaccine, the general principle is similar, with the important exception that most virus vaccines are weakened living virus preparations. The living virus is inoculated into the embryo region of hens' eggs that have been incubated for eight days. After a further six days' incubation, the embryo is removed, homogenized, and the vaccine separated. As these vaccines are unstable, they are freeze-dried and reconstituted at the time of inoculation. Poliomyelitis vaccine is usually prepared from viruses inactivated with formaldehyde. Some virus vaccines, such as measles and rubella, can now be prepared using human cells as the culture medium.

Human blood-plasma fractions. A number of proteins of therapeutic value can be concentrated from human plasma (see BLOOD, HUMAN); these include albumin, immunoglobulins or gamma-globulins that contain antibodies to various viruses, and blood coagulants.

Careful addition of ethanol to protein solutions at temperatures in the range −8° to 0° C (18° to 32° F) causes precipitation without gross damage to molecular structure. The various proteins in plasma have their minimum solubilities at differing acid or base (pH) values and salt concentrations and can be separated from each other by a series of precipitations in which these factors are controlled. Ethanol concentrations, protein concentrations, and temperatures must also be adjusted to precise values. Because protein molecules are unstable, excessive frothing or agitation, local high ethanol concentrations, or local temperature rise (due to the heat of mixing of ethanol and water) must be avoided. Precipitates are separated in refrigerated centrifuges, and the final products are freeze-dried, which removes residual ethanol and makes them more stable for storage and distribution. Albumin and immunoglobulins are sufficiently stable to be redissolved and dispensed in solution.

Difficulties in processing human blood fractions

Processing is complicated for several reasons: proteins, being unstable in heat, can be sterilized only by filtration; large volumes of some products may be given intravenously, thus precluding the use of preservatives; blood coagulant substances are extremely unstable in plasma even at refrigerator temperatures, necessitating their preparation within a few hours after collection of the blood from the donors; and certain diseases, of which blood donors may be carriers, notably serum hepatitis, can be transmitted in the fractions. Albumin is, however, stable enough to withstand heating for ten hours at 60° C or 140° F (which inactivates the hepatitis virus), and heated albumin preparations are now largely replacing whole plasma in transfusion therapy. It has been found that immunoglobulin fractions do not transmit this virus. Control of products includes tests for the presence of micro-organisms, pyrogens, and toxic substances; for pH and salt content; for protein composition; and, where appropriate, for heat stability, antibody activity, or coagulation function.

Steroid hormones. Another important group of pharmaceuticals widely distributed in plants and animals is

the steroid hormones (see also STEROIDS; HORMONE). The plant by-product called diosgenin is by far the most common starting material for the preparation of steroids. It is obtained from whole tubers of wild yams; these are broken up in a hammer mill, and the resultant mash is allowed to ferment for two days. It is then dried in the sun, hydrolyzed with mineral acid, and the diosgenin extracted with an immiscible organic solvent (one that will not mix with water), such as light petroleum or heptane.

The hormone estrone is usually manufactured from pregnant mares' urine, in which it is present almost entirely as the ethereal sulfate. This is hydrolyzed to estrone by acidification of the urine and heating, after previously covering with a layer of immiscible solvent, such as toluene. Estrogens are extracted from the solvent, and the estrone is separated from the other estrogens.

Progesterone was formerly obtained from the ovaries of sows or whales. It is now usually manufactured synthetically by a complex series of chemical reactions from various materials such as stigmasterol (obtained from soybeans) or diosgenin. Testosterone is also manufactured from diosgenin.

Further medicinal steroids, such as ethisterone, methandienone, and methyltestosterone, are obtained from an intermediate of testosterone production.

Sources of cortisone

Cortisone is manufactured from ergosterol, from progesterone (itself obtainable from stigmasterol or from diosgenin), and from other plant by-products, by routes involving both chemical and microbiological transformations. Hydrocortisone, in turn, is produced from cortisone. Deoxycorticosterone is made as the acetate from progesterone.

Vitamins. These were formerly extracted from natural sources, but now the majority are made by synthetic methods. Not only are the synthetic vitamins cheaper and purer, but their production is easier than extraction of natural vitamins.

Preparation by chemical synthesis. Two or more of the methods discussed below are usually involved in the manufacture of any one drug. There are many different possible routes for the production of pharmaceuticals. It is not often that the simplest possible theoretical pathway is the initial choice of the manufacturer; many different methods are chosen by manufacturers depending upon the starting materials and equipment available to them.

Acetylation. This is a special case of acylation, a process in which alcohols, phenols, and primary and secondary amines yield acyl derivatives by displacement of the hydrogen atom of the hydroxy, primary amino, or secondary amino group by an acyl radical. The commonest acyl radical by far is the acetyl radical, CH_3CO-. The usual acetylating agent is acetic anhydride, $(CH_3CO)_2O$, but glacial acetic acid, CH_3COOH, and acetyl chloride, CH_3COCl, are also used. With acetic acid and alcohols or phenols, it is usual to use a dehydrating agent, such as concentrated sulfuric acid or anhydrous sodium acetate, in order to eliminate the water formed in the reaction. With acetyl chloride, a base such as dimethylaniline or pyridine is added to remove hydrogen chloride as it is formed.

Acetylation is often used in drug production since it reduces the toxicity of amines. Thus the drug paracetamol is much less toxic than *p*-aminophenol. Paracetamol is manufactured from *p*-nitrophenol.

Addition. This chemical reaction can only take place with unsaturated compounds, but the addition products vary greatly in nature. The mode of addition, where there are two possibilities, is determined by the structure of the unsaturated compound and the mechanism of addition. (For detailed explanation of this reaction, see CHEMICAL REACTIONS.)

The anesthetic drug halothane is manufactured from trichloroethylene (see below *Dehydrohalogenation*) by an addition reaction with hydrogen chloride. The product is treated with hydrogen fluoride; reaction with bromine then gives halothane.

The manufacture of the anesthetic drug trichloroethylene involves, as a first step, the addition of chlorine to acetylene, followed by dehydrohalogenation.

Alkylation and arylation. Alkylation is a general term for the introduction of an alkyl group into a compound and is more specifically known as methylation, ethylation, propylation, butylation, etc., depending upon the alkyl group inserted. Alkylation of an alcohol or a phenol results in an ether, while alkylation of a thiol gives a sulfide; primary and secondary amines can also be alkylated. The usual alkylating agents are alkyl sulfates and halides. The alcohol, phenol, thiol, or amine is dissolved in sodium hydroxide solution, and dimethyl sulfate, for example, is added to yield the alkylate.

In addition to the above types of compounds, any compound containing "active hydrogen" can be alkylated. Thus, compounds containing a methylene ($-CH_2-$) or a methyne ($=CH-$) group adjacent to a group such as carbonyl ($-CO-$) contain active hydrogen.

The important drug codeine, widely used as a cough suppressant and pain-killer, occurs naturally as an alkaloid in opium, together with morphine. Demand for codeine is such that much of it is manufactured by methylation of morphine. Morphine is dissolved in potassium hydroxide solution, and the solution is treated with methyl iodide to produce codeine.

The well-known constituent of tea and coffee, caffeine, used as a stimulant drug, is produced either from tea wastes or synthetically from the organic chemicals theophylline or theobromine. These three materials are closely related, and methylation of either theophylline or theobromine or a mixture of the two results in the fully methylated caffeine.

Synthesis of barbiturates

The barbiturates, widely used as sedative–hypnotic drugs, provide many examples of alkylation and arylation (introduction of one or more aromatic radicals, such as phenyl or naphthyl, to a molecule). The general method for their synthesis is as follows. Diethyl malonate (a malonic-acid ester, which contains an active methylene group) is treated with an alcoholic solution of sodium ethoxide, and the monosodium derivative produced is converted by reaction with an alkyl halide into an alkylmalonic ester. This is either used as such in the preparation of the barbiturate, or, as is usually required, a second alkyl group, either the same or different, is introduced in the same way. The alkylmalonic or dialkylmalonic ester is then condensed (see below *Condensation and cyclization*) with urea, in the presence of sodium ethoxide, to form the sodium salt of barbituric acid. This is either used as such, since sodium barbiturates are usually very soluble in water, or the free barbituric acid is liberated by acidification. Barbiturates containing aromatic groups (phenobarbital and mephobarbital are the only common examples) must be made by a modified route, because of the relative lack of reactivity of aryl halides. For both these drugs, ethyl phenylacetate is condensed with ethyl oxalate to form ethyl oxalophenylacetate, which, when distilled, decomposes to give diethyl phenylmalonate. Ethylation is then effected to yield diethyl ethylphenylmalonate. For the manufacture of phenobarbital, this ester is condensed with urea as in the scheme above.

Amination. Two important methods of amination (the process of adding an amino ($-NH_2$) group to a compound) are reduction and ammonolysis. There are many reduction methods, but the most common is the reduction of the corresponding nitro compound by means of iron and acid. An example of ammonolysis is the preparation of aniline by heating chlorobenzene with excess aqueous ammonia, in the presence of cuprous oxide, at about 200° C (392° F) under pressure. Amination enters into the synthesis of many pharmaceuticals; a good example is amphetamine. Phenylacetone is treated with ammonia and hydrogen in the presence of a nickel catalyst, and reductive amination takes place, forming amphetamine.

Arsonation. Arsonation applied to aromatic compounds is closely analogous to nitration (see below). For example, *p*-aminophenylarsonic acid (arsanilic acid) is made by heating aniline with concentrated arsenic acid. A method other than arsonation that achieves the same result is Bart's reaction, in which a diazonium salt is coupled with an alkali arsenite, usually in the presence

of a catalyst such as copper powder. Sodium phenylarsonate, for example, can be made from benzenediazonium chloride by Bart's reaction. The drug acetarsol is manufactured from phenol by arsonation followed by nitration, reduction, and acetylation.

Carboxylation. This reaction is best exemplified by the manufacture of salicylic acid, used in external preparations as a skin softener. Dry sodium phenoxide is heated with carbon dioxide, under pressure, at about 130° C (266° F); the resulting sodium salicylate is dissolved in water, and the acid precipitated by acidification with hydrochloric acid. Salicylic acid is used in medicine as the free acid, as its acetyl derivative, aspirin (see below *Esterification*), and as its salts, of which the most important is sodium salicylate, used in the treatment of rheumatism. To obtain the latter product, a weighed amount of salicylic acid is made into a paste with water, and the equivalent quantity of pure sodium bicarbonate is added. The mixture is stirred and warmed until a clear solution is obtained, and this is evaporated to dryness; the sodium salicylate is purified by crystallization from alcohol.

Condensation and cyclization. Condensation is difficult to define, but it involves the linking together of two or more organic molecules (identical or different), with or without the elimination of a simple molecule such as water, ethanol, or a hydrogen halide. Reactions involving condensations that can be defined more precisely, such as esterification, etherification, or polymerization, are usually not described as condensations.

Hexylresorcinol, used for destroying intestinal worms, provides a simple example of a condensation. It is produced by condensing resorcinol with hexanoic acid, in the presence of anhydrous zinc chloride, at about 130° C (266° F). Water is eliminated, and the resulting ketone is reduced by means of zinc amalgam and dilute hydrochloric acid. The sedative meprobamate involves two condensations in its production. Diethyl methyl-*n*-propylmalonate is reduced with lithium aluminum hydride, $LiAlH_4$, to the corresponding dihydric primary alcohol; this is condensed with carbonyl chloride, $COCl_2$, with elimination of hydrogen chloride to give a dichloroformate; condensation of this with ammonia, again with elimination of hydrogen chloride, gives meprobamate. Cyclization is a special case of condensation; internal condensation takes place in one molecule, with resulting formation of a cyclic molecule. The compound *o*-iminodibenzyl, from which the antidepressant drugs desipramine and imipramine are produced, illustrates cyclization. The iminodibenzyl is condensed, in the presence of sodamide, $NaNH_2$, with 1-chloro-3-methylaminopropane to give desipramine or with 1-chloro-3-dimethylaminopropane to give imipramine.

Production of meprobamate

Dehydration. This reaction, as the name indicates, involves the loss of the elements of water. If this takes place internally in a molecule, the result is usually an alkene or a cyclic compound. If dehydration takes place between two molecules (condensation), the product is usually an ether.

Diethyl ether (ordinary ether), used as an anesthetic, is produced by indirect condensation of ethanol. A mixture of concentrated sulfuric acid and ethanol is heated to 140° C (284° F), when ether distills over; more ethanol is run into the mixture at the same rate as the ether distills. The ethanol esterifies with the sulfuric acid to form ethyl hydrogen sulfate. This ester, on heating with ethanol at 140° C, gives ether and sulfuric acid back again.

Vinyl ether is also used as an anesthetic and provides a more complex example of dehydration. Ethylene chlorohydrin with sulfuric acid gives 2,2′-dichlorodiethyl ether, and this is dehydrohalogenated (see below) by means of solid potassium hydroxide to produce divinyl ether.

Dehydrohalogenation. This process may be defined as removal of the elements of hydrogen halides with the resultant formation of an unsaturated bond. It is usually effected by means of a base such as calcium oxide or hydroxide. Examples of its application are the production of the anesthetic trichloroethylene and the drug tetrachloroethylene (used for destroying intestinal worms), both involving the elimination of the elements of hydrogen chloride. Acetylene, readily available and comparatively cheap, is used as the starting material. The reaction with chlorine is dangerously explosive under ordinary conditions, but the reaction can be controlled to give acetylene tetrachloride. This readily yields trichloroethylene. In turn, trichloroethylene takes up one molecule of chlorine to form pentachloroethane, and this, treated in the same way as acetylene tetrachloride, produces tetrachloroethylene.

Esterification. The addition of an acid to an alcohol generally results in esterification; that is, the production of an ester plus water. The reaction is reversible, and the rate of the reverse reaction (see below *Hydrolysis*) will increase as the concentration of water (or ester) increases. Hence, both the anhydrous acid and the anhydrous alcohol are normally used in order to produce good yields. A catalyst such as dry hydrogen chloride speeds up the attainment of equilibrium (but will not affect the point of equilibrium). Concentrated sulfuric acid may be used, but it affects the point of equilibrium since it is a dehydrating agent as well as a catalyst; it may also dehydrate the alcohol and so must not be used in too high a concentration.

Many other catalysts are employed in esterification (or hydrolytic) processes. The acid chloride or anhydride is often used instead of the acid. A simple example is the production of ethyl chloride (an anesthetic) by passing dry hydrogen chloride gas into dehydrated ethanol containing anhydrous zinc chloride.

Acetylsalicylic acid (2-acetoxybenzoic acid, or aspirin) is the most widely used of all synthetic drugs. It is usually manufactured by the action of acetic anhydride on salicylic acid. As aspirin is only slightly soluble in water, it is easily separated by dilution of the reaction mixture with water. It is readily crystallized from aqueous alcohol, benzene, or light petroleum or from mixtures of these or other solvents with glacial acetic acid (to minimize hydrolysis).

Production of acetylsalicylic acid

Halogenation. This may involve reactions of addition, substitution, or replacement of a group of atoms. The catalysts used include halogen carriers such as iron, antimony, or phosphorus, all of which exert two valences as halides. They alternately add on and give up halogen to carry on the reaction. Since the halogens form unstable interhalogen compounds, chlorine, bromine, and iodine can also be used as catalysts. Under suitable conditions, substitution of alkenes with halogens takes place, and not addition. The synthetic oral estrogen, chlorotrianisene, for example, is produced from deoxyanisoin by reaction with 4-methoxyphenyl magnesium bromide, followed by subsequent dehydration and substitutive chlorination with a solution of chlorine in carbon tetrachloride. Gamma benzene hexachloride (gammexane), a delousing agent, is manufactured by the addition (not substitution) of chlorine to benzene. The product is a mixture of four of the nine possible stereoisomers, and the gamma isomer, which is the most active as an insecticide, is separated from the others.

Hydrolysis. This term is applied to reactions wherein water effects a double decomposition with another compound, hydrogen going to one component and hydroxyl to the other:

$$XY + H_2O = HY + XOH$$

The reversal of esterification is an example of hydrolysis. Catalysts that speed up esterification also accelerate hydrolysis; many enzymes will also catalyze hydrolytic reactions.

A pharmaceutical example is the production of propylene glycol. Propene is oxidized and converted by acid-catalyzed hydrolysis to propylene glycol. The pharmaceutical substance called wool alcohol is made by hydrolysis of wool grease.

Inorganic substances. Helium, nitrogen, oxygen, sulfur, and some inorganic compounds still find a place in medical practice. These are prepared by the general methods used in inorganic chemistry; however, they must comply with pharmacopoeial standards of purity, where-

as those used for nonmedical purposes are not usually purified up to these standards.

An example is the manufacture of ferric ammonium citrate, used in iron-deficiency anemias. The first step is preparation of freshly precipitated ferric hydroxide by slow addition, with constant stirring, of a ferric-salt solution (*e.g.*, ferric sulfate) to an alkali solution such as sodium hydroxide:

$$6NaOH + Fe_2(SO_4)_3 = 3Na_2SO_4 + 2Fe(OH)_3.$$

The ferric hydroxide is collected and washed and, without being dried, is stirred with enough citric acid solution to dissolve almost all of it. A slight excess of ammonia solution is then added, and the residual ferric hydroxide filtered off. The clear, reddish-brown filtrate is evaporated to a syrup, a little ammonia being added occasionally to maintain a slight excess. The syrup is painted on glass plates, dried at a temperature below 40° C (104° F), and scraped off as scales. Ferric ammonium citrate is also produced in the form of granules. The medicinal substance is dark red in colour and contains about 21 percent iron. Green ferric ammonium citrate is also made; it contains about 16 percent iron.

Mercuration. This is usually effected by reaction with a solution of mercuric acetate in acetic acid or alcohol; substitution of a hydrogen atom by the acetoxymercuri group, $-HgOCOCH_3$, results. The topical antiseptic acetomeroctol, for example, is prepared by interaction of a complex organic chemical compound with the calculated quantity of mercuric acetate in a 50 percent aqueous ethanolic solution containing 5 percent acetic acid. The mixture is maintained at room temperature for one to two weeks.

Nitration. Medicinal nitro compounds are invariably aromatic carbocyclic or heterocyclic derivatives. These are usually nitrated with a mixture of concentrated sulfuric acid and concentrated or fuming nitric acid (mixed acid). Other nitrating agents include oxides of nitrogen and acetyl nitrate, $CH_3CO_2NO_2$.

Nitrofurantoin is prescribed as a urinary antiseptic and provides an example of nitration. Furfural is reacted with acetyl nitrate, and the resulting triacetate is treated with pyridine to give 5-nitro-2-furaldehyde diacetate. This is condensed (with elimination of acetic acid) with 1-aminohydantoin to yield nitrofurantoin.

Oxidation. This reaction is very widely used in the preparation of chemical compounds. Oxidation may be defined as the loss of electrons or hydrogen or the gain of oxygen; reduction (see below) has the opposite meanings. There are many oxidizing agents available; the choice depends upon the extent to which oxidation is desired. Air, oxygen, ozone, peroxides, chromium trioxide, chlorates, dichromates, permanganates, nitric acid, and sulfuric acid are all common examples. Conversion of the alkaloid nicotine to the vitamin nicotinic acid provides an example of oxidation (see also OXIDATION–REDUCTION REACTIONS).

Oxidizing agents

Polymerization. Polymerization reactions may be broadly divided into addition polymerization and condensation polymerization (polycondensation). The term polymerization strictly refers to addition polymerization but is often used to include all types of polymerization reactions (see also POLYMERS). Addition polymerization may be defined as the joining together of two or more molecules of a compound in such a way that the molecular weight of the polymer produced is a multiple of that of the original compound. Addition polymerization is usually induced either by heat, pressure, or a catalyst or a combination of these.

A pharmaceutical example is the preparation of paraldehyde, used as a sedative drug. A small proportion of concentrated hydrochloric acid is added to acetaldehyde, causing three molecules of the liquid to polymerize to form a molecule of paraldehyde, a saturated heterocyclic system. The reaction involves the mutual saturation of three unsaturated carbonyl groups.

Condensation polymerization involves a similar joining together of molecules, but this time a simple molecule, such as water, is condensed out. Diethylene glycol, for example, can be made by heating glycol with a dehydrating agent. Higher polyethylene glycols can be made similarly, such as polyethylene glycol 400, with a molecular weight of about 400, and polyethylene glycol 4000, with an average molecular weight of about 3,400. These two polymers are employed in the preparation of a water-soluble ointment base.

Reduction. This reaction, the opposite of oxidation, has very wide applications in the production of pharmaceuticals. Examples of reducing agents are aluminum amalgam, hydriodic acid, hydrogen and a catalyst, hypophosphorous acid, lithium aluminum hydride, metal and acid, sodium and alcohol, stannous chloride, and zinc dust and water. Reduction is employed in the synthesis of racemic menthol, an anti-itching agent. It is produced by the catalytic hydrogenation of the aromatic analogue, thymol, or of the saturated ketone, menthone, or of the unsaturated ketone, pulegone.

Sulfation. This consists of direct linkage of a sulfate group to carbon, yielding an acid sulfate, $ROSO_2OH$, or a sulfate, $ROSO_2OR$. The most usual sulfating agents are concentrated sulfuric acid, oleum, sulfur trioxide, chlorosulfonic acid, or sulfamic acid, NH_2SO_3H.

The surface-active agent sodium lauryl sulfate, for example, is produced from commercial lauryl alcohol (mainly dodecanol) by reaction with chlorosulfonic acid. The resulting lauryl hydrogen sulfate is neutralized with sodium hydroxide.

Sulfonation. This consists of direct linkage of the sulfonic-acid group, $-SO_2OH$, or salts thereof, or of the sulfonyl halide group, $-SO_2X$, to a carbon or nitrogen atom. The latter are known as *N*-sulfonates or sulfamates. The same reagents are employed as are used for sulfation, above.

In the production of sulfonamides, the intermediate 4-acetamidobenzenesulfonyl chloride is made from acetanilide and chlorosulfonic acid. Most sulfonamides are made by the condensation of this intermediate with the appropriate primary amine, followed by removal of the acetyl group by alkaline hydrolysis. The hydrolysis is not usually difficult in practice, as the $-CO-NH-$ bond is more easily broken than the $-SO_2-NH-$ linkage. After the alkaline hydrolysis, the sulfonamide is precipitated by neutralization.

Complex chemical conversions. Ascorbic acid (vitamin C) provides an example of a complex synthesis. It is manufactured from D-glucose; this is converted by means of microbiological oxidation to a keto acid, which is catalytically reduced to L-idonic acid. Microbiological oxidation converts this to 2-OXO-L-gulonic acid. The 2-OXO-L-gulonic acid is converted to its methyl ester, which is isomerized and cyclized to L-ascorbic acid.

PREPARATION OF DOSAGE FORMS

Before describing individual types of dosage forms, some consideration of the process of size reduction or comminution is necessary. The degree of comminution varies according to the particular drug and the method of extraction used. Thin leaves and certain substances such as aloes or tolu need no treatment at all. Others may be cut, crushed, broken into small pieces, or sliced thin. Finally, the drug may be powdered or ground to one of five different degrees of fineness. A number of grinding machines are employed for powdering, the most important being the disintegrator, edge-runner mill, ball mill, and end-runner mill.

Process of comminution

Solutions. Many pharmaceuticals are simply solutions of a medicament in water, alcohol, ether, glycerin, or some other solvent. Various flavoured waters (*e.g.*, chloroform or peppermint water) may also be used as solvents. After mixing, many solutions require additional treatment such as sterilization or addition of antimicrobial agents, stabilizers, or buffers.

Spirits are solutions of a volatile substance in alcohol or a mixture of water and alcohol. Elixirs are clear, sweetened hydroalcoholic (alcohol and water solvent) liquids for oral use, usually containing potent or nauseous drugs. Simple elixirs are prepared by solution, while those made from plant drugs are produced by maceration or percola-

tion (described below). Tinctures, solutions of a nonvolatile substance in alcohol or a mixture of water and alcohol, are manufactured by simple solution, maceration, or percolation.

In maceration, the drug, suitably prepared, is placed, together with the extracting solvent (known as the menstruum), in a closed vessel and left for a defined period, usually three to seven days, with occasional shaking. The product is then filtered, the residue on the filter (known as the marc) is pressed to avoid loss, and the mixed filtrates are clarified, either by subsidence (settling out) or filtration.

Percolation

In percolation, the powdered drug is first moistened with the extracting solvent, allowed to stand for a defined period, and then carefully packed in a conical or cylindrical vessel with a perforated false bottom and a tap at the base known as a percolator. Sufficient extracting solvent is added to saturate the drug, and, when liquid begins to drip from the percolator, the tap is closed and sufficient solvent added to form a layer above the drug. Maceration is then allowed to proceed for a defined period (usually 24 hours), followed by slow percolation, more solvent being added as necessary, until the desired volume of percolate is obtained, which is then clarified by subsidence or filtration.

Fluid extracts are solutions, in alcohol, water, ether, or a mixture of these, of the active part of a vegetable drug so prepared that one cubic centimetre of extract has the same strength as one gram of the dry drug. More concentrated than tinctures, fluid extracts are made by percolation or maceration.

Numerous other types of solutions are used in pharmacy, such as collodions (viscous liquids for external application) or liniments (liquids for external application, usually consisting of a drug in oil or soap solution).

Semisolid dosage forms. Ointments are preparations for external use, intended for application directly on the skin or on a surgical dressing. An ointment base is chosen depending upon whether the final product is intended for absorption by the skin or not. Oily- or fatty-base ointments may have hard, soft, or liquid paraffin bases or mixtures of these in such proportions as will render an ointment of suitable consistency. Anhydrous lanolin (wool fat), also included in this group, is particularly suitable whenever a large volume of water or aqueous solution must be incorporated into the ointment, as anhydrous lanolin forms a stable water-in-oil emulsion. Anhydrous lanolin is readily absorbed by the skin and is often mixed with paraffins when both absorption and protection are desired. Emulsifying bases (in which minute globules of fatty substances are suspended) include anhydrous lanolin and hydrophilic petrolatum, which has similar properties. (The solid particles of a hydrophilic substance strongly attract and hold molecules of water.) Other ointment bases are emulsion bases, mixtures of oily substances and water; and water-soluble bases, generally prepared from polymers of ethylene glycol.

When the active medicament in the ointment is a solid, it is powdered as finely as possible, and then incorporated by the principle of geometric dilution (see below *Solid dosage forms*). The ingredients of ointment bases are often mixed together by fusion in a water bath, followed by stirring until cold, the drug being incorporated at a suitable stage. In manufacturing practice, ointments are usually improved by being passed through a roller mill; this ensures uniformity and smoothness. Creams, with the same therapeutic purposes as ointments, are solid or semisolid emulsions of the water-in-oil or oil-in-water type, giving greasy or nongreasy preparations, respectively. Pastes, similar to ointments, are usually stiffer, less greasy, and more absorptive, due to a higher proportion of powdered ingredients, such as calcium carbonate, talc, or zinc oxide. They are made by similar methods but with a glycogelatin, paraffin, or starch basis, or a mixture of these. Jellies used for similar purposes are made with an aqueous gel such as acacia, cellulose derivatives, chondrus, gelatin, gelatinized starch, or tragacanth as a base.

Solid dosage forms. Solid dosages, such as tablets, have many advantages over other types: greater stability, less risk of chemical interreaction between different medicaments, smaller bulk, accurate dosage, and ease of production.

Powders

Powders intended for internal use are usually mixtures of two or more ingredients. If two ingredients are present in unequal quantities, then the lesser ingredient (usually the drug) is mixed with an equal weight of the greater ingredient. Next, the resulting mixture is combined with an equal weight of the greater ingredient in steps until the mixture is complete. This process of geometric dilution (or trituration) is essential in order to produce a homogeneous powder. Powders may be sold in a bulk form or in individual packets. Granules, a dosage form related to powders, are particularly suitable for the preparation of solutions or mixtures of drugs, such as antibiotics, that are unstable in the presence of water. Cachets, occasionally used for administration of powdered drugs with an unpleasant taste, consist of shells made of gelatinized starch paste, the powder being placed inside, and the cachet swallowed. More common today is the hard capsule, in which the powder is enclosed in a shell of hard gelatin. Semiliquid and liquid drugs are often enclosed in a soft capsule with a soft gelatin shell.

Pills. Before the machine-made compressed tablet, pills were a very popular solid dosage form, being prepared at the dispensing bench by the pharmacist. Today pills are rarely prescribed, though some popular types are manufactured by machine. The powdered ingredients are mixed together with a binding agent, such as acacia or tragacanth, and are then made into a plastic mass by incorporation of any liquid drugs and addition of an inert liquid. The resulting mass, known as a pill mass, is then rolled into spheres and coated with talc, gelatin, or sugar.

Tablets. Tablets, by far the most common method of administration of drugs, are only rarely made by compression of the drug alone (*e.g.*, potassium bromide tablets). Usually, the drug is mixed with suitable diluents, such as dextrin, lactose, salt, starch, or synthetic substances, designed to ensure disintegration of the tablet in the body. To prevent sticking in the machine, a lubricant such as liquid paraffin, stearic acid, talc, or a synthetic substance is usually added. Furthermore, it is essential that the tablet machines are fed with the drug mixture in a free-flowing form to ensure complete filling of the molds. To achieve this, the drug mixture is customarily granulated by mechanically forcing pellets of the mixture through a sheet of perforated metal. The granulated mixture is fed into the tablet machine, which feeds the correct dose into a cavity, the mixture then being compressed by means of a punch that fits into the cavity. To be successful, the tablet maker must choose correct diluents and lubricants, prepare suitable granules, and obtain the right degree of compression in the tablet machine. Excessive compression may mean that the tablet will not disintegrate in the body; insufficient compression results in fragile tablets that may break, causing inaccurate dosage.

Tablet coatings

Coatings of various types may be applied to the tablet—to protect the ingredients from deterioration, to hide the taste of certain drugs, to control the release of the drug from the tablet, or to produce a more attractive tablet. For sugarcoatings, a concentrated sucrose syrup containing suspended starch, calcium or magnesium carbonate, or other suitable substance is applied, each successive layer being dried before the application of the next. After the final layer is dried, it is highly polished to give an elegant finish. Sugarcoatings provide both protection and a sweet taste. The chief drawback to sugarcoating is the long time involved. This led to the development of film coating, in which a very thin transparent film, usually a cellulose derivative, is applied. Enteric coating is designed to resist solution in the stomach and to dissolve in the more alkaline intestinal fluid. Many substances have been used for enteric coatings, one of the more recent being cellulose acetate phthalate (cellacephate). In the manufacture of layered tablets, incorporating two or more drugs, a compressed tablet is fed to a second machine where another layer is compressed around it. In this way, drugs normally incompatible may be formulated in the same tablet.

Other solid dosages. Troches, also known as lozenges or pastilles, disintegrate or dissolve in the mouth, slowly releasing the active drug. The base usually consists of a mixture of sugar and gum or gelatin. Lozenges are generally manufactured by compression techniques, while pastilles are fabricated by fusion and the use of molds. Dry extracts are prepared by the methods described above for fluid extracts, followed by evaporation, usually under reduced pressure, either to a pilular consistency or to dryness. Dry extracts are usually granulated by being passed through a sieve and may be used for the preparation of tablets. Suppositories are solid, uniformly medicated masses designed for introduction into the rectum. Various bases are used in their preparation, but theobroma oil is the most common. Solid at room temperature, it melts a few degrees below body temperature. Suppositories are manufactured by the use of molds, together with fusion of the suppository mass or cold compression. Pessaries are suppositories intended for introduction into the vagina, while bougies are designed for insertion into the urethra, nostrils, or ears.

Sterile medicaments. Some drugs cannot be given orally, because they are decomposed by the digestive processes; these drugs must be administered by injection, intravenously, intramuscularly, or subcutaneously (under the skin). It is imperative that such injected solutions be free from micro-organisms or toxic agents; other pharmaceutical preparations that must be sterile include eye drops, eye lotions, eye ointments, implants, powders to be applied to wounds or body cavities, and any solution or preparation to be used in surgical operations, such as surgical dressings, ligatures, and sutures.

Sterilization methods

There are several main methods for preparing sterile products. Steam sterilization is carried out in an enclosed chamber (autoclave). The material is distributed into its final containers, which are sealed to exclude micro-organisms. The containers are then placed in the autoclave and subjected to saturated steam under pressure at a temperature of 121° C (250° F) for 15 minutes. Substances stable in heat may be sterilized at higher temperatures, for which a shorter heating period is sufficient; conversely, lower temperatures require longer sterilization periods.

Pharmaceuticals in which the solvent is not water cannot be sterilized by steam, so dry heat at temperatures of 140° C (284° F) or higher is used, with exposure times ranging from one to four hours. The temperature–time relationship is similar to that for steam sterilization. Another technique is gaseous sterilization, in which the material is exposed to a vapour or gas such as ethylene oxide; this method is much used for foods, biologicals, and medical equipment (*e.g.*, cotton wool, syringes, needles, and tubing). Radiation sterilization generally involves exposure to ultraviolet or gamma radiation or high-energy electrons; its use is limited mainly to the production of sterile medicaments and apparatus on a large scale. The final method, aseptic manipulation, is not actually a sterilization process. Here, the separate ingredients of the pharmaceutical are all available in sterile form and merely require compounding without microbial contamination.

Injections are aqueous or oily solutions, suspensions or emulsions, prepared by normal methods, with special care taken to remove all extraneous particulate matter. Injections must be sterilized by one of the methods given above. Some aqueous injections are not stable and so are prepared at the time of use by adding sterile water to the sterile drug.

Eye drops, eye lotions, and eye ointments are all prepared by general methods, but they must be sterile. The pH value (acid–alkaline content) of eye drops and eye lotions is important; a compromise must be made between the ideal value of 7.4 (almost neutral) and the value of greatest stability or therapeutic activity. Implants or pellets are prepared either by fusion and molding of the drug or by compression of the sterile crystals.

Dispersions. Aerosols were formerly defined as colloidal systems consisting of very finely subdivided liquid or solid particles dispersed in a gas. Today, the term aerosol, in general usage, has become synonymous with a pressurized package. For pharmaceutical purposes aerosols may be divided into two types. Space sprays disperse the medicament as a finely divided spray with particles not exceeding 50 microns (0.05 millimetre, or 0.002 inch) in diameter. Surface-coating aerosols produce a coarse or wet spray and are used to coat surfaces with a residual film. Propellants used in aerosols are of two main types: liquefied gases and compressed gases. The former consist of easily liquefiable gases such as halogenated hydrocarbons. The drug is dissolved in the liquefied gas or in a mixture of the gas and a suitable solvent. When these are sealed into the container, the system separates into a liquid and a vapour phase and soon reaches an equilibrium. The vapour pressure pushes the liquid phase up the standpipe and against the valve. When the valve is opened by pressing down, the liquid phase is expelled into air at atmospheric pressure and immediately vaporizes, leaving an aerosol of the drug. The pressure inside the container is maintained at a constant value as more liquid changes into vapour. When compressed gases are used as the propellant, the pressure falls steadily as the contents of the aerosol are used, and for this reason liquefied gases are used whenever possible. Pharmaceutical aerosols include solutions, suspensions, emulsions, powders, and semisolid preparations. The products include inhalation aerosols, spray-on bandages, creams, and ointments. The application of these latter to wounds and burns is obviously advantageous, as rubbing is eliminated, and the fine film produced promotes rapid absorption. Inhalation aerosols often include a metering valve, so that measured quantities of drug may be administered; these are rapidly replacing old hand sprays.

Aerosols

Sprays, solutions of drugs in aqueous or oily solutions, are applied by means of an atomizer to the mucous membranes of the nose or throat. Oily solutions are no longer considered desirable, and the ideal spray is an aqueous solution isotonic (equal in osmotic pressure) with nasal secretions and of the same pH.

Suspensions

Since some drugs are insoluble in all solvents suitable for medicinal use, they must be administered either as a solid dosage form or as a suspension. Suspensions are chemically more stable than solutions. Apart from aerosols, pharmaceutical suspensions almost always consist of a finely divided solid dispersed in a liquid. The state of subdivision varies from colloidal particles to particles that only slowly subside on standing. Suspensions should not cake on standing, and the solid phase should readily redisperse on shaking; suspensions should pour fairly easily, so the viscosity must not be too high. Gels are special suspensions, namely, those of hydrated drugs in an aqueous medium, in which the particle size is colloidal or nearly colloidal. Lotions, such as calamine lotion, are suspensions for external use only. Magmas and milks are thick, viscous, aqueous suspensions of insoluble inorganic compounds; the particle size is usually larger than in gels. Bentonite magma, for example, is produced by hydration of bentonite, a colloidal hydrated aluminum silicate, and is used as a suspending agent, as, for example, in calamine lotion. Milk (cream) of magnesia, a magnesium hydroxide mixture, is an aqueous suspension of magnesium hydroxide. Mixtures, in the wider sense, are any combination of drugs prescribed and dispensed for internal use but, in the narrow sense, are official mixtures of standard composition. Official mixtures are liquid preparations of one or more solid drugs dissolved or suspended in water; they often contain a thickening or suspending agent such as tragacanth. Thus, most mixtures may be regarded as a special type of suspension.

Mucilages are thick, viscous, aqueous solutions of gums, frequently containing a preservative such as chloroform or benzoic acid. As they are colloidal in nature, they fall between true solutions and suspensions. Acacia and tragacanth mucilages are the best known examples and are used to aid in suspending insoluble solids in liquids.

Emulsions consist of one liquid dispersed in another. Pharmaceutically, they are intended for internal use and consist of small globules dispersed in water. Oil-in-water emulsions will mix with water, whereas water-in-oil

emulsions only mix with oils. However well two immiscible liquids are mixed together, on standing they will separate into two layers. To prevent separation, an emulsifying agent is used. Emulsifying agents can be divided into three groups: finely divided solids such as bentonite and magnesium aluminum silicate; natural emulsifying agents such as cholesterol, gelatin, acacia, methylcellulose, pectin, and tragacanth; and synthetic emulsifying agents such as the anionic sodium lauryl sulfate, the cationic benzalkonium chloride, and the nonionic polyethylene glycol 400 monostearate. A preservative such as chloroform is usually added to emulsions. The production of emulsions depends on the emulsifying agent used. Equipment includes a wide variety of agitators, colloid mills, homogenizers, and ultrasonic devices.

Radioactive dosage forms. A radiopharmaceutical is a medical product incorporating a radioactive isotope. Radiopharmaceuticals are widely used for various diagnostic tests and to a lesser extent as therapeutic agents. Radioactive iodine in the form of sodium iodide has been extensively used in the diagnosis of thyroid disorders. The iodine isotopes used are iodine-131 and iodine-125, orally administered as a sodium iodide solution; the thyroid uptake of radioactive iodine is measured with a measuring instrument placed close to the thyroid gland. Vitamin B_{12} containing cobalt-57, cobalt-58, or cobalt-60 is employed for the diagnosis of pernicious anemia. Chlormerodrin injection containing mercury-203 is used for renal (kidney) scanning and brain scanning.

WORLDWIDE PHARMACEUTICAL PRODUCTION

It is impossible to produce an accurate figure for worldwide pharmaceutical production because published statistics are not available for all countries and because some nations lump pharmaceutical production under the general heading of chemicals. The Table, however, shows pharmaceutical production in selected countries for 1965 through 1970 in millions of dollars. Though the United States is first in world production by a wide margin, several other countries, notably Japan, West Germany, and France, are very important. The highest percent of increase in the 1965 to 1970 period is shown in Japan.

Production of Pharmaceuticals in Selected Countries*
(in $000,000)

	1965	1966	1967	1968	1969	1970	increase 1965–70 (percentage)
United States	3,352	3,675	4,073	4,440	4,850	...	44.7†
Japan	1,268	1,399	1,556	1,926	2,354	2,873	126.6
France	951	1,064	1,161	1,275	...	...	34.1‡
Germany, West	669	763	845	953	1,146	1,323	97.8
United Kingdom	...	...	635§	581§	...	...	...
Poland	279	316	360	409	459	520	86.4
Brazil	226	...	413	380	433	...	91.6†
Canada	156	167	182	202	226	...	44.9†
Netherlands, The	135	143	162	186	...	...	37.8‡
Sweden	58	70	80	95	109	...	87.9†
Belgium	81	86	77§	87§	...	...	7.4‡
Austria	30	32	36	42	45	53	76.7
Spain	13	17	16	15	20	...	53.8†

*Value added. †1965–69. ‡1965–68. §Turnover.

BIBLIOGRAPHY

Pharmacopoeias: WORLD HEALTH ORGANIZATION, *Pharmacopoea Internationalis*, 2nd ed. (1967, suppl. 1971); *The British Pharmacopoeia* (1968 and addenda) and *The United States Pharmacopeia XVIII* (1970), mainly as reference works and for examples.

Textbooks: L.M. ATHERDEN, *Bentley and Driver's Textbook of Pharmaceutical Chemistry*, 8th ed. (1969), an excellent source for chemical formulae and synthetic methods of production; A. BURGER (ed.), *Medicinal Chemistry*, 3rd ed., 2 vol. (1970), a comprehensive work that includes the historical development of the subject, and synthetic routes for many drugs; E.P. CLAUS, V.E. TYLER, and L.R. BRADY, *Pharmacognosy*, 6th ed. (1970), includes historical material and information on production of crude drugs, and of glandular extracts and purified hormones therefrom; W. DAVIES, *The Pharmaceutical Industry* (1967), a study of the importance, scope, and economic aspects of the worldwide industry; A. GROLLMAN, *Pharmacology and Therapeutics*, 7th ed. (1970), covers the therapeutic uses of drugs; H.E. SCHULTZE and J.F. HEREMANS, "Analytical Methods in Protein Chemistry," in *Molecular Biology of Human Proteins*, ch. 2 (1966), gives more detail on human blood products; G. SONNEDECKER, *Kremers and Urdang's History of Pharmacy*, 3rd ed. (1963), a detailed treatment of the historical development of pharmacy and pharmaceuticals.

(L.M.At.)

Pharmacological Cults

This article treats the general subject of the religious use of drugs by more or less organized groups, in the past and at the present time. It identifies and lists the main drugs used, and describes their religious functions, purposes, and history. It also considers briefly certain questions raised by the religious use of chemical substances.

NATURE AND SIGNIFICANCE

Though it is a strange idea to the modern religious worshipper, drugs have played an important role in certain religious practices in the history of religions. The ceremonial use of wine and incense in contemporary ritual is most likely a relic of the time when the psychoactive properties of these substances were designed to bring the ancient worshipper into closer touch with supernatural forces. Modern studies of the "hallucinogenic" (productive of hallucinations) or "psychedelic" ("mind-manifesting") drugs have indicated that such drugs in certain individuals under certain conditions release or bring about what they claim to be profound religious experiences of a mystical and transcendental nature. Such experiences involve an immediate subjective experience of ultimate reality or the divine, resulting from the stirring of deeply buried, unconscious, largely nonrational sensitivities. Modern students of pharmacological cults, who have participated in such ceremonies and used the drugs themselves, have often been amazed at the depth of such experiences. R. Gordon Wasson, an eminent researcher in this field, has suggested that the religious impulse itself may have had its origin in such amazement among primitive savages accidentally finding and ingesting psychoactive plants while foraging for food, though this view is not held by most scholars of religion.

The claim that drugs induce religious experience

Whatever the psychological origins of such impulses, they are viewed as religious in nature and have been structured and channelled through cultic forms. Through cultic leaders—such as shamans (spirit-mediums), witch doctors, and medicine men—as well as through tradition, the cults have specified not only how the psychoactive elements should be assimilated but also how they should be gathered and prepared, and also how participants should comport themselves outside of the ceremonies in the practical affairs of living. Western observers of primitive or non-Western cultures, such as missionaries, colonial administrators, travellers, and certain students have often looked on such practices as one might observe the play of children, noting that the primitive societies demonstrate what the observers assume to be superstition and folly. Anthropologists and other scientists, however, who have attempted to participate sympathetically in tribal rituals have reported not only the useful aspects of such practices in primitive society but have also revealed to a wider public information that is of recognized usefulness to science, medicine, religion, and social theory.

The type of drug usually encountered in cultic ceremonies is most generally classifiable as a narcotic. Few of these drugs are true narcotics, however, in the sense of being benumbing. They are usually called hallucinogens when they call attention to reactions similar to hallucinations, or psychedelics, because of their alleged capacity to reveal to the drug-user certain deep-seated truths about himself or to induce transcendental experiences. Most of the psychedelics tend to stimulate the mind rather than to benumb it, though some true narcotics, such as alcohol and opium, both stimulate and stupify the mind at different stages of their physical effect. Most of these drugs come from plants, though newly formed Western cults, composed primarily of

Caucasians in recent times, have made use of the active principles of natural drugs in synthetic form or synthetic analogs of naturally occurring compounds.

TYPES OF DRUGS AND THEIR EFFECTS

Since there are more than 100 plants known to have psychoactive properties, with more being discovered every year, it is impossible to describe all of them here. A few of the major drugs used by cults, however, will be noted. Though these drugs vary greatly in toxicity, there is much overlapping in their effects. Such factors as the personality and expectation of the participants, the setting, the nature of those in charge, and the interpretations of the experiences by various practitioners may have a more significant determining effect on the experiences themselves than the specific properties of the drugs.

At one time or another, such common substances as alcohol, tobacco, coffee, and tea have been used in religious cults for their psychoactive effects, but such use is not common today.

Hemp, mushrooms, cacti, and their derivatives. Probably the most widespread psychedelic plant from which psychedelic properties are derived is Indian hemp, *Cannabis sativa*, which grows all over the world except in very hot and very cold climates. It is used in religious practices in India and Africa (and probably elsewhere) and is also sometimes used illicitly in the United States and Europe. In its most commonly used form, it is known as marijuana, or bhang in India. As hashish, *kif*, or *charas*, it is more concentrated and stronger. But even marijuana is a mild euphoric (the producer of a feeling of well-being) that, on occasion, has been known to induce a profound religious experience. *Tetrahydrocannabinol*, a mixture of related chemicals, contains marihuana's active principle.

Marijuana

Certain mushrooms, of which the best known is *teonanacatl*, "God's flesh," as it was formerly called, are used by cults among the Indians in Latin America, especially in the Sierra Mazateca of Oaxaca province in Mexico. The chief species is *Psilocybe mexicana*, of which the main active principle is *psilocybin*, a fungus-derived (hallucinogenic and organic) compound and a tryptamine recently synthesized, in its chemical composition and activity not unlike LSD (lysergic acid diethylamide-25), a derivative synthesized from the ergot fungus that infects the fruit of rye. The *Amanita muscaria* (fly agaric) mushroom is another psychedelic mushroom that is as yet poorly studied but possibly extremely important since it may possibly be the natural source of the ritual *soma* drink of the ancient Hindus and the Zoroastrians. It is a fungus that grows only in a certain locale, characteristically near birch and pine trees in the cooler temperate climates, and is therefore restricted in its availability. Besides its psychedelic properties it is said both to increase strength and endurance and to be a soporific (sleep-inducing agent). It has the bizarre property of passing through the urine with its psychoactive properties very little diminished and is drunk in this form. This capacity may help in identifying the exact source of *soma*, as will be discussed below.

Plants with somewhat similar psychedelic properties were discovered by Indian tribes of Mexico. The tops of a species of desert cacti known as *Lophophora williamsii*, dried to form the so-called mescal buttons (to be distinguished from the mescal bean, another psychotropic, or mind-expanding, but highly poisonous cultic plant found in the same area) are ingested by widely distributed groups of Indians in Mexico, the U.S., and Canada during night-long ceremonies that have been described and studied by a number of anthropologists. Its chief active principle is an alkaloid called mescaline. Like psilocybin it is reputed, according to cultists, to produce visions and other evidences of a mystical nature such as the direct experience of the unity of all beings, eternity, the holy, the ultimate, and other mystical experiences. Despite claims of missionaries and some government agents that peyote—from the Aztec peyotl ("divine messenger")—is a degenerating and dangerous drug, scientists allegedly have found no evidence of this among the members of the Native American Church, a North American Indian cult that uses peyote in its chief religious ceremony. Worshipped as a god, peyote, like most other psychedelics, is not considered to be addictive and, far from being degenerative, is reputed by cultists and some observers to help promote morale and good behaviour among Indians within this particular cultic framework.

Other psychedelic substances. When the Spanish missionaries came to Mexico in the 16th century they described another psychedelic substance, primarily in derogatory language. The Indians called this substance *ololiuqui* ("round thing" or pellet") and venerated it highly. *Ololiuqui* has been identified as the seeds of the morning glory, mainly of *Rivea corymbosa* and the *Ipomea violacea*. Since the active principles are alkaloids that can be derived from lysergic acid, its properties are closely related to LSD-25. Thus it produces visions and mystical experiences such as are valued in cults using the seeds.

Columbus described the ceremonial snuffing of a powder encountered during one of his voyages to the West Indies. Of the many kinds of snuff, the chief variety was called *yopo*, a powerful psychedelic derived mainly from a tree, Andenanthera (or *Piptadenia peregrina*). This narcotic snuff may be confused with others, as yet not clearly differentiated.

Psychedelic snuffs and drinks

Another psychedelic substance used in South America, especially in the Amazon basin, is a drink variously called *ayahuasca*, *caapi*, and *yajé*, that is produced from the stem bark of the vines *Banisteriopsis cappi* and *Banisteriopsis inebrians*. Natives who use it claim that it has a variety of virtues, including healing powers and producing clairvoyance, among others. This drink has been certified by investigators to produce remarkable visions, often involving the sensation of flying. These visions are thought to be due to the action of *harmine*, a remarkably stable psychedelic compound that the drink contains, or related chemicals.

Other cultic drugs may be mentioned in passing. The *kava* drink, prepared from the roots of the shrub *Piper methysticum*—a variety of the pepper plant—and seemingly more of a hypnotic-narcotic than a psychedelic, is used both secularly and religiously in the South Pacific, especially in Polynesia. *Iboga*, a powerful stimulant and psychedelic derived from the root of *Tabernanthe iboga*, is used by the secret Bwiti cult in the Congo (Brazzaville) and is now said to be making converts much faster than Christianity or Islām there. Coca ("the divine plant of the Incas"), source of the narcotic cocaine, has had both religious and secular use chiefly in Peru. Datura, associated with the jimson weed, is used by both the Huichol tribe in Mexico and by other native peoples in North and South America. Datura is, however, highly toxic and dangerous. A drink prepared from the shrub *Mimosa hostilis* is used religiously in the *ajuca* ceremony of the Jurema cult in eastern Brazil, often participated in by warriors before battle as a source of glorious visions.

GOALS OF PRACTITIONERS

The drugs used by various cults for psychedelic effects were incorporated into the cults for explicit and implicit religious functions and purposes. The effects of the drugs were and are reported to enhance religious experiences similar to those that occur in religious ceremonies or acts practiced among those who do not use drugs. Controlled laboratory experiments and investigations regarding the effects of such drug use that have been performed outside of a culturally determined cultic framework help to make the cults themselves more understandable.

Ecstasy and union with the divine or sacred. The most lofty aim of the cultic use of drugs is the pure delight in what drug users describe as a direct experience of God, ultimate reality, the spirits that preside over one's destiny, or whatever the worshipper may conceive as his object of worship. As a consequence of such worship-experience there may ensue a feeling of

self-transcendence, sometimes through a melting away of the ego boundaries (with consequent loss of sense of self) and even through the terror of death, resulting in a psychological rebirth leading to a feeling of power, freedom, and the release of creative energies. In ritualized form, drugs have been used to enhance the puberty ordeal through which, in certain tribes, a youth is ushered out of the period of childhood and is certified a man. These functions of the drugs as teachers, leading participants through spiritual growth experiences, may be documented by testimonies both of Indian participants and Caucasian members of contemporary Western drug cults.

Drugs as sacraments

As a means of appropriating such experiences, the rites surrounding the assimilation of the drugs become types of sacraments in which the qualities and the gifts of the gods are appropriated, such as when a mushroom is called God's flesh. The visions, self-knowledge, energy, power, and direction reported to be secured from the rite confirm the feeling of the worshipper that he has been in the presence of God or has assimilated some of His powers. Other specific skills and benefits —either of a substantial nature or the products of magical fancy—may be extrapolated from this alleged encounter with the gods, such as sexual attractiveness, skill in hunting, protection in war, or even the ability to transcend war and to love one's enemies, depending in part on the expectations of the cult.

Purification, healing, and divination. Along with the sacramental function of the drug cults there is the concept of purification through drug use. This may take the form of certain ritualistic preparations for the ceremony, the observing of certain taboos for days before it, or as a part of the ceremony itself. Many psychedelic drugs produce nausea and the consequent vomitting may be looked on as a purging of one's faults. In more advanced cults the purification may be seen as the pure and ethical living that should both precede and follow the ceremony. Closely related to the latter may be the devotional function, viewed in the most primitive drug cults as acts of appeasement of the gods for their expected favours, and in the more developed cults as acts directed toward the needs and wishes of the god through the enhancement of the cultic experience itself. The participation in drug taking may also be viewed as an act of devotion in itself.

In some of the more primitive cults the ordinary participant is relegated to a secondary role, for the priest or shaman may be the one who principally ingests the drug and mediates benefits, real or supposed, to the worshipper or suppliant. Such is the case when the shaman mediates the healing function, when the drug is dangerous, or when it is reserved for those with the most prestige. This is also likely to occur when the divination of the answers to important questions are involved; *e.g.*, as whether to set out on a dangerous journey or where to locate lost property.

Divination through use of drugs

Drugs and magic. Sometimes the pharmacological cults do not rise much above the level of witchcraft, and ceremonies involve the expression of the participants' insecurities, anxieties, and hostilities. This is particularly true of cults operating among a marginal, competitive people, as with the *ayahuasca* cult in Iquitos, Peru, a syncretistic cult in which primitive magical beliefs are interwoven with minimal Roman Catholic features. Thus bad luck is looked on as the result of the curse of an envious neighbour. The witch doctor is sought to dispel the curse and, if possible, to return it onto its originator. At other times illness may be interpreted as the result of possession by evil spirits, and the purpose of the ceremony may be exorcism. In all of these cases the vision-producing capacities of the psychedelic drugs, influenced by the assumptions and desires of participants, will be interpreted according to the hopes and fears, prejudices, and suspicions of the participants.

Psychological goals. The literal meaning of the term psychedelic (mind-manifesting) suggests the vast amount of material (feelings, emotions, images, etc.) released by these drugs from the unconscious. This material, related as it is to the psychological needs and past history of the individual, is viewed as both of an uplifting and creative nature and, on the other hand, frightening and destructive. Though these drugs are regarded as dangerous by industrially advanced societies and their medical advisors, cults using such drugs have exhibited durability, and the societies that harbour them apparently show little or no abuse of the drugs, despite open access to the sources of supply. Perhaps these cults are so structured that basic psychological needs are fairly well met and the resulting cultural expressions are directed mainly into creative and constructive ways. Hostilities and anxieties expressed in ritual are sublimated and, therefore, less likely to be acted out. Persons who have taken the drugs maintain that they understand the closeness of feeling that may be engendered among cultic participants and so satisfy that need for comradeship that is one of man's most basic desires, while the excitement of the "trip" (drug experience) may save participants from one of mankind's most pervasive enemies, boredom.

HISTORY OF DRUG USE IN RELIGION

References to the religious use of drugs are scattered through the history of religions, and there is no doubt that the practice is ancient; its origins are lost in pre-history. A Soviet archaeologist has reported presumed ceremonial use of *cannabis* in censers for burning hemp seeds in the frozen 6th to 3rd centuries BC Scythian tombs at Pazarik in the Altai Mountains of Siberia. The ancient Greeks used wine in Dionysian rites, and circumstantial evidence points to the use of a psychedelic substance in the most solemn moments of the secret Eleusinian Mysteries of Hellenistic Greece (*c.* 300 BC–*c.* 300 AD) in the drinking of the *kykeon*—a thick gruel of unknown composition. Both the secular and religious use of the *Amanita muscaria* mushroom in Siberia probably goes back more than 6,000 years, and its cultic use has spread beyond the cool temperate climates where it grows. Evidence of the cultic use of opium in the eastern Mediterranean islands, Greece, and among the Sumerians points to dates as early as 3000 BC, though some of these facts are disputed.

Soma and haoma

One of the pharmacological mysteries is the nature both of *soma*, a sacred drink mentioned in the *Avesta*, the sacred book of the Zoroastrians, and in the Hindu *Ṛg-veda*, and the haoma plant from which it was derived. Its source may be the *Amanita muscaria* mushroom on the basis of Hindu scriptural references to the sacredness of urine drunk as the source of divine insights. But allusions to the twigs and branches of haoma, as with the *barsom* in the Haoma ceremony, suggest other plants, perhaps hemp. It is quite likely, however, that since the mushroom was not available in hot countries, other psychedelic plants may have been substituted until their identity was confused and lost. The mushroom may have been introduced into India by the Aryan invaders from the north, some of whom were from Iran.

There is less evidence for the historical use of drugs in Buddhism or Taoism, though occasionally there is some mention of it. (Some authorities believe that the Buddha may have died from an overdose of psychedelic mushrooms.) In Islām, since Muḥammad prohibited alcohol, there has been much more general use of *Cannabis*. The fanatical sect known as the Assassins, founded by Ḥasan-e Ṣabbāḥ near Baghdad in the 11th century, used hashish. Sensational and fanciful stories, however, told by Marco Polo and others, of the use of hashish to heighten the undoubted homicidal proclivities of the sect, may be untrue.

The use of kava was reported by travellers to the Pacific Islands, notably Fiji, in the 18th century, though its use must go back much farther. The same may be said for the cultic use of drugs like *iboga*, in equatorial Africa, where the discovery of such usage is very recent and where historical records are meagre. Doubtless much remains to be discovered in both of these areas.

Strangely enough, more is known of the history and practices of pharmacological cults, as well as of the varieties of botanical sources of the drugs, in the New World than in the Old. The finding of many little images sculpted in the form of mushrooms in Guatemala almost certainly points to a mushroom cult in the Mayan culture of Central America well before 1000 BC. Columbus reported the use of snuffs, as has been mentioned. The Spanish priest-historian Padre Bernardino de Sahagún reported with disapproval the religious use of mushrooms in Mexico in the 16th century, and there were reports of the widespread use of psychotropic mushrooms during the coronation of the Aztec emperor Montezuma in 1502. Suppressed by the Spaniards, the remnants of these once proud cults remain in the back country of Mexico, particularly in the province of Oaxaca. In 1955 R.G. Wasson and his companions were the first white people invited to eat the sacred mushroom at one of the night ceremonies. Consequent publicity has recently led to an influx of visitors seeking mushrooms and has also attracted the attention of Mexican federal police. Despite this attention, however, the mushroom religion is still a very living one, even though the ceremonies are now carried on less openly.

Spanish reports of about the same era as those covering Mexico allude to cultic use of various drugs in South America. Sahagún also mentioned the religious use of *ololiuqui* and peyote in his reports. The peyote cult continued to be harassed by missionaries and government agents alike toward the end of the 19th century when numerous Indian cults began to come together in an international body comprising Canadian, American, and Mexican Indians. This federation was finally incorporated in the State of Oklahoma in 1918 as the Native American Church. Despite many antipeyote laws, these Indians have steadfastly maintained their right to worship in their own way. They have suffered imprisonment and have fought these laws in the courts. For the most part the courts have upheld this right.

Use of peyote in the Native American Church

PHARMACOLOGICAL CULTS TODAY

Partly because of the recent rise of interest in psychedelic drugs, as well as the accessibility of the peyote religion, the Native American Church has been studied more thoroughly than any other pharmacological cult, mainly by American anthropologists who have written full reports of its practices. Since earlier times it has developed into a syncretic religion, differing from tribe to tribe in minor details and in the degree to which ancient Aztec practices have assimilated Christian elements. The ritual surrounding ingestion of peyote is a highly symbolic night-long ceremony in the charge of an experienced Indian. Whites privileged to attend the service universally speak of its dignity and impressiveness.

Peyote itself is highly venerated and spoken of sometimes as a gift of God, sometimes as a god itself. It is looked on partly as the possessor of magical qualities of protection and healing, and the revealer of hidden knowledge, and partly as a guide that motivates, strengthens, and guides the Indians. There is as much controversy among the Indians themselves over peyote as over LSD among whites. But the cult, claimed by some adherents to number over 300,000 members, seems to be spreading. Contrary to the assertions of missionaries, who find them hard to convert, the peyote Indians, doubtlessly influenced by white religions around them, are reported to be superior to the nonpeyote Indians in achieving their aims of brotherly love, hard work and self-reliance, family responsibility, and abstinence from alchohol.

Because psychedelic drugs, both natural and synthetic, tend to evoke an experience spontaneously recognized by many as religious and therefore of supreme value to the users, small communities of seekers have grown up wherever the drugs are generally used, most recently among whites in large centres of population. Due to what they consider to be harassment by unsympathetic elements in the cities, many such communities have moved to rural or wilderness areas. Often the harassment continues, while emphasis on spontaneity leads to poor organization. Consequently the histories of many of these communities are brief. An example is the League for Spiritual Discovery founded by Timothy Leary in 1966.

League for Spiritual Discovery and Church of the Awakening

A more durable cult is the Church of the Awakening started by two physicians, John and Louisa Aiken, in Socorro, New Mexico, in 1958. Incorporated in that state in 1963, it had a country-wide membership of 400 in 1970. The Aikens administered both peyote and mescaline sacramentally and with complete safety until prohibited by subsequently passed laws. Their application to the U.S. Department of Justice for permission to use peyote religiously was denied.

RELIGIOUS CONSIDERATIONS AND ANALYSES

It is a matter of controversy whether drug-induced religion is true religion or merely psychological phenomena. In a Harvard University doctoral experiment, W.N. Pahnke, M.D., in support of Aldous Huxley, demonstrated (1963) that drugs can induce experiences indistinguishable from certain mystical states, which the American philosopher William James (1842–1910) declared to be the heart of personal religion, partly as the result of a self-experiment with nitrous oxide. R.C. Zaehner, one of the few critics of drug-induced religion to have experimented on himself, claims that drugs may trigger a natural type of mysticism but not a true theological variety. W.T. Stace, on the other hand, acknowledges that the phenomenology of at least some drug experiences has been truly mystical.

Even if one accepts such a personal type of religion as genuine, however, there still remains the question of whether or not it is effective ethically and socially, two important aspects of most religions. This question may be answered variously, depending in part on the practices and beliefs of the particular cult studied and the context within which the drug is taken. According to some of its adherents, the peyote cult seems to have had mostly beneficial effects on its participants, though this has not been the result in some of the drug communes that apparently were hastily assembled for the sole purpose of ingesting drugs. The psychedelics involved in such communes often tend to induce transcendental experiences of unity and the holy that incline the participant toward loving relationships. Without structure, clearly defined aims, and discipline, however, any cult may eventually disintegrate or fall prey to manipulation, and this has been the experience of some drug cults. The cults that have achieved some measure of success often find the initial drug experiences valuable as a means of introducing their followers to a wider world of religious values. After initial use of drugs, the tendency may be either to use the drugs sparingly, or to give them up altogether in favour of the discipline of meditation or some other means of experiencing mystical religion.

Members of the general community tend to disapprove of drug cultists. The average citizen, accustomed to established ecclesiastical norms and forms, often feels vaguely threatened by transcendental or ecstatic expressions of religion. As a result of such critical attitudes pharmacological cults, particularly those in close contact with the larger, more conventional, or more established religious and social groups, are often harassed, legally or otherwise. Furthermore, in Western culture the word drug itself tends to connote danger and degeneracy, particularly among the general public, and even among the ranks of scholars of religion. Though few scholars of religion have had any firsthand experience with drugs, they often adopt a normative approach and tend to write off any veneration of these substances, which drug cultists consider to be sacramental, as magic, superstition, or witchcraft. The worshipper in a drug cult, on the other hand, tends to view drugs as a means of grace.

Nativistic pharmacological cults appear to be growing,

"Hippie" culture

at least in some areas, such as in the case of the Bwiti, or of the peyote religion. Besides such nativistic cults an amorphous movement popularly known as the "hippie" culture, both in the U.S. and elsewhere, has been growing since the middle of the 20th century. Though not acknowledged as a religious movement by the average Western churchgoer, the hippie movement is regarded by some persons as a very dynamic religious movement in terms of its active search for a mystical approach to life and the experience of the divine or sacred through meditation. In this search the hippie movement is influenced, to a great extent, by its use of drugs.

Due to the religious drug-users' demand for the right to practice religion in their own way, a legal crisis has been developing in the U.S. and elsewhere. This legal battle apparently has been won by one group, the Native American Church. Opponents, however, generally claim that to allow anyone to use drugs for such religious purposes is dangerous to society, and that the right to use drugs under such a guise would simply become an occasion for irresponsible indulgence. Many of the drug users, however, have performed what history shows is a characteristic act of those who are deeply committed to any religious ideal or practice—they have defied the law and thus have suffered the consequences.

BIBLIOGRAPHY. W.H. CLARK, *Chemical Ecstasy* (1969), a description and discussion of drug-induced religion; R.H. BLUM *et al., Drugs I: Society and Drugs* (1969), historical and social aspects of drug use; W. HEDGEPETH and D. STOCK, *The Alternative* (1970), an excellent journalistic account of contemporary communes; D.H. EFRON, B. HOLMSTEDT, and N.S. KLINE (eds.), *Ethnopharmacologic Search for Psychoactive Drugs* (1967), a review by leading international experts, including material on cults; R.E. SCHULTES and A. HOFMANN, *The Botany and Chemistry of Hallucinogens* (1971), an authoritative work in the indicated fields; W. LA BARRE, *The Peyote Cult*, enl. ed. (1969), and J.S. SLOTKIN, *The Peyote Religion* (1956), detailed and thorough references on the peyote cult; V.P. W.T. STACE, *Mysticism and Philosophy* (1960), a very clear account of mysticism that alludes to drugs; and R.G. WASSON, *Mushrooms, Russia, and History*, 2 vol. (1957), and R.G. WASSON, *Soma: Divine Mushroom of Immortality* (1968), required readings for those who wish to be thoroughly informed on mushroom cults; R.C. ZAEHNER, *Mysticism, Sacred and Profane* (1957), mainly a criticism of the views of Aldous Huxley on drugs.

(W.H.C.)

Pharmacy

Pharmacy is the science and art concerned with the collection, preparation, and standardization of drugs. Its scope includes the cultivation of plants that are used as drugs, the synthesis of chemical compounds of medicinal value, and the analysis and standardization of medicinal agents. The science that embraces all available knowledge of drugs with special reference to the mechanism of their action in the treatment of disease is pharmacology. Obviously this broad science has many subdivisions, such as toxicology (the study of poisons) and therapeutics (the use of drugs in the treatment of disease). According to the foregoing description, pharmacy is one of the subdivisions or specialties of pharmacology. Pharmacists are responsible for the manufacture of the dosage forms of drugs, such as tablets, capsules, and sterile solutions for injection. They compound physicians', dentists', and veterinarians' prescriptions for drugs.

Historical development. The beginnings of pharmacy are ancient. When early man expressed a juice from a succulent leaf to apply to a wound, he was practicing this art. In the Greek legend, Asclepius, the god of the healing art, delegated to Hygieia the duty of compounding his remedies. She was his apothecary or pharmacist. The physician-priests of Egypt were divided into two classes: those who visited the sick and those who remained in the temple and prepared remedies for the patients.

In ancient Greece the art of healing recognized a separation between the duties of the physician and those of the pharmacist. The same separation prevailed among the Roman practitioners. During the Middle Ages in Europe, the separation of the two groups became less distinct, and many physicians undertook the preparation of their own medicines. The Moorish influence in Europe, however, again brought about the practice of separate duties for the pharmacist and physician. The trend toward specialization was later reinforced by a law enacted by the city council of Bruges, Belgium, in 1683, forbidding physicians to prepare medications for their own patients. In France an office of apothecary general in the army was established in the middle of the 18th century. During the same period in America, Benjamin Franklin took an important step in keeping the two professions separated when he appointed an apothecary to the Pennsylvania Hospital.

Ancient and medieval pharmacy

Education in pharmaceutical science. The history of pharmaceutical education has closely followed that of medical education. As the training of the physician underwent changes from the apprenticeship system to formal educational courses, so did the training of the pharmacist. The first college of pharmacy was founded in the United States in 1821 and is now known as the Philadelphia College of Pharmacy and Science. Other institutes and colleges were established soon after in the United States, Great Britain, and continental Europe. Later, many universities organized schools and colleges of pharmacy within their courses of instruction. Colleges of pharmacy as independent organizations or as schools of universities now operate in most developed countries of the world.

The course of instruction leading to a degree in pharmacy (bachelor of science in pharmacy) was extended from four to five years in 1960. The first and frequently the second year of training, embracing general education subjects, are often provided by a school of arts and sciences. Many institutions, in addition, offer graduate courses in pharmacy and cognate sciences leading to the degrees of master of science and doctor of philosophy in pharmacy, pharmacology, or related disciplines. These advanced courses are intended especially for those who are preparing for careers in research, manufacturing, or teaching in the field of pharmacy.

Several schools of pharmacy have now adopted a six-year professional course leading to the degree of doctor of pharmacy. This professional training includes many subjects common to the medical curriculum and involves training on the hospital wards. In this service the professionally trained pharmacist is expected to give advice to the physician in the techniques of administering medication and the possible interaction of drugs in the patient, along with expected side effects.

Since the treatment of the sick with drugs encompasses a wide field of knowledge in the biological and physical sciences, it is obvious that an understanding of these sciences is necessary for adequate pharmaceutical training. The basic five-year curriculum in the colleges of pharmacy of the United States, for example, embraces physics, chemistry, biology, bacteriology, physiology, pharmacology, and many other specialized courses such as dispensing pharmacy. As the pharmacist is engaged in a business as well as a profession, special training is provided in merchandising, accounting, computer techniques, and pharmaceutical jurisprudence. All other countries requiring licenses to practice offer the same basic curriculum with minor variations.

The basic five-year curriculum

Legislation. Before one is permitted to practice pharmacy in those countries in which a license is required, an applicant must be qualified by graduation from a recognized college of pharmacy, meet specific requirements for experience, and pass an examination conducted by a board of pharmacy appointed by the government. The passing of this board examination carries with it the legal right to practice pharmacy. The holder is then designated a registered or licensed pharmacist.

Pharmacy laws generally include the regulations for the practice of pharmacy, the sale of poisons, the dispensing of narcotics, and the labelling and sale of dangerous drugs. The pharmacist sells and dispenses drugs within the provisions of the food and drug laws of the country in

which he practices. These laws recognize the national pharmacopoeia (a treatise on products used in medicine, their purity, dosages, and other pertinent data) as the standard for drugs. The pharmacopoeias of the different nations are compiled and published according to the respective legal national procedures. The World Health Organization of the United Nations published the first volume of the *Pharmacopoea Internationalis* in the early 1950s. Its purpose is to standardize, as far as possible, the medication with drugs throughout the world. There are two volumes and a supplement. This pharmacopoeia is intended to supply standards, strengths, and nomenclature for those countries that have no national pharmacopoeia: volume 1 contains monographs for basic chemicals and drugs of plant origin; volume 2 contains monographs for medicinal agents and specifications for dosage forms.

Organizations. There are numerous national and international organizations of pharmacists. In Britain the Pharmaceutical Society of Great Britain, established in 1841, is typical of the pharmaceutical organizations. In the United States the American Pharmaceutical Association, established in 1852, is a society that embraces all pharmaceutical interests. Among the international societies is the Fédération Internationale Pharmaceutique, founded in 1910 and supported by 48 national societies, for the advancement of the professional and scientific interests of pharmacy on a worldwide basis. The Pan American Pharmaceutical and Biochemical Federation includes the pharmaceutical societies in the various countries in the Western Hemisphere.

International organizations

There are also other international societies in which history, teaching, and the military aspects of pharmacy are given special emphasis.

Research. Pharmaceutical research, conducted mainly in the schools of pharmacy and in the laboratories of the pharmaceutical manufacturing houses, embraces the organic chemical synthesis of new chemical agents for use as drugs and is also concerned with the isolation and purification of plant constituents that might be useful as drugs. Research in pharmacy also includes methods for assay and standardization of medicinal agents. These phases of research in pharmacy are primarily an application of the science of chemistry to drug products, but the pharmacist also engages in research peculiar to the profession. This research involves formulation of dosage forms of medicaments, a study of their palatability, methods of standardization, and their stability. Another facet of pharmaceutical research that has attracted wide medical attention is the problem of the "availability" to the body (bioavailability) of various dosage forms of drugs. Exact methods of determining levels of drugs in blood and organs have revealed that slight changes in the mode of manufacture or the incorporation of a small amount of inert ingredient in a tablet may diminish or completely prevent its absorption from the gastrointestinal tract, thus nullifying the action of the drug. Ingenious methods have been devised to test the bioavailability of pharmaceutical dosage forms. Although these in vitro, or test-tube, methods are useful and indicative, the ultimate test of bioavailability is the response of the patient to dosage form of the drug.

The results of research in pharmacy are usually published in journals, such as the *Journal of Pharmacy and Pharmacology* (London), the *Journal of the American Pharmaceutical Association* and the *Journal of Pharmaceutical Sciences* (Washington), the *American Journal of Pharmacy* and the *American Journal of Hospital Pharmacy* (Philadelphia), and the *Pharmaceutica Acta Helvetiae* (Zürich).

BIBLIOGRAPHY. ARTHUR OSOL *et al.* (eds.), *Remington's Pharmaceutical Sciences*, 14th ed. (1970), a comprehensive treatise on pharmacy and its allied sciences; *The United States Dispensatory and Physician's Pharmacology*, 26th ed. (1967), an authoritative and descriptive treatise on drugs and their actions; J.C. KRANTZ, JR. and C. JELLEFF CARR, *The Pharmacologic Principles of Medical Practice*, 7th ed. (1969), deals with the mechanism of drug actions in the treatment of diseases; J.B. SPROWLS and H.M. BEAL (eds.), *American Pharmacy*, 6th ed. (1966), a commentary on the training and service of the pharmacist and his role in the health program of the community.

(J.C.K.)

Phase Changes and Equilibria

The existence of a chemical substance, whether element or compound, in different forms or physical states is well-known. Water, for example, is commonly recognized in the forms of solid (ice), liquid, and gas or vapour. Besides these three familiar forms, various others are known. All are called phases of matter. Many substances, for example, in addition to their liquid and vapour phases occur in several solid forms, defined in terms of their crystal structure (*e.g.*, rhombic and monoclinic sulfur, white and violet phosphorus, hexagonal and rhombohedral ice). Less familiar phases of matter include liquid crystals, which have some properties of solids and some properties of liquids, and the supercritical fluid state, which has properties of both gases and liquids.

Varieties of phase forms

GENERAL CONSIDERATIONS

When matter in one form is transformed into another form, a phase change is said to take place. Whether the sample consists of a complicated mixture of many substances or of a single substance, the transformation may be a reversible one. To state the simplest case, in a reversible system, if form *A* is changed into form *B* by raising the temperature, form *B* will be transformed into form *A* by lowering the temperature, assuming that all other factors are constant. If both forms are present, the temperature will determine which form changes into the other. Because the energy of an atom is distributed over a wide spectrum, the whole of a sample of a solid substance, for example, will not be transformed instantly to liquid at the melting point. In fact, while some of the molecules or atoms are changing from solid to liquid, others will be changing from liquid to solid. In any time intervals, the amount of phase *A* changing into phase *B* can be compared to the amount of phase *B* changing into phase *A*, and the proportions can be expressed in terms of the actual number of molecules involved—a quantity that can be calculated by using atomic weights of the elements and actual weights of the substances. It is possible, at a given temperature, to find that both *A* and *B* forms persist indefinitely in a fixed proportion. That is, the rate of change of *A* into *B* equals the rate of change of *B* into *A*, thus creating a system of phases that is in equilibrium. If even the slightest amount of heat is added to the system, the proportions will change until again the rate in one direction equals the rate in the other. Equilibrium is a frequently unwanted condition because it prevents reactions from going to completion and, for example, lowers yields of desired products. Phase changes are of many kinds, and systems in equilibrium are also of various types. The few variable forces that bring about phase changes in any system include temperature, pressure, and concentration of substances.

Phase number. When one substance or a mixture of two or more substances in a given proportion is brought to equilibrium at a specified temperature and pressure, a phase or combination of such phases of matter will result. Such a system, as it is called, is described by enumerating the phases that are present, the composition but not the quantity of each phase, and the temperature and pressure of the system. A phase is defined as a homogeneous, mechanically separable, portion of a system, all parts of the phase being identical in intensive properties; *i.e.*, those qualities that characterize a phase, such as density and refractive index, but do not depend on the actual quantity of phase present. Although a phase must be homogeneous, it need not be continuous: crushed ice, for example, is a single phase since the intensive properties of one piece of ice are indistinguishable from those of every other piece. A system consisting of one substance throughout may have one, two, or three phases, but a complex system may have many liquid and solid phases, but normally only one gas phase because gases mix in all proportions. Thus water, held at 25° C (77° F) under a pressure of one atmosphere, will consist entirely of liq-

Definition of phase

uid water, but if the temperature is raised above 100° C or 212° F (the pressure still being maintained at one atmosphere) it will consist only of vapour. Liquid and vapour phases can coexist only at 100° C when the pressure is one atmosphere. If approximately equal weights of the compound diethyl ether and water are mixed and brought to equilibrium at 25° C (77° F) and one atmosphere, they will not mix completely and two liquid phases will result, one of which is 1.3 percent water while the other is 94 percent water. Similarly, under appropriate conditions, one can form, by mixing common salt (formula NaCl) and water (formula H_2O), a system consisting of four phases: a vapour (H_2O), a liquid (NaCl dissolved in H_2O), and two solids, pure water as ice and a compound of salt and water with twice as many water molecules as salt (formula $NaCl \cdot 2H_2O$). The equilibria in the aforementioned systems are essentially physical, but chemical equilibria are common, as in the system formed from solid calcium carbonate ($CaCO_3$) that, under suitable conditions, partly breaks down into gaseous carbon dioxide (CO_2) and solid calcium oxide (CaO); the system consists of two solid phases—one being calcium carbonate and the other calcium oxide—and one gas phase carbon dioxide, in proportions that remain constant until temperature, pressure, or volume are changed. All such equilibria are included in the term phase equilibria. The study of phase equilibria concerns the problem of determining what phases coexist at equilibrium under what conditions of temperature or pressure for a given combination of substances in given proportion. Equilibrium involving more than one phase is called heterogeneous equilibrium. The number of phases in a system is designated as beta (symbol β).

Distinction between components and substances present

Component number. The number of components, or the order, of the system, is designated as alpha (symbol α). Alpha is not necessarily the same as the number of substances used to form the system but is rather the smallest number of independently variable (explained directly below) substances required to account for the algebraic composition of each and every phase present. In this sense, a system containing only water has only one component, there being no recognition of the presence of molecular species other than water, such as the hydronium ion (H_3O^+) and the hydroxide ion (OH^-), because, although both ions are always present, their concentrations are not independently variable. In the above-mentioned ether–water example, the system has two components, ether and water, because both are independently variable and both are needed to describe the composition of the two phases formed. In the system formed from calcium carbonate, however, even though there are three substances present, there are only two components, calcium oxide and carbon dioxide, because the composition of the $CaCO_3$ phase can be described as 50 percent CaO and 50 percent CO_2, that of the CaO phase as 100 percent CaO, 0 percent CO_2, and that of the gas phase as 0 percent CaO, 100 percent CO_2, these being mole percentages (one mole of a substance is its molecular weight in grams). Other selections of components could be made—*e.g.*, CaO and $CaCO_3$—but these would require negative percentages. In another example, barium sulfate ($BaSO_4$) and sodium carbonate (Na_2CO_3) react when melted together to produce barium carbonate ($BaCO_3$) and sodium sulfate (Na_2SO_4), all solids. Thus, the system involves four substances: $BaSO_4$, Na_2CO_3, $BaCO_3$, and Na_2SO_4, related through the reaction $BaSO_4 + Na_2CO_3 \rightleftharpoons BaCO_3 + Na_2SO_4$, but the system has only three components because the composition of all four can be described in terms of percentages of only three of them. It is, however, not the identity of the components themselves but the actual number of components that is of importance in the discussion immediately below.

The value of alpha corresponding to a given combination of substances can depend on the range of conditions over which it is studied. Thus, a system of oxygen (molecules containing two atoms, O_2) and ozone (molecules containing three atoms of oxygen, O_3) in any proportions has two components under conditions in which neither is converted to the other, but becomes a one-component system when they interact according to the equation $3O_2 \rightleftharpoons 2O_3$. When the reaction is at equilibrium, according to the laws of chemical reaction the concentrations of O_2 and O_3 are no longer independently variable; *i.e.*, one cannot change without the other also changing. A system consisting of solid ammonium chloride (NH_4Cl) in equilibrium with its vapour has one component, even if the vapour dissociates into ammonia (NH_3) and hydrochloric acid (HCl), provided the NH_3 and HCl are present in equal molar quantities. If, however, excess of either NH_3 or HCl is present, alpha will equal two. Systems with alpha (α) = one, two, three, or four are said to be unary, binary, ternary, or quaternary, respectively.

Degrees of freedom. The description of any one system is facilitated by a statement of the number of degrees of freedom, or its variance, symbolized as F, that it possesses. This is the (minimum) number of variables of temperature (symbol T), pressure (symbol P), and concentration that must be fixed in order to define the system—*i.e.*, in order to fix the intensive properties of all the phases. At equilibrium, there is only one pressure variable because all the phases are at the same total pressure; there is only one temperature variable because all the phases are at the same temperature. The maximum number of independent concentration variables for each phase will be $\alpha - 1$ because fixing $\alpha - 1$ concentrations automatically fixes the concentration of the remaining component. In a system containing only water vapour, for example, the properties of the vapour will not be defined until both T and P are fixed, so that F equals two. There is no concentration variable when α equals one. For a system containing liquid water and water vapour in equilibrium, only one of the two variables, T and P, can be fixed arbitrarily—fixing one fixes the other—therefore F equals one. If ice, liquid water, and water vapour are all present, no variables can be fixed arbitrarily—the latter are beyond any control—and therefore $F = 0$. For a system consisting of water vapour and liquid water saturated with solid salt (NaCl), fixing any one of the following will define the system: T, P, concentration of salt in the liquid solution, concentration of water in the liquid solution; F is therefore unity. (The concentrations of NaCl in the H_2O vapour and of H_2O in the solid NaCl are not determinative variables because experimentally none of them is present.)

The variables of a system

Phase rule. The value of F is readily calculated by means of a rigorous thermodynamic equation known as the phase rule, which states that the number of degrees of freedom equals the number of components minus the number of phases, plus two, or $F = \alpha - \beta + 2$. Systems for which F equals zero, one, two, or three are invariant, univariant, bivariant, and trivariant, respectively.

The phase rule was derived by a U.S. physicist, J. Willard Gibbs, in some papers on thermodynamics published from 1873 to 1878, and represents an important milestone in the history of science. Its usefulness for chemists in the systematization of heterogeneous equilibria was brought to light two or three decades later. Its validity is restricted to systems in real equilibrium and assumes that the intensive properties of every phase of a system are determined only by its temperature, pressure, and composition. The influence of such factors as surface exposure, electric, magnetic, and gravitational fields is assumed to be negligible.

PHASE DIAGRAMS

Derivation of phase diagrams

The subject of phase equilibria depends on graphical representation. With any system, the first step of experimental study is to establish what phases result under a variety of specific conditions by bringing together various proportions of the given substances. It is not yet possible to predict such information, at least in a quantitative sense. The results of the experiments are then plotted and embodied in phase diagrams that are graphical representations in two or more dimensions of the relationships among the variable factors.

Because phase changes are so much a part of the physical world, phase diagrams are pertinent to everyday existence: the melting of ice, the boiling of water, the forma-

tion of fog, the setting of cement, for example, are well known phase changes, the details of which can be predicted from phase diagrams. In the laboratory, processes such as extraction, crystallization, distillation, precipitation, the use of freezing mixtures, and the identification of substances all involve phase changes describable by phase diagrams. Many industrial operations—such as the manufacture of ceramics, alloys, glass, and salts—depend on knowledge of phase diagrams. Finally, geological studies of the history of minerals in igneous, sedimentary, and metamorphic rocks in Earth and moon surfaces are based on previously established phase equilibria.

While phase diagrams ideally describe only systems at equilibrium, the attainment of equilibrium is often a very slow process. In the interpretation of phase behaviour, therefore, the application of phase diagrams must be made with caution and an awareness that time may be an additional important factor.

Unary (one-component) systems. When only one component is present, there are considered to be no concentration variables, every possible phase being 100 percent of that component. The two variables of pressure and temperature suffice to determine the state of the system. Thus all possible phase relations can be shown by a graph in which pressure is plotted against temperature (see Figure 1). Since the number of components alpha

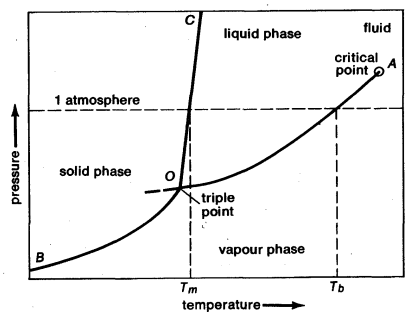


Figure 1: Phase equilibria for the solid, liquid, and vapour states of a substance (see text).

(α) equals one, and the number of degrees of freedom (F) equals two, one, and zero, by the phase rule, when the number of phases, beta (β) equals one, two, and three, respectively, systems with only one phase are bivariant (have two independent variables) and are represented by areas and defined by the two variables; two-phase systems are univariant (have one independent variable) and are represented by lines; three-phase systems are invariant (have no variables) and are represented by points. The lines represent the limits of the condition of bivariance. Three such lines (OA, OB, OC) meet at an invariant point, such as point O, known as a triple point. The substance thus will consist of a single liquid phase if pressure and temperature are such that the plotted point lies in the area partly enclosed by COA, designated liquid; the substance will consist of a single solid phase if the plotted point lies in the area partly enclosed by BOC; similarly, if the plotted point lies in the area AOB the substance will exist as a single vapour phase. Liquid and vapour phases can coexist only when pressure and temperature are such as to lie on OA, the vaporization curve. Solid and vapour phases can coexist only along OB, the sublimation curve (*i.e.*, the solid phase transforms reversibly into the gas without passing through a liquid phase), and solid and liquid can coexist only along OC, the fusion curve. Solid, liquid, and vapour can coexist only at the unique pressure and temperature of point O. AO can be extended for a short distance into the region for solid and, as such, represents a metastable (less stable) liquid–vapour equilibrium that, on being disturbed, yields the more stable solid-vapour equilibrium. The vaporization curve terminates at A, the critical point, beyond which liquid is no longer distinguishable from vapour, a condition called the supercritical fluid state. The sublimation and fusion lines, on the other hand, are often found to terminate in other triple points. OA gives the variation of the vapour pressure of the liquid with temperature, OB the variation of the sublimation pressure of the solid with temperature, and OC the variation of the melting point of the solid with pressure. The one-atmosphere dotted line intersects OC at temperature T_m, and OA at temperature T_b, T_m and T_b being, respectively, the normal melting and boiling points. (If, as with carbon dioxide, point O is above the one-atmosphere line, there are no such points—only a normal sublimation point.) OC is nearly vertical, because the melting point is only slightly sensitive to changes in pressure. The slopes of any of these univariant transition lines at any point thereon are governed by thermodynamic principles embodied in the Clausius–Clapeyron relation, which states that if ΔH is the heat absorbed when a given mass of one of the two phases is transformed to the other phase (under equilibrium conditions at that point), and ΔV is the corresponding increase in volume, then the ratio of a small change in pressure (dP), to a small change in temperature (dT) will equal the heat absorbed divided by the product of the temperature and the volume change, thus: $dP/dT = \Delta H/T\Delta V$. For transitions along OA and OB, ΔH and ΔV have the same sign, so the lines have a positive slope. This is also true usually along OC, but for exceptional cases, including water, ΔH and ΔV have opposite signs and, therefore, dP/dT is negative. Thus the melting point of ice is lowered by an increase in pressure. Point O for water is at 4.58 torr (one torr equals the pressure of $\frac{1}{760}$ of the atmosphere at standard conditions) and 0.0100° C (32.0180° F). (The melting point of ice, measured in the presence of air at ambient or normal atmospheric pressure, is 0.00° C [32.00° F], because the pressure is about 760 torr and because air is dissolved in the water.)

Thermodynamics of univariant transition lines

Polymorphic substances. Many substances are polymorphic—*i.e.*, they exist in more than one crystalline form—a property well illustrated by water. When such crystalline forms are stable under a given pressure and temperature, each will be represented by an area, as in Figure 2, that shows several polymorphic forms of ice (the

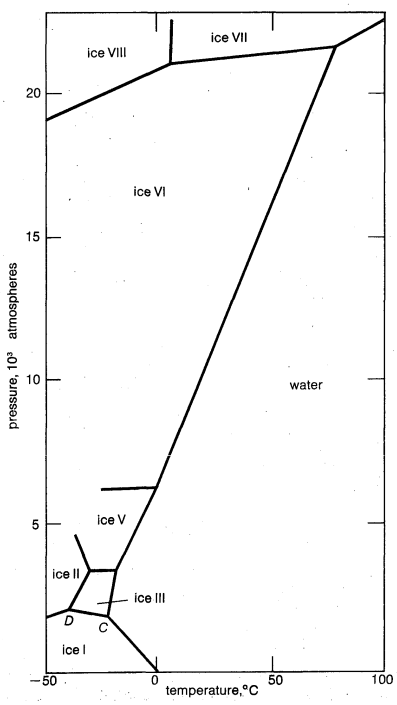


Figure 2: Equilibrium relations for high-pressure forms of ice.

various forms being identified by roman numerals) and is an extension to higher pressures of Figure 1 as applied to the substance water. Two such forms occupying adjacent areas may be said to be enantiotropic; *i.e.*, they are directly interconvertible and can coexist stably at pressures and temperatures anywhere on the line separating the two areas. In contrast, some substances, such as phosphorus,

exist in solid forms (*e.g.*, white phosphorus), that are not known to be stable under any conditions; on standing, they change spontaneously to a more stable form (*e.g.*, white phosphorus changes to red phosphorus). Such metastable forms are said to be monotropic, but the terms enantiotropic and monotropic are both chiefly of historic interest. Figure 2 shows the variety of ice forms existing under appropriate conditions of temperature and pressure. Ice IV, known only for heavy water, is metastable and is not shown. Ice VII has a density of nearly 1.7 grams per cubic centimetre and requires pressures of at least 21,700 atmospheres to be stable.

Other well-known substances exhibiting polymorphism are sulfur, carbon, iron, tin, silica, carbon tetrachloride, *p*-dichlorobenzene, ammonium chloride, and potassium nitrate. Rhombic and monoclinic sulfur are stable below and above 95.5° C (203.9° F), respectively, at one atmosphere pressure, but this temperature (at which they coexist) increases with pressure along a line similar to a fusion line and called a transition line. Crystalline carbon is known in both graphite and diamond forms, each having its own area of stability and bearing to each other a relation similar to that of ice I and ice III (Figure 2), in which ice I is replaced by graphite and ice III by diamond. With carbon, however, *DC* slopes upward (instead of down as with water) to the right, rising from approximately 22,000 atmospheres at room temperature to 200,000 atmospheres at 4,000° C (7,200° F). Thus diamonds can be formed from graphite only under immense pressures. (The temperature must also be high enough for an appreciable transition rate.) Diamonds remain unchanged at ambient pressure and temperature only because their conversion to graphite is extremely slow.

Liquid crystals. Finally, mention must be made of certain compounds, such as *p*-azoxyanisole, the solid form of which melts sharply at one temperature to give a milky, "liquid crystal" or mesomorphic phase that is fluid but anisotropic (with different properties in different directions) and, at a higher temperature, becomes clear and isotropic (uniform in all directions). Such substances, known since 1888, have only recently become of wide interest because of their unique properties.

Variables of binary system

Binary (two-component) systems. When two components are present there are, in addition to the pressure and temperature variables, two concentration variables, only one of which is independent because fixing the percent of one component in a phase fixes that of the other. Thus, the complete description of the phase behaviour requires a three-dimensional model, referred to three axes for ***P***, ***T***, and composition, respectively. A typical portion of such a model for a system of two volatile liquid components, *A* and *B*, that are mutually soluble (miscible) in all proportions, is shown in Figure 3. Along *AB* is plotted the concentration or composition in weight or mole percent (points *A* and *B* representing pure *A* and *B*, respectively). Because alpha equals two, then *F* equals three, two, one, or zero according as beta equals one, two, three, or four. Systems consisting of a single phase are thus represented in the model by volumes. If one regards Figure 3 as depicting a system that, within a limited range of pressure and temperature, shows a liquid phase and vapour phase or both, then *ab* on the *A* face of the model is part of the line *OA* of Figure 1 for component *A*, and *cd* is part of the line *OA* of Figure 1 for component *B*. Within the model lie two curved surfaces that meet along *ab* and *cd* and that, between *ab* and *cd*, envelop a region of space such that a section (slice) through the model at any temperature *T'*, called an isothermal (uniform temperature) section, gives a lenticular area (P_A to P_B). The space (volume) below the lower surface represents systems consisting entirely of vapour and containing both *A* and *B*, and the volume above the upper surface systems consisting entirely of liquid containing both *A* and *B*. The space between the surfaces represents systems consisting of both liquid and vapour phases. Every point on the upper surface is joined by an imaginary line to a corresponding point on the lower surface, these lines—called tie lines—being parallel to the *AB* axis. The two points that terminate each tie line give the values of pressure, temperature, and composition of the two coexisting phases. Because beta equals two in this region, *F* equals two. Regions of space crossed by such tie lines may be called ruled volumes. If Figure 3 covered a wider range of temperature and pressure, then other such pairs of curved surfaces could appear. Intersection of such pairs give ruled surfaces, which, in turn, intersect along lines parallel to *AB*, the details of which are available in extended works on the subject.

It is usually more convenient to consider sections of the solid model rather than the model itself. Figure 4, left,

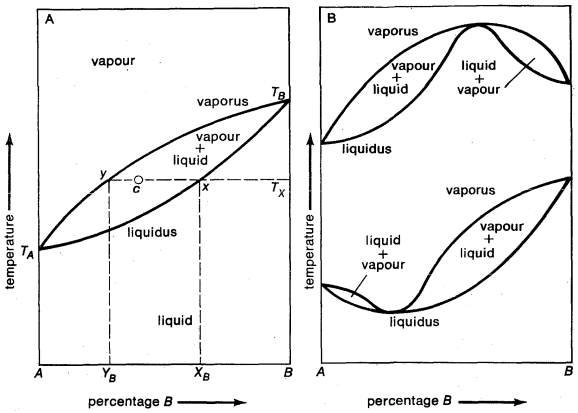


Figure 4: Three types of binary liquid–vapour relations at constant pressure (boiling-point diagrams) the curves in B have maxima and minima, respectively (see text).

is an isobaric section of Figure 3 taken at pressure *P'*, implying that the experimenter has fixed the pressure (throughout the system) at the value *P'*, conveniently one atmosphere. What was a volume in Figure 3 becomes an area in Figure 4, left, and what was a surface becomes a line. Thus, what were trivariant volumes are now areas, labelled "vapour" and "liquid"; what was a ruled volume is now a ruled area, marked "vapour + liquid," crossed by tie lines such as *yx* joining every point on the upper line (vaporus) to a corresponding point on the lower line (liquidus). With pressure fixed arbitrarily (isobaric) in this way, and composition and temperature variable, one-phase systems are said to be isobarically bivariant, two-phase systems isobarically univariant, etc. An isobarically invariant system, obtained by taking an isobaric section of the appropriate solid model (not shown) is represented by a (three-phase) tie line to be illustrated later (Figure 6 [left]). In like manner, one may draw isothermal sections of Figure 3, such as Figure 5 taken at temperature *T'*. P_A and P_B are the vapour pressures of *A* and *B*, respectively, at *T'*. The similarity of Figures 4, left, and 5 is evident. Such qualitative mirror-image relationships between isotherms and isobars are common.

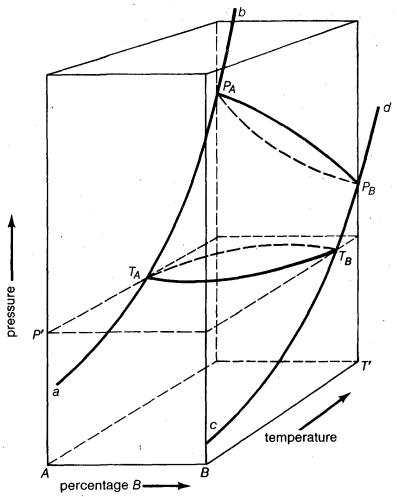


Figure 3: Portion of a typical solid model for a binary system.

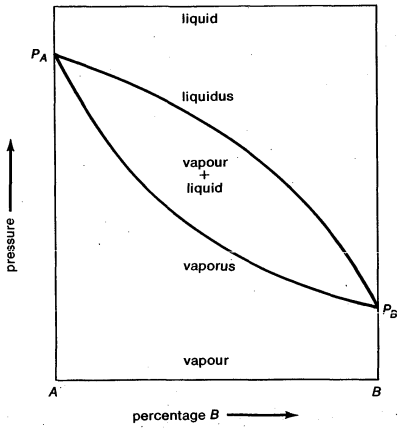


Figure 5: Binary liquid–vapour relations at constant temperature.

Simplest binary system. Figure 3 represents the simplest possible type of binary system and gives only the liquid–vapour relations. The labelling of the areas of the typical isobaric section, Figure 4, is self-explanatory. (T_A and T_B are boiling points of A and B.) In a closed system, an overall composition such as that of c, held at temperature T_X and pressure P', will yield a liquid of composition X_B and a vapour of composition Y_B. The ratio of the intercepts yc/cx on the tie line yx, gives the ratio of the quantity of liquid to the quantity of vapour phase. This principle is valid for all two-phase tie lines. If P' is the barometric pressure, the lower curve gives the boiling points of all solutions at that pressure. A solution of composition X_B boils at temperature T_X, and the vapour it gives has composition Y_B. Thus, the vapour is richer in A than the liquid. Continued removal of vapour from the boiling liquid causes the temperature and liquid and vapour compositions to change following the liquidus and vaporus lines, respectively, to the right. Such boiling-point diagrams are pertinent to fractional distillation of liquids. In general terms, most liquid solutions do not continue to boil at a fixed temperature but, because the vapour is richer in the more volatile components than the liquid, the temperature rises toward the boiling point of the least volatile component. Simultaneously both the remaining liquid and the vapour being formed become richer in the less volatile component. By proper arrangement of cooling apparatus, the vapours can be separated into samples or fractions such that each fraction has a different ratio of components. Usually the aim is to isolate one of the components, and, therefore, a specific fraction richest in that component will be distilled again, further reducing the unwanted components. This process of separating liquids by distilling them and trapping fractions for further distillation is fractional distillation.

Separation of liquids

Frequently, the solid model is such that its isobaric sections have the appearance of Figure 4, right (top and bottom), which show maxima or minima in the curves. At such unique points the two curves touch tangentially, and the liquid solutions distill unchanged in composition. This phenomenon is called azeotropy. Examples of liquid pairs belonging to this type, shown in Figure 4, left and right (top and bottom), are, respectively, methyl alcohol–water, acetone–chloroform, ethyl alcohol–water. Azeotropy prevents complete separation of a solution into its pure components by simple fractional distillation.

The comparative simplicity of Figures 3, 4, and 5 results from the assumed miscibility of liquids A and B and of their vapours. Similar diagrams result in considering equilibria between solid and liquid phases if solid A and solid B are miscible (continuous series of solid solutions) and if liquid A and liquid B are miscible. The previous discussion can be made pertinent by replacing the words liquid by solid, vapour by liquid, boiling by melting, and distillation by crystallization wherever they occur. Figure 4 then shows possible melting-point diagrams at constant pressure. The constant-pressure requirement is usually unimportant in melting-point diagrams.

More complex systems. A common solid–liquid relationship is one in which two components, A and B, have no miscibility as solids but are completely miscible as liquids. If the solids are not very volatile, an isobaric section at ambient pressure often has the appearance of Figure 6 (left), in which T_A and T_B are the respective melting points of A and B. The labelling of the various areas indicates, as before, what phases result when known mixtures are brought to equilibrium at a given temperature. For example, a mixture of composition X_B at temperature T_X will yield a liquid of composition l and solid B, lT_X being a tie line. The line T_BE is thus the solubility curve of solid B. A solution of composition l begins to freeze at temperature T_X. With further cooling, solid B continues to separate, and the liquid composition follows the line lE to E, whereupon the liquid solidifies without change in temperature to give an intimate mixture of solids A and B. Thus the line T_BE also shows the lowering of the freezing point of B by A. T_AE has an analogous meaning. The line cEd is a (three-phase) tie line where three two-phase areas come together and represents the coexistence of the three phases solid A, solid B, and liquid E. Because alpha equals two and beta equals three, and because pressure is constant, the system is isobarically invariant and therefore defines itself as long as all three phases are present. Hence the temperature is fixed at T_E and the liquid composition at E. Any attempt to dislodge the system from this state will be resisted if equilibrium is maintained. For example, if heat is added in an attempt to raise the temperature, solid A and solid B will melt and dissolve in each other to give more liquid of composition E, and this process, which is accompanied by the absorption of heat, will nullify the external heat input. Such a reversible phase reaction (which can be represented by the equation solid A + solid $B \rightleftharpoons$ liquid E) at an invariant point in a binary system, characterizes a system said to be at a (binary) eutectic point. (This statement applies even when solids A and B are solid solutions instead of pure substances.) The mixture of solids A and B in the proportions required by this equation is called the eutectic mixture, and it melts as though it consisted of one substance. Temperature T_E is the lowest temperature at which a liquid phase can form. Liquid, and solids A and B, can coexist at, and only at, temperature T_E. With vapour present, the system is at a quadruple point. The creation of a constant low-temperature environment by mixing the solids ice and salt is based on this eutectic phenomenon, as is also the use of low-melting alloys. Thus, the temperature of a mixture of solid potassium iodide and ice cannot rise above −23° C (−9° F), which is the T_E for this system, as long as both solids are present. The use of salt to melt ice on the streets, the purification of water by partial freezing, the use of sharpness of melting as a partial criterion of purity of a substance, the process of zone refining, and the determination of mixed melting points are all related to diagrams of the type represented by Figure 6 (left).

Ice–salt mixture

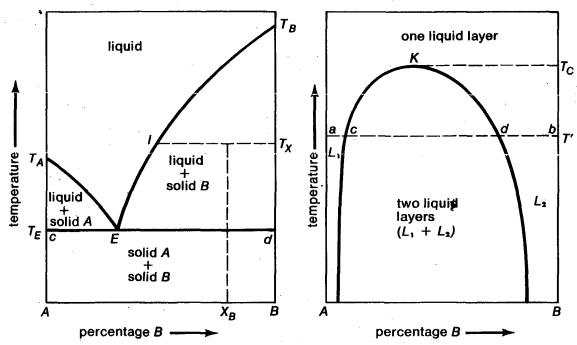


Figure 6: (Left) Binary system with eutectic and pure components as solid phases. (Right) System showing partial miscibility with upper critical solution point; L_1 = solution of B in A; L_2 = solution of A in B.

Between the extremes of complete miscibility and complete immiscibility lies partial miscibility, as exemplified

by the liquids ether and water mentioned earlier. Liquids *A* and *B* of Figure 6 (right) are in this category. Addition of *B* to *A* at temperature *T'* means following the line *acdb*. With small additions, only one layer is present (*a* to *c*); with larger additions, two layers (points *c* and *d*) result. Total compositions very rich in *B* again give only one layer (*d* to *b*). With rise in temperature, the miscibility gap (*cd*) often closes so that above T_C complete miscibility exists. Point *K* is an upper critical solution point. Less commonly, the gap closes with falling temperature, and in some examples it closes in both directions. Analogously, two solids may form solid solutions showing complete miscibility at one temperature and partial miscibility (discontinuous series of solid solutions) at another. The solids sodium chloride and potassium chloride, for example, are completely miscible above 500° C (900° F), yet only partially so below this temperature.

Complex variations. Many systems in their solid–liquid or liquid–vapour relations, combine the features just described. Thus a boiling-point diagram can involve partial miscibility of the liquid components with or without azeotropy. A melting-point diagram can combine partial miscibility of the liquids with immiscibility or with partial miscibility of the solids. Melting in systems with partial miscibility of the solid components may show eutectic behaviour, similar to that discussed above, or peritectic behaviour, outlined below. One or both of the components can be polymorphic. Solid, liquid, and vapour phases may appear in the same diagram. Supercritical phenomena may be encountered. Intercomponent compounds may form. Some of these are referred to below.

Components *A* and *B*, themselves frequently compounds, may form addition compounds (*C* in Figure 7).

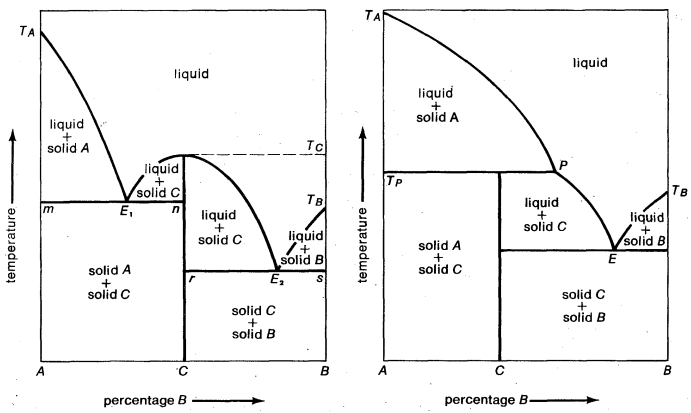


Figure 7: *Systems with intercomponent compounds* (*C*). (Left) Congruently melting compound. (Right) Incongruently melting compound.

In Figure 7 (left), which resembles two figures in juxtaposition—each similar to Figure 6 (left)—compound *C* melts congruently at T_C; *i.e.*, it melts to a liquid of the same composition. Lines mE_1n and rE_2s both represent eutectic points. In Figure 7 (right), *C* melts at T_P to a liquid, *P*, of a different composition and solid *A*; *i.e.*, *C* melts incongruently at T_P. The coexistence of solid *A*, solid *C*, and liquid *P* at temperature T_P, represented by the horizontal line at *P*, is an invariant system at constant pressure. Under these conditions, the system is said to be at a (binary) peritectic point (*cf.* eutectic point), characterized by the reversible phase reaction that can be written in equation form: solid $C \rightleftharpoons$ solid A + liquid *P*, which occurs on the addition or removal of heat and maintains the temperature at a constant value. The system consisting of copper and zinc, which includes ordinary brass, shows a series of intermetallic compounds and solid solutions; the system iron–carbon, which includes steel, shows the polymorphism of iron, compound formation, and solid solution.

Variables in ternary systems

Ternary and higher order systems. With a third component present, there are a maximum of four independent variables. It is thus necessary, for graphical representation, either to keep one variable constant or to ignore the effect of one of the variables (the one to which the system is least sensitive, commonly the pressure). This leaves the temperature and two independent concentration variables to be plotted as a solid model. In the absence of displacement reactions (in which parts of two components change places with each other in chemical reaction), the latter usually consists of a triangular prism as in Figure 8 (left),

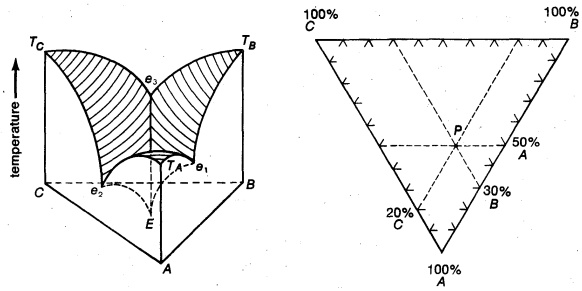


Figure 8: (Left) Freezing-point relations for a ternary system with pure components as solid phases. (Right) Method of plotting ternary composition with triangular coordinates.

the base of which is equilateral. Composition is plotted on triangular coordinates as illustrated in Figure 8 (right), in which point *P* represents the composition 50 percent *A*, 30 percent *B*, 20 percent *C* (either in weight or in mole percent). Temperature is measured vertically upward. Figure 8 (left) shows a typical solid–liquid relationship—that of three immiscible solids that form no compounds and that melt at T_A, T_B, T_C to give liquids that are completely miscible. Each of the three binary systems (*A–B*, *A–C*, *B–C*), which comprise the ternary, then shows the phase behaviour of Figure 6 (left) as seen on the vertical faces of Figure 8 (left). There are three curved surfaces sloping downward and inward, which intersect along the lines e_1E, e_2E, and e_3E, where e_1, e_2, and e_3 are the eutectic liquids of the respective binary systems. These surfaces give the (initial) freezing points of all solutions, the order in which the solids separate on cooling, and other details. Point *E*, a ternary eutectic liquid, gives the composition and temperature of the first liquid to form on heating, or last to remain on cooling. The cumulative effect on the lowering of the minimum temperature of melting through the addition of more components is evident. The curved surfaces may also be regarded as solubility surfaces. Figure 8 (left) is purposely simple, but other systems can be complicated by the presence of phenomena such as formation of compounds, partial or complete miscibility in the solid state, partial miscibility in the liquid state, ternary peritectic behaviour, presence of a vapour phase, etc. An extensively studied example is the complex system calcium oxide–aluminum oxide–silicon dioxide, $CaO–Al_2O_3–SiO_2$, because of its relation to the setting of portland cement.

When both pressure and temperature are held constant —*e.g.*, conveniently at one atmosphere and 25° C (77° F)—the phase relationships can be portrayed by a horizontal (isothermal) section of the triangular prism. The resulting plane triangular diagrams are widely used and illustrated in Figure 9. In this figure, for the system copper(II) chloride–potassium chloride–water; *i.e.*, $CuCl_2–KCl–H_2O$, at 75° C (167° F), the lines *ab*, *bc*, *cd*, and *de* are sections of curved surfaces of the prism (not

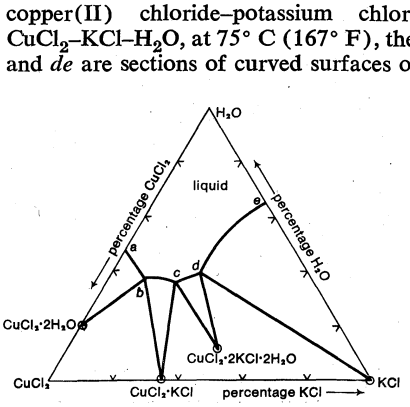


Figure 9: Phase diagram of $CuCl_2 \cdot KCl \cdot H_2O$ at 75° C.

shown) and give the compositions of liquids saturated with solid $CuCl_2 \cdot 2H_2O$, $CuCl_2 \cdot KCl$, $CuCl_2, \cdot 2KCl \cdot 2H_2O$ and KCl, respectively. Because tie lines join points representing coexisting phases, every point on the curved edge of each wedge-shaped region is joined by an imaginary tie line to the corresponding opposite corner of the wedge. Such diagrams are useful in determining, for example, the conditions under which these solids can be prepared at 75° C. Diagrams similar to Figure 8 (left) form the basis for purification by recrystallization where, for example, *C* is the substance to be purified, *B* is the impurity to be removed, and *A* is the solvent.

Figure 10 is for three nonreacting liquids, *A*, *B*, and *C*,

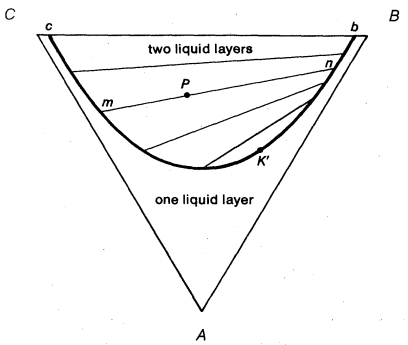


Figure 10: Typical ternary liquid system at constant temperature.

Three nonreacting liquids

in which the pairs *A–B* and *A–C* are miscible but the pair *B–C* is only partially so. Total compositions anywhere above the "binodal curve," *cK'b*, give two layers. For example, composition *P* gives layers of compositions *m* and *n*, respectively, *mn* being a tie line, several of which are shown. Below *cK'b*, only one phase is possible. As drawn, the slope of the tie lines indicates that there is more *A* in the *C*-rich layer than in the *B*-rich one. *K'*, where the tie lines have shrunk to a point, is called a plait point. Addition of sufficient *A* to the partially miscible *B–C* pair produces complete miscibility. Such systems are important in extraction processes.

Isothermal ternary diagrams often combine the phenomena typified by Figures 9 and 10. They may also involve the complications mentioned earlier. Systems with more than one binodal curve and systems with solid solutions have also been widely studied. When double-displacement reactions occur among the components a triangular diagram is not adequate.

Systems of four, five, and more components have been investigated but none completely. A historic study by a Dutch-born chemist, J.H. van't Hoff, of the solubility relations of the chlorides and sulfates of sodium, potassium, magnesium, and calcium in water was undertaken in connection with the utilization of the famous Stassfurt (now in East Germany) salt deposits and the determination of their geological history.

BIBLIOGRAPHY

Historical: J.W. GIBBS, "On the Equilibrium of Heterogeneous Substances," *Trans. Conn. Acad. Arts Sci.*, vol. 3 (1875–78), on the deduction of the phase rule.

College texts: W.F. SHEEHAN, *Physical Chemistry*, 2nd ed. (1970); G.W. CASTELLAN, *Physical Chemistry*, 2nd ed. (1971).

Technical: A. FINDLAY, A.N. CAMPBELL, and N.O. SMITH, *The Phase Rule and Its Applications*, 9th ed. (1951); R. HAASE and H. SCHONERT, "Solid-Liquid Equilibrium," in *The International Encyclopedia of Physical Chemistry and Chemical Physics*, topic 13, vol. 1 (1969); A.W. FRANCIS, *Liquid–Liquid Equilibriums* (1963); J.E. RICCI, *The Phase Rule and Heterogeneous Equilibrium* (1951); L. BREWER, "Viewpoints of Stability of Metallic Structures," in P.S. RUDMAN, J. STRINGER, and R.I. JAFFEE (eds.), *Phase Stability in Metals and Alloys* (1967).

Miscellaneous: Chapters on "Heterogeneous Equilibria and Phase Diagrams," in *A. Rev. Phys. Chem.*, vol. 1–7 (1950–56), and vol. 10 (1959); A.M. ALPER (ed.), *Phase Diagrams, Materials Science and Technology*, vol. 1, 2, 3 (1970); E.M. LEVIN, H.F. MCMURDIE, and F.P. HALL, *Phase Diagrams for Ceramists* (1956); D. EISENBERG and W. KAUZMANN, *The Structure and Properties of Water* (1969); F.P. BUNDY, "Direct Conversion of Graphite to Diamond in Static Pressure Apparatus," *Science*, 137:1057–1058 (1962).

(N.O.S.)

Phenomenology

In the 20th century, Phenomenology is mainly used as the name for a philosophical movement the primary objective of which is the direct investigation and description of phenomena as consciously experienced, without theories about their causal explanation and as free as possible from unexamined preconceptions and presuppositions. The word itself is much older, going back at least to the 18th century, when the Swiss-German mathematician and philosopher Johann Heinrich Lambert applied it to that part of his theory of knowledge that distinguishes truth from illusion and error. In the 19th century the word became associated chiefly with the *Phänomenologie des Geistes* (1807; Eng. trans., *Phenomenology of Mind*, 2nd ed., 1931), by G.W.F. Hegel, who traced the development of the human spirit from mere sense experience to "absolute knowledge." The so-called Phenomenological movement did not get under way, however, until early in the 20th century. But even his new Phenomenology includes so many varieties that a comprehensive characterization of the subject requires their consideration.

CHARACTERISTICS OF PHENOMENOLOGY

In view of the spectrum of Phenomenologies that have issued directly or indirectly from the original work of the Austrian-born German philosopher Edmund Husserl, it is not easy to find a common denominator for such a movement beyond its common source. But similar situations occur in other philosophical as well as non-philosophical movements.

Essential features and variations. Although, as seen from Husserl's last perspective, all departures from his own views could only appear as heresies, a more generous assessment will show that all those who consider themselves Phenomenologists subscribe, for instance, to his watchword, *Zu den Sachen selbst* ("To the things themselves"), by which they meant the taking of a fresh approach to concretely experienced phenomena, an approach, as free as possible from conceptual presuppositions, and the attempt to describe them as faithfully as possible. Moreover, most adherents to Phenomenology hold that it is possible to obtain insights into the essential structures and the essential relationships of these phenomena on the basis of a careful study of concrete examples supplied by experience or imagination and by a systematic variation of these examples in imagination. Some Phenomenologists also stress the need for studying the ways in which the phenomena appear in men's object-directed ("intentional") consciousness.

"To the things themselves"

Beyond this merely static aspect of appearance, some also want to investigate its genetic aspect, exploring, for instance, how the phenomenon intended—for example, a book—shapes ("constitutes") itself in the typical unfolding of experience. Husserl himself believed that such studies require a previous suspension of belief in the reality of these phenomena, whereas others consider it not indispensable but helpful. Finally, in existential Phenomenology, the meanings of certain phenomena (such as anxiety) are explored by a special interpretive ("hermeneutic") Phenomenology, the methodology of which needs further clarification.

Contrasts with related movements. It may also be helpful to bring out the distinctive essence of Phenomenology by confronting it with some of its philosophical neighbours. In contrast to Positivism and to traditional Empiricism, from which Husserl's teacher at Vienna, Franz Brentano, had started out and with which Phenomenology shares an unconditional respect for the positive data of experience ("We are the true positivists," Husserl claimed in his *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* [1913; "Ideas for a Pure Phenomenology and Phenomenological Philosophy"]), Phenomenology does not restrict these data to the range of sense experience but admits on equal

terms such non-sensory ("categorial") data as relations and values, as long as they present themselves intuitively. Consequently, Phenomenology does not reject universals; and, in addition to analytic a priori statements, whose predicates are logically contained in the subjects and the truth of which is independent of experience (*e.g.*, "All material bodies have extension"), and the synthetic a posteriori statements, whose subjects do not logically imply the predicate and the truth of which is dependent on experience (*e.g.*, "My shirt is red"), it recognizes knowledge of the synthetic a priori, a proposition whose subject does not logically imply the predicate but one in which the truth is independent of experience (*e.g.*, "Every colour is extended"), based on insight into essential relationships within the empirically given.

In contrast to phenomenalism, a position in the theory of knowledge (epistemology) with which it is often confused, Phenomenology—which is not primarily an epistemological theory—accepts neither the rigid division between appearance and reality nor the narrower view that phenomena are all that there is (sensations or permanent possibilities of sensations). These are questions on which Phenomenology as such keeps an open mind—pointing out, however, that phenomenalism overlooks the complexities of the intentional structure of men's consciousness of the phenomena.

In contrast to a Rationalism that stresses conceptual reasoning at the expense of experience, Phenomenology insists on the intuitive foundation and verification of concepts and especially of all a priori claims; in this sense it is a philosophy from "below," not from "above."

In contrast to an Analytic philosophy that substitutes simplified constructions for the immediately given in all of its complexity and applies "Ockham's razor," Phenomenology resists all transforming reinterpretations of the given, analyzing it for what it is in itself and on its own terms.

Phenomenology shares with Linguistic Analysis a respect for the distinctions between the phenomena reflected in the shades of meaning of ordinary language as a possible starting point for phenomenological analyses. Phenomenologists, however, do not think that the study of ordinary language is a sufficient basis for studying the phenomena, because ordinary language cannot and need not completely reveal the complexity of phenomena.

In contrast to an Existential philosophy that believes that human existence is unfit for phenomenological analysis and description, because it tries to objectify the unobjectifiable, Phenomenology holds that it can and must deal with these phenomena, however cautiously, as well as other intricate phenomena outside the human existence. (H.Sp.)

ORIGIN AND DEVELOPMENT OF HUSSERL'S PHENOMENOLOGY

Husserl's earlier development

Basic principles. Phenomenology was not founded; it grew. Its fountainhead was Edmund Husserl, who held professorships at Göttingen and Freiburg im Breisgau and who wrote *Die Idee der Phänomenologie* (Eng. trans., *The Idea of Phenomenology*, 1964) in 1906. Yet, even for Husserl, the conception of Phenomenology as a new method destined to supply a new foundation for both philosophy and science developed only gradually and kept changing to the very end of his career. Trained as a mathematician, Husserl was attracted to philosophy by Franz Brentano, whose descriptive psychology seemed to offer a solid basis for a scientific philosophy. The concept of intentionality, the directedness of the consciousness toward an object, which is a basic concept in Phenomenology, was already present in Brentano's *Psychologie vom empirischen Standpunkte* (1874): "And thus we can define psychic phenomena by saying that they are those phenomena which, precisely as intentional, contain an object in themselves." Brentano dissociated himself here from Sir William Hamilton, known for his philosophy of the "unconditioned," who had attributed the character of intentionality to the realms of thought and desire only, to the exclusion of that of feeling.

The point of departure of Husserl's investigation is to be found in the treatise *Der Begriff der Zahl* (1887; "The Concept of Number"), which was later expanded into *Philosophie der Arithmetik: Psychologische und logische Untersuchungen* (1891). Numbers are not found ready-made in nature but result from a mental achievement. Here Husserl was preoccupied with the question of how something like the constitution of numbers ever comes about. This treatise is important to Husserl's later development for two reasons: first, because it contains the first traces of the concepts "reflection," "constitution," "description," and the "founding constitution of meaning," concepts that later played a predominant role in Husserl's philosophy; and second, because it reflected two events—(1) a criticism of his book by Gottlob Frege, a seminal thinker in logic, who had charged him with confusing logical and psychological considerations, and (2) Husserl's discovery of the *Wissenschaftslehre* (1837; Eng. trans., *Logic and Scientific Methods*, 1971) by Bernard Bolzano, a Bohemian mathematician, theologian, and social moralist, and his view concerning "truths in themselves"—which led Husserl to an analysis and critical discussion of psychologism, the view that psychology could be used as a foundation for pure logic, which he clearly felt to be no longer possible.

In the first volume of *Logische Untersuchungen* (1900–01; Eng. trans., *Logical Investigations*, 1970), entitled *Prolegomena*, Husserl began with a criticism of psychologism. And yet he continued by conducting a careful investigation of the psychic acts in and through which logical structures are given; these investigations, too, could give the impression of being descriptive psychological investigations, though they were not conceived of in this way by the author. For the issue at stake was the discovery of the essential structure of these acts. Here Brentano's concept of intentionality received a richer and more refined signification. Husserl distinguished between perceptual and categorical intuition and stated that the latter's theme lies in logical relationships. The real concern of Phenomenology was clearly formulated for the first time in his *Logos* article, "Philosophie als strenge Wissenschaft" (1910–11; Eng. trans., *Philosophy as Rigorous Science*, 1965). In this work Husserl wrestled with two unacceptable views: naturalism and historicism.

Anti-naturalism

Naturalism attempts to apply the methods of the natural sciences to all other domains of knowledge, including the realm of consciousness. Reason becomes naturalized. Although an attempt is then made to find a foundation for the human sciences (*Geisteswissenschaften*) by means of experimental psychology, it proves to be impossible, because in so doing one is unable to grasp precisely what is at stake in knowledge as found in the natural sciences.

What a philosopher must examine is the relationship between consciousness and Being; and in doing so, he must realize that from the standpoint of epistemology, Being is accessible to him only as a correlate of conscious acts. He must thus pay careful attention to what occurs in these acts. This can be done only by a science that tries to understand the very essence of consciousness; and this is the task that Phenomenology has set for itself. Because clarification of the various types of objects must follow from the basic modes of consciousness, Husserl's thought remained close to psychology. In contradistinction to what is the case in psychology, however, in Phenomenology, consciousness is thematized in a very special and definite way, viz., just insofar as consciousness is the locus in which every manner of constituting and founding meaning must take place. In man's intuition, conscious occurrences must be given immediately in order to avoid introducing at the same time certain interpretations. The nature of such processes as perception, representation, imagination, judgment, and feeling must be grasped in immediate self-givenness. The call "To the things themselves" is not a demand for realism, because the things at stake are the acts of consciousness and the objective entities that get constituted in them: these things form the realm of what Husserl calls the phenomena.

Thus, the objects of Phenomenology are "absolute data grasped in pure, immanent intuition," and its goal is to discover the essential structures of the acts (*noesis*) and the objective entities that correspond to them (*noema*).

Anti-historicism

On the other hand, Phenomenology must also be distinguished from historicism, a philosophy that stresses the immersion of all thinkers within a particular historical setting. Husserl objected to historicism because it implies relativism. He gave credit to Wilhelm Dilthey, author of "Entwürfe zur Kritik der historischen Vernunft" ("Outlines for the Critique of Historical Reason," in *Gesammelte Schriften,* 12 vol. [1914–36], vol. 5, 6), for having developed a typification of world views, but he doubted and even rejected the skepticism that flows necessarily from the relativity of the various types. History is concerned with facts, whereas Phenomenology deals with the knowledge of essences. To Husserl, Dilthey's doctrine of world views was incapable of achieving the rigour required by genuine science. Contrary to all of the practical tendencies found in world views, Husserl demanded that philosophy be founded as a rigorous science. Its task implies that nothing be accepted as given beforehand but that the philosopher should try to find the way back to the real beginnings. This is tantamount to saying, however, that he must try to find the way to the foundations of meaning that are found in consciousness. Just as for Immanuel Kant the empirical has merely relative validity and never an absolute, or apodictic, validity, so for Husserl, too, what is to be searched for is a scientific knowledge of essences in contradistinction to a scientific knowledge of facts.

Basic method. The basic method of all Phenomenological investigation, as Husserl developed it himself—and on which he worked throughout his entire lifetime—is the "reduction": the existence of the world must be put between brackets, not because the philosopher should doubt it but merely because this existing world is not the very theme of Phenomenology; its theme is rather the manner in which knowledge of the world comes about. The first step of the reduction consists in the phenomenological reduction, through which all that is given is changed into a phenomenon in the sense of that which is known in and by consciousness; for this kind of knowing —which is to be taken in a very broad sense as including every mode of consciousness, such as intuition, recollection, imagination, and judgment—is here all-important. There are several reasons why Husserl gave a privileged position to intuition: among them is the fact that intuition is that act in which a person grasps something immediately in its bodily presence and also that it is a primordially given act upon which all of the rest is to be founded. Furthermore, Husserl's stress on intuition must be understood as a refutation of any merely speculative approach to philosophy.

This reduction reverses—"re-flects"—man's direction of sight from a straightforward orientation toward objects to an orientation toward consciousness.

Eidetic and transcendental reductions

The second step is to be found in the eidetic reduction. To get hold of consciousness is not sufficient; on the contrary, the various acts of consciousness must be made accessible in such a way that their essences—their universal and unchangeable structures—can be grasped. In the eidetic reduction one must forego everything that is factual and merely occurs in this way or that. A means of grasping the essence is the *Wesensschau,* the intuition of essences and essential structures. This is not a mysterious kind of intuition. Rather, one forms a multiplicity of variations of what is given, and while maintaining the multiplicity, one focusses attention on what remains unchanged in the multiplicity; *i.e.,* the essence is that identical something that continuously maintains itself during the process of variation. Husserl, therefore, called it the invariant.

To this point, the discussion of reduction has remained within the realm of psychology, albeit a new—namely, a phenomenological—psychology. The second step must now be completed by a third, the transcendental reduction. It consists in a reversion to the achievements of that consciousness that Husserl, following Kant, called transcendental consciousness, although he conceived of it in his own way. The most fundamental event occurring in this consciousness is the creation of time awareness through the acts of protention (future) and retention (past), which is something like a self-constitution. To do phenomenology was for Husserl tantamount to returning to the transcendental ego as the ground for the foundation and constitution (or making) of all meaning (German *Sinn*). Only when a person has reached this ground can he achieve the insight that makes his comportment transparent in its entirety and makes him understand how meaning comes about, how meaning is based upon meaning like strata in a process of sedimentation.

Husserl worked on the clarification of the transcendental reduction until the very end of his life. It was precisely the further development of the transcendental reduction that led to a division of the Phenomenological movement and to the formation of a school that refused to become involved in this kind of system of problems (see below *Phenomenology of essences*).

Basic concepts. In an effort to express what it is to which this method gives access, Husserl wrote:

Intentionality and constitution theory

> In all pure psychic experiences (in perceiving something, judging about something, willing something, enjoying something, hoping for something, etc.) there is found inherently a being-directed-toward. . . . Experiences are intentional. This being-directed-toward is not just joined to the experience by way of a mere addition, and occasionally as an accidental reaction, as if experiences could be what they are without the intentional relation. With the intentionality of the experiences there announces itself, rather, the essential structure of the purely psychical.

The phenomenological investigator must examine the different forms of intentionality in a reflective attitude because it is precisely in and through the corresponding intentionality that each domain of objects becomes accessible to him. Husserl took as his point of departure mathematical entities and later examined logical structures, in order finally to achieve the insight that each being must be grasped in its correlation to consciousness, because each datum becomes accessible to a person only insofar as it has meaning for him. From this position, regional ontologies, or realms of being, develop: for instance, those dealing with the region of "nature," the region of "the psychic," or the region of "the spirit." Moreover, Husserl distinguished formal ontologies—such as the region of the logical—from material ontologies.

In order to be able to investigate a regional ontology, it is first necessary to discover and examine the founding act by which realities in this realm are constituted. For Husserl, constitution does not mean the creation or fabrication of a thing or object by a subject; it means the founding constitution of its meaning. There is meaning only for consciousness. All founding constitution of meaning is made possible by transcendental consciousness. Speaking of this transcendental motif, Husserl wrote:

> It is the motif of questioning back to the last source of all achievements of knowledge, of reflection in which the knower reflects on himself and his knowing life, in which all the scientific constructs which have validity for him, occur teleologically, and as permanent acquisitions are kept and become freely available to him.

In the realm of such transcendental problems, it is necessary to examine how all of the categories in and through which one understands mundane beings or purely formal entities originate from specific modes of consciousness. In Husserl's view, the temporalization must be conceived as a kind of primordial constitution of transcendental consciousness itself.

Understood in this way, Phenomenology does not place itself outside the sciences but, rather, attempts to make understandable what takes place in the various sciences and, thus, to thematize the unquestioned presuppositions of the sciences.

Husserl's notion of the life-world

In his last publication, *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: Einee Einleitung in die phänomenologische Philosophie* (1936; Eng. trans., *The Crisis of European Sciences and Transcendental Phenomenology,* 1970), Husserl arrived at the life-world—the world as shaped within the immediate experience of each man—by questioning back to the foundations that the sciences presuppose. In *Die Krisis* he analyzed the grounds that had led to the European crisis of culture and philosophy, which found its

immediate expression in the contrast between the great successes of the sciences of nature and the failure of the sciences of man. In the modern era, scientific knowledge had become fragmented into an objectivistic–physicalist and a transcendental knowledge. Until recently this split could not be overcome. It led, rather, to the attempt to develop the sciences of man in accordance with the procedures used in the exact sciences of nature (naturalism) —an attempt doomed to failure. In opposition to this attempt, Husserl wished to show that in the new approach one must reflect on the activities of the scientists.

> As the immediately given world, this merely subjective world, was forgotten in the scientific thematization, the accomplishing subject, too, was forgotten and the scientist himself was not thematized.

Husserl demonstrated this point by using the example of Galileo and his mathematization of our world. The truth characteristic of the life-world is by no means an inferior form of truth when compared to the exact, scientific truth but is, rather, always a truth already presupposed in all scientific research. That is why Husserl claimed that an ontology of the life-world must be developed—*i.e.*, a systematic analysis of the constitutive achievements the result of which is the life-world, a life-world that, in turn, is the foundation of all scientific constitutions of meaning. The stimulating change that occurred here consists in the fact that truth is no longer measured after the criterion of an exact determination. For what is decisive is not the exactness but, rather, the part played by the founding act.

It is in this connection that, rather abruptly, historicity, too, became relevant for Husserl. He began to reflect upon the emergence of philosophy among the Greeks and on its significance as a new mode of scientific knowledge oriented toward infinity; and he interpreted the philosophy of René Descartes, often called the father of modern philosophy, as the point at which the split into the two research directions—physicalist objectivism and transcendental subjectivism—came about. Phenomenology must overcome this split, he held, and thus help mankind to live according to the demands of reason. In view of the fact that reason is the typical characteristic of man, mankind must find itself again through Phenomenology.

LATER DEVELOPMENTS OF PHENOMENOLOGY

Phenomenology of essences. A different type of Phenomenology, the Phenomenology of essences, developed from a tangential continuation of that of the *Logische Untersuchungen*. Its supporters were Husserl's students in Göttingen and a group of young philosophers in Munich, originally students of Theodor Lipps, a Munich psychologist and philosopher—students who had turned away from Lipp's psychologism and discovered powerful support in Husserl. The Phenomenological movement, which then began to take shape, found its most tangible expression in the publication of the *Jahrbuch für Philosophie und phänomenologische Forschung* (1913–30), a Phenomenological yearbook with Husserl as its main editor, the preface of which defined Phenomenology in terms of a return to intuition (*Anschauung*) and to the essential insights (*Wesenseinsichten*) derived from it as the ultimate foundation of all philosophy.

Fruits of the movement in broader application

The 11 volumes of the *Jahrbuch* contained, in addition to Husserl's own works, the most important fruits of the movement in its broader application. Of the co-editors, Alexander Pfänder contributed chiefly to the development of phenomenological psychology and pure logic but developed also the outlines of a complete Phenomenological philosophy. Moritz Geiger applied the new approach particularly to aesthetics and Adolf Reinach to the philosophy of law. The most original and dynamic of Husserl's early associates, however, was Max Scheler, who had joined the Munich group and who did his major Phenomenological work on problems of value and obligation. A Polish philosopher, Roman Ingarden, did major work in structural ontology and analyzed the structures of various works of art in its light; Hedwig Conrad-Martius, a cosmic Realist at the University of Munich, worked intensively in the ontology of nature; and others made comparable contributions in other fields of philosophy. None of these early Phenomenologists, however, followed Husserl's road to transcendental Idealism; and some tried to develop a Phenomenology along the lines of Realism.

Heidegger's hermeneutic Phenomenology. Martin Heidegger, one of Germany's foremost philosophers at the middle of the 20th century, was inspired to philosophy through Brentano's work *Von der mannigfachen Bedeutung des Seienden nach Aristoteles* (1862; "On the Multifarious Meaning of Being According to Aristotle"). While he was still studying theology, from 1910 to 1911, Heidegger encountered Husserl's *Logische Untersuchungen*. From then on he pursued the course of Phenomenology with the greatest interest, and from 1916 he belonged to the narrow circle of students and followers of the movement. The typical character of the Phenomenological intuition was at that time the focus of Husserl's seminar exercises. To be sure, there appeared very early a difference between Husserl and Heidegger. Discussing and absorbing the works of the important philosophers in the history of metaphysics was, for Heidegger, an indispensable task, whereas Husserl repeatedly stressed the significance of a radically new beginning and—with few exceptions (among them Descartes, Locke, Hume, and Kant)—wished to bracket the history of philosophy.

Heidegger's basic work, *Sein und Zeit* (1927; Eng. trans., *Being and Time,* 1962), which was dedicated to Husserl, strongly acknowledged that its author was indebted to Phenomenology. In it, Phenomenology was understood as a methodological concept—a concept that was conceived by Heidegger in an original way and resulted from his questioning back to the meanings of the Greek concepts of *phainomenon* and *logos*. *Phainomenon* is "that which shows itself from itself," but together with the concept of *logos*, it means "to let that which shows itself be seen from itself in the very way in which it shows itself from itself." This conception of Phenomenology, which relied more on Aristotle than on Husserl, constituted a change that was later to lead to an estrangement between Husserl and Heidegger. For in *Sein und Zeit* there is no longer a phenomenological reduction, a transcendental ego, or an intuition of essences in Husserl's sense. Heidegger's new beginning was, at the same time, a resumption of the basic question of philosophy: that concerning the meaning (*Sinn*) of Being. His manner of questioning can be defined as hermeneutical in that it proceeds from the interpretation of man's situation. What he thematized is, thus, the explanation of what is already understood.

Heidegger's return to ontology

At the heart of *Sein und Zeit* lies Heidegger's analysis of the one (the man) who asks the question—who is capable of asking the question—concerning Being, who precisely through this capability occupies a privileged position in regard to all other beings, viz., that of *Dasein* (literally, "being there"). By conceiving of *Dasein* as being-in-the-world, Heidegger made the ancient problem concerning the relationship between subject and object superfluous. The basic structures of *Dasein* are primordial moodness (*Befindlichkeit*), understanding (*Verstehen*), and logos (*Rede*). These structures are, in turn, founded in the temporalization of *Dasein*, from which future, having-been (past), and present originate. The two basic possibilities of man's existing (from the Latin *ex* and *sistere*, "standing out from") are those in which *Dasein* either comes to its self (called authenticity) or loses itself (called inauthenticity); *Dasein* is inauthentic, for example, when it lets the possibilities of the choice for its own "ek-sisting" be given to it by others instead of deciding for itself. Heidegger's concept of care (*Sorge, cura*) has nothing to do with distress (*Bekümmernis*) but includes the unity of the articulated moments of man's being-in-the-world.

The hermeneutic character of Heidegger's thought manifested itself also in his interpretation of poetry, in which he discovered a congenial spirit in Friedrich Hölderlin, one of Germany's greater poets, of whose poetry he inaugurated a completely new interpretation; but it manifested itself equally well in his interpretation of metaphysics,

which Heidegger tried to envision as an occurrence determined by the forgottenness of Being, an occurrence in the centre of which man finds himself and of which the clearest manifestation is to be found in "technicity," the attempt of modern man to dominate the earth by controlling beings that are considered as objects.

The concept of transcendental consciousness, which was central for Husserl, is not found in Heidegger—which clearly shows how Heidegger, in *Sein und Zeit*, had already dissociated himself from Husserl's Phenomenology.

Other developments. Eugen Fink, for several years Husserl's collaborator, whose essay "Die phänomenologische Philosophie Edmund Husserls in der gegenwärtigen Kritik" (1933) led to a radicalization of Husserl's philosophical, transcendental Idealism, later turned in another direction, one that approached Heidegger's position and divorced itself at the same time from that of Husserl.

Ludwig Landgrebe, who was Husserl's personal assistant for many years, published in 1938 *Erfahrung und Urteil* ("Experience and Judgment"), the first of Husserl's posthumous works devoted to the genealogy of logic. Among German-language scholars, Landgrebe remained closest to Husserl's original views and has developed them consistently in several works.

DISSEMINATION OF PHENOMENOLOGY

Phenomenology in various countries. Following upon the work of Husserl, Phenomenology spread into a worldwide movement.

In France. One of the first French authors to become familiar with Husserl's philosophy was Emmanuel Lévinas, a pluralistic Personalist, who combined ideas from Husserl and Heidegger in a very personal way. Similarly, Jean-Paul Sartre, the leading Existentialist of France, took his point of departure from the philosophies of Husserl and Heidegger. His first works, *L'Imagination* (1936; Eng. trans., *Imagination: A Psychological Critique*, 1962) and *L'Imaginaire: Psychologie phénoménologique de l'imagination* (1940; Eng. trans., *The Psychology of Imagination*, 1950), remain completely within the context of Husserl's analyses of consciousness. Sartre explains the distinction between perceptual and imaginative consciousness with the help of Husserl's concept of intentionality, and he frequently employs the method of ideation (*Wesensschau*).

Jean-Paul Sartre's use of the method

In *L'Etre et le néant* (1943; Eng. trans., *Being and Nothingness*, 1956), an essay on Phenomenological ontology, it is obvious that Sartre borrowed from Heidegger. Some passages from Heidegger's *Was ist Metaphysik?* (1929; Eng. trans., *What is Metaphysics?*, 1949), in fact, are copied literally. The meaning of nothingness, which Heidegger in this lecture made the theme of his investigations, became for Sartre the guiding question. Sartre departs from Heidegger's analytic of *Dasein* and introduces the position of consciousness (which Heidegger had overcome).

The distinction between being-in-itself (*en-soi*) and being-for-itself (*pour-soi*) pervades the entire investigation. The in-itself is the opaque, matter-like substance that remains the same, whereas the for-itself is consciousness permeated by nothingness. The influence of the Idealist G.W.F. Hegel becomes apparent when the author tries to interpret everything in a dialectical way; *i.e.*, through a tension of opposites. The dialectic of men's being-with-one-another is central: thus, seeing and being-seen correspond to dominating and being-dominated. The basic characteristic of being-for-itself is bad faith (*mauvaise foi*), which cannot be overcome because facticity (being-already) and transcendence (being-able-to-be) cannot be combined.

The Phenomenological character of Sartre's analyses of consciousness consists in the way in which he elucidates certain modes of behaviour: love, hatred, sadism, masochism, and indifference. Although Sartre sees and describes these forms of behaviour strikingly and precisely, he limits himself to those modes that fit his philosophical interpretation. The significance of psychology, recognized by Husserl, emerges again in Sartre and leads to a demand for an Existential psychoanalysis.

Sartre's definition of man as a being of possibilities that finds or loses itself in the choice that it makes in regard to itself refers to Heidegger's definition of *Dasein* as a being that has to materialize itself. For Sartre, freedom is the basic characteristic of man; thus Sartre belongs to the tradition of the great French moralist philosophers.

In his later works, as in his *Critique de la raison dialectique* (1960; Eng. trans., *Search for a Method*, 1963), Sartre turned to Marxism, though he developed a method of understanding that was influenced by hermeneutics. Here the choice made by the individual is limited by social and psychological conditions. Sartre's outstanding two-volume interpretation of Flaubert, *L'Idiot de la famille; Gustave Flaubert de 1821–1857* (1971), is an example of this new method of understanding and interpretation, which combines Marxist elements with interpretations of a highly personal nature taken from depth psychology.

The Phenomenology of Maurice Merleau-Ponty

Maurice Merleau-Ponty (died 1961), who, together with Sartre and his associate Simone de Beauvoir, a writer and novelist, was an important representative of French Existentialism, was at the same time the most important French Phenomenologist. His works, *La Structure du comportement* (1942; Eng. trans., *Structure of Behaviour*, 1963) and *Phénoménologie de la perception* (1945; Eng. trans., *Phenomenology of Perception*, 1962), were the most original further developments and applications of Phenomenology to come from France. Merleau-Ponty gave a new interpretation of the meaning of the human body from the viewpoint of Phenomenology and, connected with this, of man's perception of space, the natural world, temporality, and freedom.

Starting from Husserl's later phenomenology of the life-world, Merleau-Ponty anchored the phenomena of perception in the phenomenology of the lived body (the body as it is experienced and experiences), in which the perceiving subject is incarnated as the mediating link to the phenomenal world. Such a phenomenology of human "presence" in the world was also to offer an alternative to the rigid dichotomy between Idealism and Realism, in which consciousness and world could be reciprocally related. Phenomenology thus became a way of showing the essential involvement of human existence in the world, starting with everyday perception.

Although it is true that Merleau-Ponty was originally close to Husserl in his thought, he later developed noticeably in the direction of Heidegger, a change that became particularly manifest in *L'Oeil et l'esprit* (1964; "The Eye and the Spirit").

Paul Ricoeur, a student of the volitional experience, whose translation of Husserl's *Ideen zu einer reinen Phänomenologie* brought Husserl closer to the younger French generation, writes in a Phenomenological vein but with the intention of further developing Husserl's conception of Phenomenology. Ricoeur's two-volume *Philosophie de la volonté* (1950–60; "Philosophy of the Will") also deals with the problems involved in the theological concept of guilt.

Suzanne Bachelard, who in 1957 translated Husserl's *Formale und transzendentale Logik: Versuch einer Kritik der logischen Vernunft*, has pointed to the significance of Husserl for modern logic; and Jacques Derrida, an original French thinker on the limits of thought and language, has combined Phenomenology and Structuralism in his interpretation of literature.

Husserl scholarship at Cologne and Louvain

In Germany. After World War II, interest in Phenomenology sprang up again in its own homeland. The influence of Ludwig Landgrebe in Cologne has been particularly felt, as have the activities of the Husserl Archives in Cologne, with editions by Walter Biemel, who also published *Philosophische Analysen zur Kunst der Gegenwart* (1968; "Philosophical Analyses of Contemporary Art") and essays on the relationships between Husserl and Heidegger. The circle around Gerhard Funke in Mainz, author of *Phänomenologie—Metaphysik oder Methode?* (1966), has also had a positive influence.

In other European countries. In Belgium, at the Cath-

olic University of Louvain, are located the entire posthumous works of Husserl, as well as his personal library. Thanks to the initiative of H.L. Van Breda, founder of the Husserl Archives, several scholars worked intensively on the manuscripts for several decades. By 1972, 12 volumes of collected works had been published. Van Breda was also the director of the *Phaenomenologica* series —totalling 42 volumes by 1972—in which the most important publications in the field of Phenomenology (taken in a very broad sense) have been published. Thus, mainly through Van Breda's efforts, Louvain has become the most important centre for Phenomenology. Van Breda also organized international colloquia on Phenomenology. The influence of Alphonse de Waelhens, a Belgian philosopher of fresh experience and author of *Phénoménologie et vérité* (1953) and *Existence et signification* (1958), also bears mentioning.

In The Netherlands, Stephan Strasser, who is oriented particularly toward phenomenological psychology, is especially influential. And in Italy, the Phenomenology circle centres around Enzo Paci. The Husserl scholar Jan Patocka, a prominent expert in Phenomenology as well as in the metaphysical tradition, is influential in Czechoslovakia; in Poland, Roman Ingarden has represented the cause of Phenomenology; and there are also important representatives in Yugoslavia, Portugal, England, South America, Japan, and India.

In the United States. In the United States, Phenomenology has lived a rather marginal existence for quite some time, notwithstanding the meritorious journal of *Philosophy and Phenomenological Research* founded by Husserl's student Marvin Farber, who is also the author of *The Foundation of Phenomenology* (1943). In recent years, however, a noticeable change has taken place, chiefly because of the work of two scholars at the New School for Social Research in New York, Alfred Schütz, an Austrian-born sociologist and student of human cognition (died 1959), and Aron Gurwitsch, a Lithuanian-born philosopher. Schütz came early to Phenomenology, developing a social science on phenomenological basis. Gurwitsch, author of *Théorie du champ de la conscience* (1957; Eng. trans., *The Field of Consciousness*, 1964), came to Phenomenology through his study of the Gestalt psychologists Adhemar Gelb and Kurt Goldstein. While in Paris, Gurwitsch influenced Merleau-Ponty. The essays on Phenomenology published by Gurwitsch in the United States are among the best. His comprehensive knowledge ranges from mathematics, via the natural sciences, to psychology and metaphysics. The work *The Phenomenological Movement* (2nd ed. 1965), by Herbert Spiegelberg, an Alsatian–American Phenomenologist, is the movement's first encompassing historical presentation.

The work of Farber, Schütz, and Gurwitsch

Phenomenology in other disciplines. Of greater significance is the role of Phenomenology outside of philosophy proper in stimulating or reinforcing phenomenological tendencies in such fields as mathematics and the biological sciences. Much stronger was its impact on psychology, in which Franz Brentano and the German Carl Stumpf had prepared the ground and in which the U.S. psychologist William James, the Würzburg school, and the Gestalt psychologists had worked along parallel lines. But Phenomenology has probably made its strongest contribution in the field of psychopathology, in which the German Karl Jaspers, a foremost contemporary Existentialist, stressed the importance of phenomenological exploration of a patient's subjective experience. Jaspers was followed by the Swiss Ludwig Binswanger and several others. The Phenomenological strand is also very pronounced in American Existential psychiatry, and has affected sociology, history, and the study of religion.

PROGNOSTICATION

At the turn of the fourth quarter of the 20th century, it remained to be seen whether Phenomenology could make solid contributions to philosophical knowledge. To this end, it needed to develop rigorous standards, which had not always been observed by some of its most brilliant practitioners, such as Max Scheler, and which were likely to be violated in a philosophy the ultimate appeal of which had to be made to intuitive verification. With this proviso, Phenomenology may well be qualified not only to become a bridge for better international communication in philosophy but also to shed new light on philosophical problems old and new, to reclaim for philosophy parts of man's everyday life-world that have been abandoned by science as too private and too subjective, and, finally, to give access to layers of man's experience unprobed in everyday living, thus providing deeper foundations for both science and life.

BIBLIOGRAPHY. Most of the classical works on Phenomenology were written by Husserl himself (see the bibliography of the HUSSERL article). In EDMUND HUSSERL, *Logische Untersuchungen*, 2nd ed., 2 vol. (1913–21; Eng. trans., *Logical Investigations*, 2 vol., 1970), one of the fundamental texts on Phenomenology, the phenomenological method is applied in the area of logic. The following works appeared in the journal *Jahrbuch für Philosophie und phänomenologische Forschung* (1913–30); Husserl's *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* (1913; Eng. trans., *Ideas: General Introduction to Pure Phenomenology*, 1931, reprinted 1969), through which Phenomenology established itself as transcendental philosophy and received worldwide reaction; *Vorlesungen zur Phänomenologie des inneren Zeitbewusstseins* (1928; Eng. trans., *The Phenomenology of Internal Time-Consciousness*, 1964); and *Formale und transzendentale Logik* (1929; Eng. trans., *Formal and Transcendental Logic*, 1969); MAX SCHELER, *Der Formalismus in der Ethik* (1916); and MARTIN HEIDEGGER, *Sein und Zeit* (1927; Eng. trans., *Being and Time*, 1962). The following works appeared in the series *Husserliana* (Husserl's collected works): *Cartesianische Meditationen und Pariser Vorträge* (1950; Eng. trans., *The Paris Lectures*, 1964), which contains the text of the Paris lectures of 1929 and the subsequent elaboration; *Die Idee der Phänomenologie* (1950; Eng. trans., *The Idea of Phenomenology*, 1964), an introduction to Phenomenology in five lectures from 1907; *Ideen*, 3 vol. (1950–52); *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie* (1954; Eng. trans., *The Crisis of European Sciences and Transcendental Phenomenology*, 1970), Husserl's later work (1934–37), significant for the problems regarding the life-world; *Erste Philosophie*, 2 vol. (1956–59), a critical history of ideas and a theory of reduction presented in a series of lectures, 1923–24; *Phänomenologische Psychologie* (1962), lectures from 1925, a different wording of the *Encyclopædia Britannica* article of 1927, with remarks by Heidegger and the Amsterdam addresses of 1925; *Zur Phänomenologie des inneren Zeitbewusstseins* (1966), text on the problem of time from 1893–1917 with lectures (in the middle) from 1905 on the inner time-consciousness (first edited by Heidegger); *Analysen zur passiven Synthesis* (1966), a phenomenological analysis having sensation as its subject matter; and *Philosophie der Arithmetik* (1970), early manuscripts from 1890 to 1901. See also JOSEPH J. KOCKELMANS, *Edmund Husserl's Phenomenological Psychology* (1967); and RUDOLF BOEHM, *Vom Gesichtspunkt der Phänomenologie* (1968).

In the series "Phaenomenologica" works are published that are written from a Phenomenological perspective. HERBERT SPIEGELBERG, *The Phenomenological Movement*, 2nd ed., 2 vol. (1965), worthy of particular mention, also appeared in this series. Since 1904 MARVIN FARBER has edited the journal *Philosophy and Phenomenological Research* (by no means, however, exclusively dedicated to Phenomenology). In England *The Journal of the British Society for Phenomenology* first appeared in 1970. In the United States the following journals have appeared: *Journal of Phenomenological Psychology* (semi-annual); *Research in Phenomenology* (1971); and *Analecta Husserliana: The Yearbook of Phenomenological Research.*

(W.Bi.)

Philadelphia

Philadelphia has been described either as the elegant but rather jaded great lady or as the overage and sickly spinster of American cities. A more realistic portrait of contemporary Philadelphia, however, must show it as a very new and vigorous city, arising in gracious counterpoint to the deep serenity of an older city that, for nearly three centuries, has provided gentle but firm intellectual, economic, and humanitarian direction to the nation at whose birth it played midwife.

An overview of the city

Philadelphia, the fourth-largest city in the United States and largest in Pennsylvania, displays many characteris-

tics of a small town. Its many trees, parks, and other open spaces and its quiet pace of life reflect in many ways the genteel Quaker heritage bestowed on the city by its founder, William Penn. Nearly everywhere are dignified reminders of the colonial and Revolutionary city and of Benjamin Franklin, a Philadelphian by adoption, who left his imprint on innumerable ongoing institutions, both cultural and commercial, in the city.

Beneath this facade, however, Philadelphia represents a physical presence of national and international stature. Though the population of the city proper slid to below 2,000,000 during the 1960s, it is the focus of more than 2,750,000 other persons in five counties in eastern Pennsylvania and three in neighbouring New Jersey. Its harbour, the largest freshwater port in the world, is the major element in the official agglomeration of Delaware River ports known as Ameriport, collectively the third-busiest shipping centre in the world. The enormous industrial production of the city and the surrounding metropolitan area, at least half of it representing products unknown as recently as 1945, represents a continuation of Philadelphia's early leadership in the Industrial Revolution and in American commerce and finance generally. Lying in the midst of the vast urban community stretching down the Atlantic Seaboard, the Philadelphia of the 1970s is an integral part of the turbulent fabric of contemporary social and economic life as well as a tranquil oasis joining together the spirit of America, past and present. (For information on related topics, see the articles PENNSYLVANIA; UNITED STATES, HISTORY OF THE.)

THE GROWTH OF THE METROPOLIS

Penn's plan

The founding ideals. William Penn acquired the province of Pennsylvania in 1681 from King Charles II of England as a place where his fellow Quakers could enjoy freedom of worship and a chance to govern themselves and develop their own way of life. The King made the grant, signing it on March 4, 1681, and proclaiming it the following April 2, partly to settle a debt owed to Penn's father, Admiral Sir William Penn, upon his death and also to complete the settlement of this Middle Atlantic region with Englishmen. Penn sent his cousin William Markham to take charge of affairs of government and also to lay out the city Penn named Philadelphia, city of "brotherly love," the name betokening his idealistic concepts. From England, Penn wrote in 1681 asking that "the Rivers and Creeks be sounded on my side of the Delaware River . . . in order to settle a great Towne, and be sure to make your choice where it is most navigable, high, dry, and healthy." He wanted every house to be placed in the middle of its own plot to provide ground about it "that it may be a greene Country Towne, which will never be burnt, and always be wholesome."

Penn had little chance to enjoy his city. He was forced to return to England in 1684 and was unable to return until 1699. By then, Philadelphia was a flourishing town with many shops and trading houses, as well as several hundred dwellings and some 10,000 people clustered close to the riverfront. Penn's governor declared the city already was the equal of New York "in trade and riches." Penn's policies throughout the colony of religious toleration and the right of the people to take part in the government, in addition to growing prosperity, soon began to attract thousands of English, Germans, and Scots-Irish settlers, and they all came by way of Philadelphia. Most of this immigration moved into the interior and, by 1770, even west of the Allegheny Mountains. This created new sources of farm, forest, and mineral wealth that Philadelphia could tap to add to its prosperity.

Economic growth. Philadelphia by the 1770s had grown to at least 30,000 persons in the central city, and it was the third most important business centre in the British Empire, overshadowed only by Liverpool and London. This position was due in large measure to the city's site at the confluence of the Delaware and Schuylkill rivers, which provided the city access to inland farm regions and to the coal and ore resources that supplied the early iron industry.

A visitor in 1756 wrote that "Everybody in Philadelphia deals more or less in trade," a tribute not merely to Philadelphia's location but also to the shrewd business talents of the Quaker merchants. A large and profitable system of triangular trade involved foodstuffs and wood products, such as lumber and barrel staves, that went from Philadelphia to the West Indies and there were exchanged for sugar, rum, and other West Indian products. These were carried to English ports, where they in turn were exchanged for English manufactures to be brought back to Philadelphia. The prosperous farm country of interior Pennsylvania supplied the Philadelphia merchant with his goods for the West Indies, and a profitable coastal trade existed with other colonies and directly with England. By the 1750s, Philadelphians had invested heavily in the flourishing charcoal–iron industry. Anthracite coal became a great new mineral resource of Pennsylvania 100 years later, and Philadelphia capital played a leading part in this industry as well as in the mining of bituminous coal farther to the west. Philadelphia continued its leadership in foreign commerce until about 1810, when New York, with an even more advantageous location, took over this position. It did not surrender its position as financial capital of the nation, however, until the 1880s.

Philadelphia's leadership in commerce and culture

Shipyards had flourished along the Delaware since colonial days. Most of what came to the city was raw material for manufacture, and Philadelphia became a major centre of the early Industrial Revolution in the United States. In 1785 Oliver Evans invented the first grist mill operated entirely by mechanical power. The city was a pioneer in textile manufacturing and took the raw iron from inland furnaces and made it into tools and implements, such as saws, huge iron castings for cotton-mill machinery, and the first American-built steam locomotives. By 1860 the value of Philadelphia's manufactures ran into several hundred million dollars, about 30 percent of the national total. Textiles, ships, iron products, leather, refined sugar, and boots and shoes were leaders, giving important aid to the Union in the Civil War.

Cultural dominion. Prosperity was translated into personal and community wealth, and it was natural that Philadelphia should assume early leadership in the arts, in science, and in culture. The young Benjamin Franklin came from Boston to become a great American leader in scientific and intellectual affairs. Philadelphia had the nation's first free library, its first hospital, and its first learned society, the American Philosophical Society—all founded by Franklin. Along with Franklin, there were men such as Benjamin Rush, the great physician, and David Rittenhouse, a paper manufacturer and early Philadelphia aristocrat, and many others. The city excelled in printing and publishing: by 1776 there were 23 printers and newspapers with circulations of from 500 to 3,000 copies. Fine private and public buildings were erected. One of them was Andrew Hamilton's Independence Hall, originally—and still—better known by Philadelphians as "the statehouse." Led by Benjamin West, Charles Willson Peale, and Gilbert Stuart, oil painting flourished in the colonial and Federal periods.

This leadership in commerce, thought, and culture, as well as its central location, brought to Philadelphia the delegates who formed the First Continental Congress in 1774 and a year later the Second Continental Congress, which proclaimed the Declaration of Independence and governed throughout the Revolution. The Constitutional Convention met in Philadelphia in 1787 and framed the federal Constitution, and the city served as the capital of the United States from 1790 to 1800. The nation celebrated its centennial in Philadelphia in 1876.

Political evolution. In the decades just before and following the Civil War, Pennsylvania fell into the hands of Philadelphia-based Republican political machines that, like their counterparts (usually Democratic) in other cities, were becoming increasingly sophisticated in methods of manipulating the political processes, especially through the newer immigrant groups, and profiting from the economic life of the state and city. They were instrumental in electing governors and U.S. senators, and, in the Depression years of the 1930s, the hard-hit city re-

Political machines and civic malaise

fused aid from the Democratic administration in Washington, D.C. Powerful bankers and industrial and business leaders, living for the most part outside the city, favoured this form of government because it kept taxes low, imposed little or no regulation on business, and maintained an aura of social calm through benign neglect or quiet but forceful repression.

About 1900 Philadelphia had been described as "corrupt but content," a status quo that Philadelphians were indeed content with until 1939, when a group known as the Young Turks and influenced by the nationwide New Deal of the Democratic Party began to agitate for charter reform and a city planning commission. Women and blacks were brought into the city's political life for the first time. In spite of continued machine domination of the city, the group began to realize its goals—with a "Better Philadelphia Exhibition" in 1947, a coalition of top business and financial leaders the following year, and a new charter in 1951. The election in 1952 of Joseph S. Clark as mayor signalled the moment for a new Philadelphia to leap off the drawing boards.

THE CONTEMPORARY CITY

Present-day Philadelphia spreads over about 128½ square miles (333 square kilometres) from the Delaware on the east across Schuylkill and beyond. The coterminous boundaries of Philadelphia city and county remain essentially as defined by the Consolidation Act of 1854.

The city and its people. *Physical layout.* The gridiron of streets in central Philadelphia, the first U.S. city to be so arranged, follows the original plans of William Penn. Midway between the two rivers, Penn Square, occupied by City Hall, is the centre of the plan. Four shady, fountained squares—Logan, Franklin, Washington, and Rittenhouse—are placed in each quadrant. Westward from Penn Square along John F. Kennedy Boulevard is the new Penn Center, a prime symbol of rejuvenated Philadelphia since the 1950s. The multilevel complex comprises high-rise offices and hotels, with interior courts and malls and underground walkways lined with shops and restaurants.

Benjamin Franklin Parkway provides a splendid vista as it cuts diagonally northwestward from Penn Square through the grid, encircling Logan Square and leading into Fairmount Park. The nation's largest landscaped park within city limits and the centre of the Centennial Exposition of 1876, Fairmount is one of the most frequent foregrounds for photographs of Philadelphia's skyline, adding to the city's reputation for shaded, sculpted elegance. Once a section of outlying estates of the wealthy, the park contains many fine old mansions maintained by the city as museums. Through it winds the narrow valley of Wissahickon Creek, whose rugged beauty has inspired generations of poets and painters.

Historic preservations and restorations

The oldest sections of Philadelphia—Southwark, Society Hill, and the Independence Hall area—lie to the east, along and inland from the Delaware. Southwark is the oldest, having been settled by Swedes in 1643. Those of its ancient and dilapidated houses that have escaped bulldozing for riverfront expressways resemble the edifices of Society Hill before its restoration in the 1950s. This latter area contains some of the city's finest old houses, taverns, and churches, though some high-rise apartment buildings, here as elsewhere, strike a dissonant note in the hearts of many Philadelphians.

Independence National Park, established in 1951, contains not only the hall itself with its famous Liberty Bell but also many other buildings used by the Revolutionary and early federal governments. Nearby, Elfreth's Alley, dating from 1690, contains 33 houses that make up the oldest continuously inhabited street in America. Included in the complex are Carpenters' Hall, site of the meeting of the First Continental Congress, and Philosophical Hall, home of the American Philosophical Society.

Throughout these areas and elsewhere, domestic architecture is characterized primarily by two- and three-story red-brick structures fronting directly on the sidewalks and containing impressive examples of colonial design both outside and inside. The Philadelphia Historical Commission has certified about 75 restored buildings predating 1800. Restoration of the old but habitable has been more characteristic than wholesale demolition and rebuilding throughout most of the city. Even in its downtown section, Philadelphia continues to be a city of shops rather than of huge merchandising outlets. It has nothing approaching New York's Fifth Avenue as a street of large stores. Chestnut Street is an attractive street of shops but has only two department stores. Pine and other streets are noted for such specialties as antique shops. Major features of the skyline are the massive contours of the city's many banking and financial institutions. Overall, downtown Philadelphia has about 350 office buildings.

The long stretches of Philadelphia north, south, and west of the intersection of Broad and Chestnut streets, a "city hub" of sorts just below Penn Square, contain numerous distinctive sections, often identified for generations with various ethnic groups that have filled the city during its long history. Among the more interesting is the Germantown section of North Philadelphia, settled in Penn's time by Germans and the home in the 18th century of wealthy Philadelphians fleeing the periodic yellow-fever epidemics of the riverside city. Much of North Philadelphia comprises a massive black and Puerto Rican ghetto. South Philadelphia contains notable Italian and other sections settled by European immigrants mainly in the 19th century.

The Philadelphians. Though Philadelphia has had most of the characteristics of an ethnic and racial pot nearly from its start, it lacks the steaming hurly-burly visible everywhere in its behemoth neighbour, New York, some 90 miles northeast. The Philadelphians by and large are not a street people, and their orientation has been so traditionally toward the home that the city became known as a "city of homes." Philadelphians are also great joiners, giving the city more social and other clubs than any other in America. Many of these are based in ethnic neighbourhoods, others are city wide, and still others serve the affluent "Main Liners" who reside in the plush suburbs that grew up in past centuries along the main line of the Pennsylvania (now Penn Central) Railroad.

Quaker tradition and ethnic mix

Much of this tradition dates from Philadelphia's early years, from the beliefs and attitudes not only of the Quakers but also of such German pietist sects as the Mennonites, who stressed personal religious experience rather than institutionalized formulations. A basic sobriety underlying these tenets led to many stringent laws that remain in both city and state. On the other hand, these groups strove for tolerance in all matters. Pennsylvania was among the few colonies admitting Catholics and Jews, and the Quakers long were leaders in seeking justice and the alleviation of inhumanity in racial and other human relations. Philadelphia still has innumerable small, endowed charities of Quaker origin to provide the poor with coal or food—as well as one of the oldest and strongest municipal commissions on human relations.

In ethnic mix, these early Protestant sects were challenged increasingly during the 1800s by Catholics of Irish, German, Polish, Slavic, and Italian origin. In 1840 the municipal tranquillity was jarred by anti-Catholic, nativist riots. By the early 1970s more than one-third of the Philadelphians and well over one-half the public-school population was black. The suburbs, by contrast, are almost 95 percent white. Problems of substandard and insufficient housing, delinquency, public health, crime, drug abuse, and other ills, by no means confined to the ghettoes, have somewhat dimmed the city's image. Efforts at public housing date back to 1937. Other municipal and federal bodies have continued efforts at urban renewal for both blacks and poor whites, usually following the pattern of rehabilitating neighbourhoods rather than razing them. Unusual is the fact that nearly 40 percent of the black families own their homes.

For many Philadelphians, a serious new problem was created during the late 1960s by the increasing militancy of many young blacks. This led in 1971 to signs of a backlash in the election as mayor of a tough law and order oriented former police commissioner; his support was surprisingly large, however, within the black community, in which there was fear of militant action.



Central Philadelphia and (inset) its metropolitan area.

An economic overview

The city's economy. More than 150 of Philadelphia's business enterprises were established in the 1700s, but the city's economy has changed greatly in recent decades. Such old and once-dominant industries as textile mills met stiff competition from the South. Industry was given a boost by World War II but began to decline later; thousands of jobs were lost in the 1960s. Nonmanufacturing and service industries, however, continued to increase employment into the 1970s. Although such old industries as textiles were declining, new industries, such as electrical products and appliances, grew. The area was a centre for phonograph and record making, along with radio and television set manufacturers. Apparel, food and food products, printing and publishing, and electrical machinery are leaders in manufacturing. A decline was evident also in the migrating of both people and business from the city to the suburbs. The Valley Forge area became a nest of business and industrial locations. Much of this planning and relocation is under the sponsorship of the public Philadelphia Industrial Development Corporation.

Research facilities. George Washington approved the first U.S. patent in 1790, to a Philadelphian named Samuel Hopkins for a better way of making potash. In today's economy, research and invention are keys to progress. In West Philadelphia is located the University City Science Center and Institute, a nonprofit project of over 20 universities and colleges with a multimillion-dollar commitment to research. General Electric, E.I. du Pont de Nemours & Company, Westinghouse, Monsanto Chemicals, Rohm and Haas, and the Philadelphia Electric Company are other large companies with extensive research programs within the metropolitan area. American Viscose, at Marcus Hook, is the world's largest producer of rayon. U.S. Steel's Fairless plant, on the Delaware at Morrisville, is the largest integrated steel plant built at one time. The area is a centre for research on and refining of petroleum. The computer age was born out of experiments at the University of Pennsylvania, where the first electronic computer was operated in 1945. More than 60 percent of the output of the Campbell Soup Company, at Camden, are foods newly developed in the last decade. A unit of Boeing Company, at Eddystone, is the most advanced helicopter research and design plant in the world.

Financial institutions. From its early days until about 1880, Philadelphia was the financial capital of the United States but lost this position to New York. It was the birthplace of American banking, and the first building and loan association was founded there. The city continues to be a major centre of banking, with its financial institutions remaining the oldest and among the most prosperous in the nation. Philadelphia also supplies the stock and exchange services demanded by modern business. The Philadelphia Commercial Exchange was set up in 1868 as an outgrowth of the older Corn Exchange Association of 1863 to regulate the then-flourishing grain and flour trade. The Philadelphia Bourse was organized in 1891, and the Maritime Exchange was founded in 1875. All three agencies are now housed in the Bourse Building. Philadelphia had the oldest stock exchange in the U.S., founded in 1790. The Philadelphia–Baltimore–Washington Stock Exchange was joined by Pittsburgh in 1969.

Government and municipal rejuvenation. The city–county consolidation in 1854 was a result of the inability of a colonial-type government by committees to adapt to the needs of a growing city for new public services; *e.g.*, better streets, police, transportation, sanitation, and schools. Ironically, Philadelphia had, in fact, led most cities in providing urban improvements, but they were too few and too slow, and the breakdown of law and order in the 1840s forced changes. Until the early 1950s the standard type of U.S. city government prevailed in Philadelphia, comprising an elected mayor and a city council elected from among candidates presented by the political parties in the political divisions, or wards. As in other large cities, this form was subject to pressures for special favours, such as street-railway franchises and public-works contracts; bribery and corruption were both inevitable and rampant.

The Greater Philadelphia Movement

The reform movement that began in 1939 was joined in 1948 by top business and financial leaders, who organized the Greater Philadelphia Movement, and together they took up the battle for an entirely new city charter. This document effectively removed the city council from its administrative role and increased the staff and powers of the mayor. A strong civil-service commission improved professional employment. The council, seven of whose members were elected by all the voters and ten by districts, was to review and approve taxes and budgets and enact ordinances. Philadelphia now had the first modern big-city charter in the United States; its approval by a 2 to 1 popular vote made it an expression of the desire of Philadelphians for better, more efficient, and honest city government.

The first mayors under the new charter were Joseph S. Clark and Richardson Dilworth, men devoted to making it work. From wealthy Republican families, both were lawyers who had revolted against the corruption and inefficiency of city government and had become Democrats. Men of the highest qualifications were selected for key positions, planning was made a virtue, and a $150,000,000 plan was launched at once for improvements in sewage and sanitation, playgrounds, lighting and streets, police and fire protection, and other basic services long neglected. The ambitious plan for a Penn Center involved removal of the old Pennsylvania Railroad "Chinese Wall" of overhead tracks that ran into the heart of the city at Broad Street and replacing it and Broad Street Station.

Renewal planning gave close attention to the need for better housing: the Philadelphia Housing Authority and the Philadelphia Redevelopment Authority, with federal and private aid, approached Philadelphia's housing in terms of improving neighbourhoods rather than total clearance followed by new high-rise apartments.

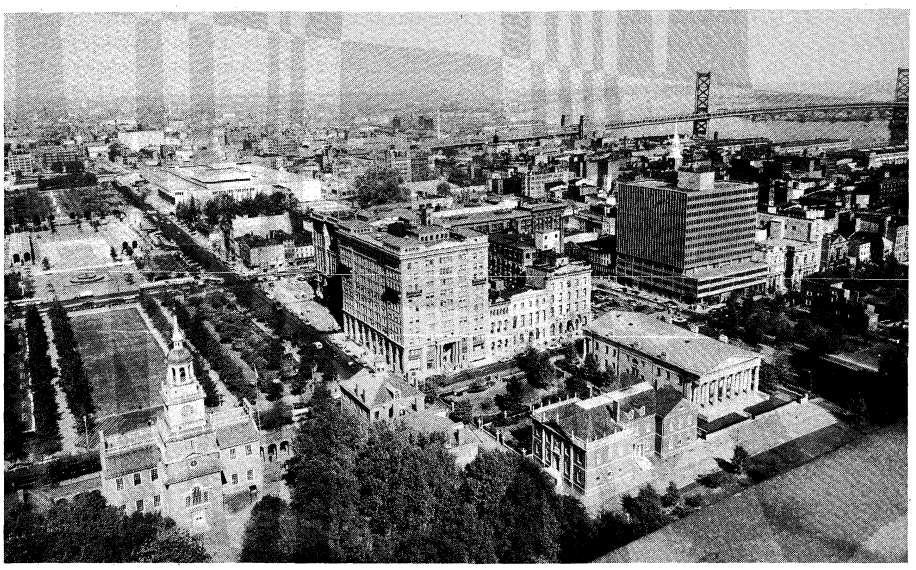
The preservation and restoration of the Independence Hall area was begun in cooperation with the National Park Service, and the restoration of neighbouring Society Hill was put in motion. This was not new housing for the poor but an effort that bore fruit to attract people in higher income brackets to restore or build houses in this historic inner-city district. Plans for improvement of the historic waterfront as a Penn's Landing area were launched.

Much of the work of the new city government is carried out through such housing redevelopment, parking, and other authorities. The Old Philadelphia Development Corporation was organized in 1956 as a quasi-public and private body to aid the new development plans. Proposals for Franklin Town, a development of housing and offices on a 50-acre tract north of the Benjamin Franklin Parkway, were launched in 1971. As in other similar undertakings, private enterprise was made an integral part of the project.

Public services. In 1970, more than 35,000 persons were on the municipal payroll, exclusive of the independent public-school unit. The largest single group was the police, who gave Philadelphia one of the best records of any American city in the reported crime rate. Street maintenance, fire protection, water supply, and recreation followed in number employed.

Though the Philadelphia metropolitan area has a per capita income that is the highest of any such area in the state, a large percentage of Philadelphians live below the poverty level. City, state, and federal agencies administer a full range of ameliorative social-service programs. The Philadelphia General Hospital cares for the indigent, and a department of public health operates a variety of services through ten health districts. With its Quaker tradition, Philadelphia is famous for its humanitarian concern and has a great variety of privately supported child-care, hospital, and other social services.

Transportation. A complex system of public and private trolley and bus lines was consolidated in 1963 by the state legislature's creation of the Southeastern Pennsylvania Transportation Authority (SEPTA) to plan, develop, and coordinate a regional transportation system in five southeastern counties and to fund projects by the sale



Philadelphia, with Independence Hall, Library Hall, and the Second Bank of the U.S. in the foreground. In the rear are the low-lying U.S. Mint on the left and the Benjamin Franklin Bridge.
H. Armstrong Roberts

of bonds. This system also contracts with the Reading and Penn Central railroads to service 12 radial commuter lines. Express buses provide rapid service between the inner and outer city.

A joint New Jersey–Pennsylvania commission operates five toll and 13 free bridges over the Delaware. The Schuylkill is bridged at a number of points and has a subway tunnel. The New Jersey–Pennsylvania–Delaware port authority administers Ameriport and the Benjamin Franklin and Walt Whitman bridges over the Delaware, the latter, with approaches, six miles in length. In the 1960s and early 1970s, it has carried out major improvement and expansion programs throughout the entire port area. Philadelphia International Airport, less than seven miles from the city centre, provides good nationwide and increasing numbers of international flights.

Food supply. The unique Food Distribution Center, a nonprofit corporation managed by a board of directors representing city government and private enterprise, is a prime example of how Philadelphia has joined the work of the private and public sectors to serve the best interests of both. Covering 388 acres, it is a food-industry park handling in a unified operation every food-marketing facility from ripening bananas to smoking fish; it comprises well over 100 stores plus warehousing and processing plants.

Cultural life and institutions. Every great city has a certain character not always easy to define. Much of the essence of Philadelphia lies in the features described previously—its small-town atmosphere, its parks and tree-shaded downtown square and streets, and its innumerable memorials to the American past, as well as its teeming riverside and factories and its diverse business institutions. There are other factors as well that contribute in their way to an understanding of Philadelphia's culture, considered in its broadest implications to comprise the life-styles of its people.

Parks and playgrounds. In addition to Fairmount Park, Philadelphia has Pennypack Park in the northeast, a semiwilderness setting with bridle paths, bird-watching trails, and an abundance of deer and other wildlife. More than 100 parks with playgrounds are located in all areas of the city, including the ghettoes.

Recreational facilities

Philadelphians have always been a sports-loving group, whether passive or participatory, though professional teams have always been the object, like the city itself, of a good-natured deprecation that is tolerated in natives but not in strangers. Fox hunting in the surrounding countryside is of old Quaker origins, while the Schuylkill is a major rowing site for collegiate and older individuals and crews, and Germantown harbours remnants of a once-lively city-wide enthusiasm for cricket.

The arts. In colonial days, Philadelphia was known as the "Athens of America," and it retains a high place in the artistic achievement of the nation. The Academy of Music, opened in 1857, is the home of the Philadelphia Orchestra, which is among the very finest orchestras of the world and was the first to broadcast and make records. The Robin Hood Dell Orchestra has presented outdoor concerts in Fairmount Park since 1930.

Philadelphia was the nation's theatrical centre until well after the Revolution, its stages having hosted the greatest players of Europe and America during three centuries. The Walnut Street Theatre, opened in 1809, is the oldest playhouse in active use in the English-speaking world. The Playhouse in the Park opened in 1952 as the first city-owned and -operated theatre of its kind. Professional theatre revolves generally around new plays trying out on their way into New York or road companies of proved Broadway hits.

Museums and galleries. Philadelphia was a pioneer in museums of all kinds. Charles Willson Peale's museum was housed in Independence Hall in the 1800s, but today the art museums are led by the Philadelphia Museum of Art. One of the world's great museums, it houses priceless collections of art from the Western Middle Ages onward and from the Orient. Others include the unique Rodin Museum, featuring the largest collection of sculptures by Auguste Rodin outside of Paris. The Historical Society of Pennsylvania has some 700 portraits by Benjamin West, James Copley, Gilbert Stuart, the four Peales, and other early American painters. The output of visual artists in Philadelphia and vicinity has been by and large conservative, though Thomas Eakins gained fame for work beyond his American contemporaries, and Mary Cassatt was among the few women in the Impressionist school of the late 19th century.

Other museums include the Civic Center Museum, originally organized in 1894 as the Commercial Museum to promote trade exhibitions. The Academy of Natural Sciences is the oldest of its kind in the U.S. The Franklin Institute Science Museum is full of marvellous things that move and can be moved, while the University Museum is a major feature of the University of Pennsylvania. There are, finally, many small museums housed in the restored buildings throughout the city.

Education. Philadelphia's public-school system is an independent governmental unit operating under a board of education appointed by the mayor on the recommendation of a citizens' nominating panel. The city's institutions of higher learning are headed by the privately endowed University of Pennsylvania, which was founded in 1740, opened the nation's first medical school (1765), and was the first designated "university" in the U.S.

Institutions of higher education

(1779). Temple University (1884) has acquired the reputation of "the people's college" and is known for work in the arts and social sciences. Catholic colleges include LaSalle and St. Joseph's. Philadelphia is also a centre for medical education, with several general schools and specialized schools in pharmacy, podiatry, optometry, and osteopathy. The Curtis Institute of Music (1924) is a leading music school, while the Pennsylvania Academy of Fine Arts (1805) is the oldest art school in the United States.

The media. *The Philadelphia Inquirer* and *The Philadelphia Evening Bulletin* are the city's major dailies, though *The Philadelphia Courier* has a long history of service to the black community. Several weeklies serve sections of the city or various linguistic and religious groups. The magazine *Philadelphia* appeared first in 1909; since the 1950s it has published lively and often controversial stories and commentary on the city's social and political life. The numerous television and radio channels include both commercial and "public" broadcasting.

Prospects. The Quakers represent only a very small percentage of Philadelphians in the 20th century, but they continue to exert, by both tradition and present-day activity, an influence far exceeding their numbers. It seems unlikely that Philadelphia will soon lose the pervading quiet and dignity that they more than any other force have implanted on the city. It is too early to estimate the impact of the militancy of the city's ghettoes or of the possibility of stern responses from the various power structures. In terms of ongoing civic renewal, successors of mayors Clark and Dilworth have been accused of a lessening appetite for the goals of the original reformers. Certainly the test of years under the charter of 1951 has shown the need for improvement and for change to most still newer circumstances; Philadelphia's renaissance, however—in its spiritual as well as its physical body—continues to command attention from American urban leaders.

BIBLIOGRAPHY

General literature: *Bulletin Almanac* (annual), an indispensable review; *Delaware Valley, U.S.A.* (annual), a review of the economy and transportation by the *Philadelphia Inquirer*; *Official Handbook for Visitors* (n.d.), an excellent city-published guide; *Philadelphia Magazine*, a periodical that often features a sharply critical and satirical approach.

Historical works: CARL and JESSICA BRIDENBAUGH, *Rebels and Gentlemen: Philadelphia in the Age of Franklin* (1942), an extremely good picture of colonial and Revolutionary times and people; BROOKE HINDLE, *The Pursuit of Science in Revolutionary America, 1735–89* (1956), the best single source for early science in Philadelphia; AMERICAN PHILOSOPHICAL SOCIETY, *Historic Philadelphia from the Founding Until the Early Nineteenth Century* (1953), a study in depth; HORACE MATHER LIPPINCOTT, *Early Philadelphia: Its People, Life, and Progress* (1917), a fine study that is both colourful and informative; *Pennsylvania Magazine of History and Biography* (quarterly), the prestigious journal of the Historical Society of Pennsylvania (well indexed); SYLVESTER K. STEVENS, *Pennsylvania, Birthplace of a Nation* (1964), the first scholarly one-volume history, fully integrates Philadelphia's growth with that of the state.

Specialized studies: NATHANIEL BURT, *The Perennial Philadelphians: The Anatomy of an American Aristocracy* (1967), an excellent study of Philadelphia society; MAWELL STRUTHERS BURT, *Philadelphia, Holy Experiment* (1945), a good general picture with great feeling for the city; W.E.B. DU BOIS, *The Philadelphia Negro: A Social Study* (1899, reprinted 1967), good for the subject; JEANNE R. LOWE, *Cities in a Race with Time* (1967), an excellent study of urban problems and renewal, includes a good history of Philadelphia; GEORGE TATUM, *Penn's Great Town: 250 Years of Philadelphia Architecture* (1961), a valuable review of the cultural heritage; WILLIAM S. VARE, *My Forty Years in Politics* (1933), a first-hand account of city machine politics.

(S.K.S.)

Philip II Augustus, of France

Philip II Augustus, the first of the great Capetian kings of medieval France, destroyed the English-based Angevin Empire and greatly extended the influence of his dynasty. Born in Paris on August 21, 1165, the son of Louis VII of France and Adela of Champagne, Philip Augustus was crowned at Reims on November 1, 1179, so as to be associated as king with his father, who had fallen mortally ill. His uncles of the House of Champagne—Henry I, count of Champagne, Guillaume, archbishop of Reims, and Thibaut V, count of Blois and Chartres—hoped to use the youthful king to control France. To escape from their tutelage, Philip, on April 28, 1180, married Isabella, the daughter of Baldwin V of Hainaut and the niece (through her mother) of Philip of Alsace the count of Flanders, who promised to give the King the territory of Artois as her dowry. When Henry II of England arrived in Normandy, perhaps with the intention of responding to an appeal by the House of Champagne, Philip II entered into negotiations with him and, at Gisors on June 28, 1180, renewed an understanding that Louis VII had reached with him in 1177. As a result, the House of Champagne was politically isolated and Philip II was making all decisions for himself and acting as he saw fit when his father died, on September 18, 1180, leaving him sole king in name as well as in fact.



Philip II, Augustus seal of majesty, showing the king crowned and enthroned, from a document of 1180. In the Archives Photographiques, Paris.
Archives Photographiques, Paris

When the Count of Flanders allied himself with the Champagne faction, there followed a serious revolt against the King. In the Peace of Boves, in July 1185 (confirmed by the Treaty of Gisors in May 1186), the King and the Count of Flanders composed their differences (which had been chiefly over possession of Vermandois, in Picardy), so that the disputed territory was partitioned, Amiens and numerous other places passing to the King and the remainder, with the county of Vermandois proper, being left provisionally to Philip of Alsace. Thenceforward the king was free to run against Henry II of England.

Territorial expansion. *The struggle against the Plantagenets.* Henry's French possessions, the so-called Angevin Empire—consisting of Normandy, Maine, Anjou, and Touraine, with Aquitaine in the hands of his son, the future Richard I of England, and Brittany ruled by another son, Geoffrey (died 1186)—all were a constant menace to the French royal domain. Furthermore, there were long-standing disputes over the Vexin (between Normandy and the Île-de-France), Berry, and Auvergne.

Philip II launched an attack on Berry in summer 1187 but then in June made a truce with Henry, which left Issoudun in his hands and also granted him Fréteval, in Vendômois. Though the truce was for two years, Philip found grounds for resuming hostilities in the summer of 1188. He skillfully exploited the estrangement between Henry and Richard, and Richard did homage to him voluntarily at Bonmoulins in November 1188. Finally, by the Treaty of Azay-le-Rideau, or of Colombières (July 4, 1189), Henry was forced to renew his own homage, to confirm the cession of Issoudun, with Graçay also, to Philip, and to renounce his claim to suzerainty over Auvergne. Henry died two days later.

Philip Augustus and the Third Crusade

Richard, who succeeded Henry as king of England, had already undertaken to go on crusade (the Third Crusade) against Saladin in the Holy Land, and Philip now did likewise. Before his departure, he made the so-called Testament of 1190 to provide for the government of his kingdom in his absence. On his way to Palestine, he met Richard in Sicily, where they promptly found themselves at variance, though they made a treaty at Messina in March 1191. Arriving in Palestine, they cooperated against the Muslims at Acre, until Philip fell ill and made his illness a pretext for returning to France, quite determined to settle the succession to Flanders (Philip of Alsace had just died on the crusade) while Richard was still absent. Thus, by the end of 1191, Philip II was back in France.

In spite of promises he had made in the Holy Land, Philip at once prepared to attack the Plantagenet possessions in France. Informed of this, Richard also left the crusade but was taken prisoner while on his way back by the duke of Austria, Leopold V of Babenberg. Philip did everything he could to prolong his rival's captivity, but Richard was at last set free (1194) and went to war against Philip. The French king suffered a number of defeats (from that at Fréteval in July 1194 to that at Courcelles in September 1198) in a series of campaigns that were occasionally punctuated by negotiations. It was fortuitous for Philip, however, when Richard was killed in April 1199.

Richard's brother John was by no means as formidable a fighter. Moreover, his right to Richard's succession could be contested by Arthur of Brittany, whose father had been senior to John. To secure the succession, therefore, John came to terms with Philip: by the Treaty of Le Goulet (May 22, 1200), in return for Philip's recognition of him as Richard's heir, he ceded Évreux and the Norman Vexin to Philip; agreed that Isoudun and Graçay should be the dowry of his niece Blanche of Castile, who was to marry the future Louis VIII (Philip's son by Isabella of Hainaut); and renounced any claim to suzerainty over Berry and Auvergne. Shortly afterward, however, John entered into conflict with the Lusignan family of Poitou (in Aquitaine), who appealed to Philip as overlord. When he was summoned to appear before the royal court as a vassal of the French crown, John did not present himself, and Philip, in April 1202, pronounced John's French fiefs forfeit and undertook to carry out the sentence himself. He invaded Normandy, overran the northeast, and laid siege to Arques, while Arthur of Brittany, the son of Geoffrey, who died some years before, campaigned against John's supporters in Poitou; but John, marching south from Maine, captured Arthur at Mirebeau (August 1). In fury, Philip abandoned the siege of Arques and marched southwestward to Tours, ravaging John's territory on his way before returning to Paris. Guillaume des Roches, the powerful seneschal of Anjou, who had taken John's side, came to terms with Philip in March 1203.

Philip's conquest of Normandy

Resuming operations against Normandy, Philip occupied the towns around the great fortress of Château-Gaillard, to which he laid siege in September 1203, having overruled Pope Innocent III's attempts to mediate. John, who is reported to have murdered Arthur of Brittany in April, retired to England in December, and Château-Gaillard fell to Philip in March 1204. Rouen, the Norman capital, surrendered in June, after 40 days' resistance.

After his conquest of Normandy, Philip subdued Maine, Touraine, Anjou, and most of Poitou with less difficulty (1204–05), though the castles of Loches and Chinon held out for a year. He sought to secure his conquests by lavishing privileges on the towns and on the religious houses but otherwise left the local barons in power. Unrest, however, was endemic in Poitou, and in June 1206 John landed at La Rochelle. After a campaign in the south, he turned north toward the Loire. At Thouars in October 1206, he and Philip made a two-year truce, leaving John in possession of the reconquered Poitevin lands. In the following year, however, Philip invaded Poitou again; and, after a further campaign in 1208, only the south and part of the west of Poitou remained loyal to John (with Saintonge, Guyenne, and Gascony).

Philip next hoped to exploit the dispute between John and Pope Innocent III. While Innocent was threatening to declare John unfit to reign (1212), plans were being made for a French landing in England and for the accession of Philip's son Louis to the English throne. The plans had to be dropped when John made his submission to the Pope (1213). Throwing himself into schemes for revenge, John formed a coalition against France: the Holy Roman emperor Otto IV, the Count of Flanders (Ferrand, or Ferdinand, of Portugal), and the Count of Boulogne (Raynald, or Renaud, of Dammartin) were to invade the Capetian territory from the northeast while John attacked from the west, with the help of his Poitevin barons. John landed at La Rochelle in February 1214 and advanced into Anjou but was put to flight by Louis at La Roche-aux-Moines on July 2; his confederates were completely defeated by Philip in the decisive Battle of Bouvines on July 27. The Anglo-Angevin power in France and the coalition had both been broken in one month. Thus Philip, who, in 1213, had transferred Brittany to his cousin Peter of Dreux, was left without any significant opposition to his rule in France.

The Battle of Bouvines

The northeastern frontier and Languedoc. It was not only at the Plantagenets' expense that Philip enlarged the royal domain. His claim to Artois through his first marriage and his gains by the settlement of 1185–86 have been mentioned above, and he subsequently proceeded, step by step, to acquire the rest of Vermandois and Valois. His insistence on his suzerainty over vacant fiefs and on his tutelage over minors and heiresses was particularly effective with regard to Flanders, where two successive Flemish counts, Philip of Alsace (died 1191) and Baldwin IX (died *c.* 1205) had left no male issue.

Though he did not personally take part in the crusade proclaimed by Pope Innocent III against a Cathari religious sect in Languedoc, Philip Augustus allowed his vassals and knights to carry it out. Simon de Montfort's capture of Béziers and Carcassonne (1209) and his victory at Muret over Raymond VI of Toulouse and Peter II of Aragon (1213) prepared the way for the eventual annexation of eastern Languedoc to the royal domain six years after Philip's death and for the union of northern and southern France under Capetian rule.

Internal affairs. *The matrimonial conflict with the papacy.* Several years before he tried to take advantage of the papacy's quarrel with John of England, Philip Augustus had himself been in dispute with Rome. After the death (1190) of Isabella of Hainaut, he had married Ingeborg, sister of the Danish king Canute VI, on August 14, 1193, and on the next day, for a private reason, had resolved to separate from her. Having procured the annulment of his marriage by an assembly of bishops in November 1193, he took a Tirolese lady, Agnes, daughter of Bertold IV of Meran, as his wife in June 1196. Denmark, meanwhile, had complained to Rome about the repudiation of Ingeborg, and Pope Celestine III had countermanded it in 1195; but Celestine died (1198) before he could resort to coercion against Philip. The next pope, Innocent III, was sterner: in January 1200 he imposed an interdict on France. Philip, therefore, in September 1200, had to submit, pretending to be reconciled with Ingeborg. In fact, he refused to cohabit with her and kept her in semicaptivity till 1213, when he accepted her beside him—not as his wife but at least as his queen. Agnes had died in 1201, after bearing two children to Philip: Marie, countess of Namur (1211) and duchess of Brabant (1213), by successive marriages; and Philip, called Hurepel; count of Clermont.

Submission to Innocent III

Philip's relations with his subjects. Throughout his reign, Philip Augustus kept a close watch over the French nobility, which he brought effectively to heel. He maintained excellent relations with the French clergy, leaving the canons of the cathedral chapters free to elect their bishops and favouring the monastic orders. He knew, too, how to win the support of the towns, granting privileges and liberties to merchants and frequently aiding their struggles to free themselves from the seignorial au-

thority of the nobles. In return, the communes helped financially and militarily. Most of all, Philip gave his attention to Paris, not only fortifying it with a great rampart but also having its streets and thoroughfares put in order. For the countryside, he multiplied the number of *villes neuves* ("new towns"), or enfranchised communities.

The Capetian monarchy's hold on the huge royal domain as well as on the kingdom as a whole was considerably strengthened by Philip's institution of a new class of administrative officers, the royal *baillis* and the seneschals for the provinces, who were appointed by the king to supervise the conduct of the local *prévôts* ("provosts"), to give justice in his name, to collect the revenues of the domain for him, to call up the armed forces, and so forth.

Conclusion. Philip II Augustus died at Mantes on July 14, 1223. Knowing his own strength, he was the first of the Capetians not to have his eldest son crowned and associated with him during his lifetime; in fact, his conquests and strong government made him the richest and most powerful king in Europe, and prepared the way for France's greatness in the 13th century. (See also FRANCE, HISTORY OF.)

BIBLIOGRAPHY. For a concise outline of the reign of Philip Augustus, read the corresponding chapters of two famous publications: *The Cambridge Medieval History*, vol. 6; and BRIAN TIERNEY and SIDNEY PAINTER, *Western Europe in the Middle Ages, 300–1475* (1970). For the relationship of Philip Augustus to Capetian policy and the development of the monarchy, see R. FAWTIER, *Les Capétiens et la France* (1942); or C. PETIT-DUTAILLIS, *La Monarchie féodale en France et en Angleterre, X^e–XIII^e siècle*, 2nd ed. (1950). Three books that give a very circumstantial history of the reign are: A. CARTELLIERI, *Philipp II August, König von Frankreich*, 4 vol. (1899–1922); C. PETIT-DUTAILLIS and P. GUINARD, *L'Essor des états d'occident* (1937); and A. LUCHAIRE, *Louis VII, Philippe-Auguste, Louis VIII* (1901).

(M.Pa.)

Philip IV the Fair, of France

An accomplished statesman, yet scrupulously attentive to his conscience, Philip IV the Fair, king of France from 1285 to 1314, established the French monarchy's prestige and authority over matters temporal and spiritual within Christian Europe and confirmed the supreme power of the king within the realm of France. He also instituted important reforms in the conduct of government. His determination to defeat any challenge to his authority within his kingdom was demonstrated in his lengthy duel with Pope Boniface VIII and in his relations with the Holy Roman Empire and other western European powers of the late 13th and early 14th centuries. The claim that the French monarch should by right serve as arbiter of Christian orthodoxy and morality was established not only by Philip's appeal to a general council of the church against Boniface and his campaign against the pope's reputation, but also by his assault on the crusading order of the Knights Templars. Within France Philip's reign saw the beginnings of large-scale consultation of royal subjects in assemblies, the issuance of important reform charters, and the development of policies permitting the marshalling of the realm's forces to meet attack.

Early years. Born at Fontainebleau in 1268 while his grandfather was still ruling, Philip, the second son of Philip III the Bold and grandson of St. Louis (Louis IX), was not yet three when his mother, Isabella of Aragon, died on her return from the crusade on which Louis IX had perished. The motherless Philip and his three brothers saw little of their father, who, stricken by Isabella's death, threw himself into campaigning and administrative affairs. His troubled childhood and the series of blows he suffered explain in some measure the conflicting elements in his adult personality. In 1274 his father married Marie de Brabant, a beautiful and cultivated woman, and with her arrival at court intrigue began to flourish. In the same year the two-year-old Jeanne, heiress of Champagne and Navarre, was welcomed as a refugee. Reared with the royal children, she would, when she was 12, become the bride of Philip the Fair.



Philip IV the Fair, detail of the statue from his tomb by an unknown artist, 14th century. In the abbey church at Saint-Denis, France.
Archives Photographiques, Paris

In 1276 Philip's older brother Louis died, and the shock of this event, which suddenly made Philip heir of the kingdom, was compounded by persistent rumours of poisoning and suspicions that Philip's stepmother intended to see Isabella's remaining sons destroyed. Vague allegations were circulated that Louis's death was linked with certain unspecified "unnatural acts" of his father. These rumours, never satisfactorily put to rest, together with the unexpected change in Philip's fortunes, apparently served to arouse in him feelings of insecurity and mistrust. Consequently, Philip turned elsewhere in search of a model for his own conduct. He found it in Louis IX, whose memory was increasingly venerated as the number of miracles attributed to him mounted. Reports of Louis's exacting standards of rulership and his saintly virtues were reinforced by the precepts of the religious advisers who surrounded the adolescent Philip. A more self-confident person might have been able to discriminate realistically among the sometimes artificially exaggerated stories and the utopian ideals. Philip, however, became convinced that it was his God-given duty to attain the lofty goals of his grandfather.

Influence of Louis IX

When Philip was 16, he was knighted and married to Jeanne of Navarre. In 1285 he accompanied his father to the south on a campaign to install Philip's brother Charles on the throne of Aragon. But he had no sympathy with the enterprise, which was backed by his stepmother and aimed against the King of Aragon, his late mother's brother. When his father died in October 1285, Philip immediately abandoned the venture.

Wars with England and Flanders. In the first years of Philip's reign the Aragonese affair was settled, and Philip intensified his predecessors' efforts to reform and rationalize the administration of the realm. He dispatched investigators to inquire into the conduct of royal officials and into infringements upon royal prerogatives. Philip persisted in seeking such reforms, which strengthened the monarchy's position but angered the nobles, townsmen, and ecclesiastics who had profitted from the laxer policies of earlier kings.

War with England began in 1294, initiating a ten-year period of conflict that severely strained Philip's resources. There had been some naval clashes, but full-scale war might well have been avoided had not Philip, perhaps in a fit of youthful bravado, decided to demonstrate his power over England's mighty Edward I, his vassal, for control of the duchy of Gascony. Philip's victories in 1297 may have satisfied his ambitions; but they brought no territori-

Repercussions of the war with England

al gain, for the many lands that Philip had seized were returned. Nevertheless the war was significant. First, the peace treaty of 1303 stipulated that Philip's daughter Isabella should marry the future Edward II of England—an alliance that led to years of peace between the two kingdoms. Second, during the war, Philip's vassal Guy of Dampierre, count of Flanders, had allied himself with Edward I, a move that Philip considered base treachery and that resulted in a breach between the two that persisted until long after Philip's death.

Before the peace, but after fighting with England had ceased, Philip moved to crush the Flemings, only to see a host of his nobles fall at Courtrai in 1302. This stunning defeat was redeemed two years later at Mons-en-Pévèle, where Philip showed great personal courage. Finally, in 1305, Philip forced Flanders to accept a harsh peace treaty that exacted heavy reparations and humiliating penalties. In financing the prolonged effort against Flanders, Philip had held assemblies and had bartered privileges and promises of reform for support in the war. Most important, in abandoning the property taxes that earlier had been levied for defense, he enforced the principle that all must fight to defend France but might purchase exemption if they wished. This successful policy was later employed as a regular wartime expedient by the French monarchy.

Conflict with the papacy. Philip's rupture with Boniface VIII can be considered a third consequence of the English war. Because the hostilities interfered with papal plans for a crusade, Boniface intervened aggressively and sometimes tactlessly to promote peace. In February 1296 he issued the bull *Clericis Laicos*, prohibiting lay taxation of clergy without papal approval. Both Edward I and Philip, affronted by this threat to their authority and their treasuries, responded with retaliatory measures, forcing Boniface to retreat and, in July 1297, to proclaim the legitimacy of clerical taxation without the pope's permission when the ruler attested its necessity. To mollify Philip, Boniface supported him against the Flemings and canonized Philip's grandfather Louis IX in 1297. The Pope's position in France was even weaker than he realized, for, at least as early as 1297, his enemies had spread charges against him, such as that he questioned the immortality of the soul or that he had plotted the death of his predecessor on the papal throne. Philip, for his part, had not rejected these stories out of hand.

New grounds for dispute developed in 1301, when Philip arrested Bernard Saisset, bishop of Pamiers, as a suspected traitor. Boniface, his confidence buoyed after the triumphant papal jubilee of 1300, determined not only to force Philip to send Saisset to Rome, but also to launch a frontal attack on the king's authority. In December 1301 Boniface suspended Philip's right to tax ecclesiastics and summoned the French clergy to Rome to discuss the King's governance and the state of the French Church. Saisset was permitted to go to Rome, but Boniface's other measures encountered immediate resistance. Philip had a papal bull burned and ceremoniously invoked curses on any of his sons who dared subordinate the kingdom to any power other than God's. In April 1302 he rallied public support in a large assembly. Undaunted by his humiliating defeat by the Flemings at Courtrai and by Boniface's declaration of the universal supremacy of the Roman pontiff in the bull *Unam Sanctam*, Philip held additional assemblies in the spring of 1303. He issued his own grand ordinance of reform that included remedies for administrative weaknesses enumerated by the Pope. Then, in response to appeals by his ministers Guillaume de Nogaret and Guillaume de Plaisians, Philip pledged to see Boniface judged for the heretical words and criminal and immoral deeds with which he had been charged.

Arrest of Boniface VIII

Boniface's plans to issue a personal sentence of excommunication against Philip were forestalled when Nogaret appeared in Anagni and seized Boniface on September 7, 1303. Nogaret probably intended to press Boniface to submit to conciliar judgment, but his plans were frustrated when troops, rallied by Boniface's Italian enemies, turned to violence and pillage. Two days later townsmen of Anagni freed the Pope, whose death the following month saved him from having to appear before a council to answer Philip's charges. Accusations against Boniface's memory, however, proved a useful negotiating device in dealing with his successor, Benedict XI, and even more so with Clement V, the Gascon-born pope who pleased Philip by transferring the papal curia from Rome to Avignon, a city near Philip's realm. Charges against Boniface were pressed until 1311, when Clement declared Philip's zeal praiseworthy and nullified all the offensive bulls that Boniface had issued after November 1301.

The years of peace between 1304 and 1313 allowed Philip not only to pursue his posthumous vendetta against Boniface but also to formulate plans to establish himself as the undisputed champion of orthodoxy. Before 1304 Philip had given thought to the moral implications of his deeds, taking steps to indemnify those harmed by coinage changes and advancing approved justifications of other wartime measures. After 1304 his scrupulousness became more apparent. He secured bulls from Pope Clement that absolved him from future crusading vows, annulled any obligation to return sums wrongfully taken from ecclesiastics for his wars, and—after lengthy negotiation—declared the Flemings subject to ecclesiastical sanctions if they failed to observe the peace treaty of 1305. In 1306 reformation of the coinage, long advocated by the church, was begun, and in later years Philip sought his subjects' advice concerning monetary policy. Philip's intensified preoccupation with questions of conscience and morals may have been prompted by the death in April 1305 of Queen Jeanne, a determined woman of scholarly inclinations and a devotee of Saint Louis. Philip never married again, and in the months after his wife's death he considered abdicating and assuming the kingship of the Holy Land as head of a consolidated crusading order.

Persecution of the Jews and Knights Templars. Philip soon found more practical ways of venting his grief and proving his dedication to God, while also advancing the material interests of a realm that was impoverished by long years of warfare. In 1306 he expelled all Jews from France, seizing their property and confiscating the monies owed to them. Philip's devotion to St. Louis was witnessed that same year by elaborate ceremonies in his honour, and Louis's anti-Semitic proclivities might have inspired Philip to act against the Jews, whose usefulness as a source of regular revenue had in any case been exhausted by his earlier repeated impositions.

Campaign to crush the Knights Templars

Similarly mixed motivations influenced Philip to attack the Knights Templars, the wealthy, powerful, independent crusading order that had long acted as the French monarchy's financial agent. Philip's newfound interest in uniting the crusading orders made him mistrustful of the Templars' opposition to such plans. Thus, he was receptive to charges of heresy and sodomy presented against them in 1305. His tentative overtures to Clement V were fruitless, but, with the support of Nogaret and his own Dominican confessor—who was also the papal inquisitor in France—Philip decided in September 1307 to seize all Knights Templars in France and to exhort his fellow rulers to follow his lead. At first dubious and reluctant, Clement V eventually supported Philip; he had been told of outraged anti-Templars appeals voiced in a large assembly of Philip's subjects, and he had heard damning confessions from the mouths of representatives of the order, many of whom had been tortured. Far-flung tribunals had gathered enough materials to cast doubts on the Templars' dedication, and, although not condemned as heretical, the order was quashed and its property assigned to the Knights Hospitallers. Whatever Philip's reasons for launching his attack against the order, his action brought him substantial financial gain despite Clement V's repeated attempts to protect the Templars' interests.

During the interwar years Philip asserted his independence of the Holy Roman Empire in diplomatic exchanges with imperial princes and the emperors themselves. He also fortified his eastern border by arranging marriages for his sons that extended French influence over the county of Burgundy and by exerting his authority over the city of Lyons. Less successful were his attempts to implement

his dream of gaining control of the empire. He failed to have his brother Charles elected emperor in 1308, nor did he succeed on behalf of his second son Philip in 1313.

Last years. This disappointment presaged the troubles of Philip's last year as king. In June 1313 his fortunes had reached a high point. Having knighted his sons, taken the crusader's cross, and issued coinage-reform ordinances, he witnessed the triumphal departure of his sons against the Flemings, who had been excommunicated for their failure to observe the treaty of 1305. When the Flemings capitulated and accepted a truce, Philip magnanimously ordered the return of money collected for the army. After these successes and demonstrations of royal grace, 1314 brought only troubles. Philip's stubborn resolve to defend morality and the faith was shown when, with royal acquiescence, the grand master of the Templars, Jacques de Molay, was burned at the stake after retracting his earlier confession. Far graver were the public trial and execution of two young squires convicted of adultery with Philip's daughters-in-law.

Scandal of the King's daughters-in-law

The death of Clement V was another blow, but an even greater one was the Flemings' fresh uprising. With his subjects' agreement, Philip mounted another campaign, but the negotiation of a truce caused difficulties. There were rumours of treason involving one of Philip's ministers and insistent demands for the restitution of all money collected for the war. Philip, who needed the funds, delayed, and a coordinated movement of opposition developed. Nobles in northern and eastern France presented their grievances to the King, who by early November was lying ill at Poissy, St. Louis's birthplace, recovering from a minor stroke. Having regained strength, he travelled to his own birthplace, Fontainebleau, and there, a day before his death on November 29, 1314, he halted collection of his last tax and provided for a crusading expedition to be conducted in his name.

Assessment. Philip was aloof, sternly reserved, and physically striking—tall, fine-featured, blond—hence his name, "the Fair." Frankly enjoying the trappings of monarchy, he built and maintained the palaces befitting his estate, but his chief concern was for the dignity of his office. He believed that he and his kingdom were guarded by God and chosen as special defenders of the Catholic faith. He valued the curative powers that he believed God's grace had given him. Proud of his orthodoxy and inspired by a priestly sense of mission, he was receptive to accusations of sin and heresy levelled against others, and he prosecuted ruthlessly those charged with such failings. A visionary idealist, he also possessed a fundamental sense of the possible and the necessary, and those of his grandiose schemes that exceeded the bounds of practicality were abandoned. As authoritarian and determined as he sometimes appeared, he lacked the firmness and independence of judgment that a surer sense of confidence in himself might have afforded him. He often turned for guidance to the intelligent and able ministers whom he promoted and protected, and he solicited the support of his subjects, who were called together with unprecedented frequence to endorse and sometimes to help construct royal policy. Such consultation may have been politically expedient, but, like the expressions of reassurance obtained from spiritual authorities, it also served to convince Philip of the legitimacy of his actions.

Philip was, by all accounts, an impressive ruler. Some Italian writers detested him and his line, but most contemporaries acknowledged his piety and good intentions, although many of them disapproved of his wars, his taxes, and his influential advisers. In the later years of the 14th century, when memory of these sources of complaint had faded, his reign was considered a golden age of freedom from burdensome and unjust taxation. Historians of recent times have differed in their judgments, and nationalistic and ecclesiastical biases have affected many assessments. It now seems possible to conclude, however, that the most striking accomplishments of his reign bear the ultimate impress of the complex personality of a scrupulous, suspicious, rigorous, visionary, determined man who fervently believed himself called by God to act as judge of the morals and welfare of his subjects.

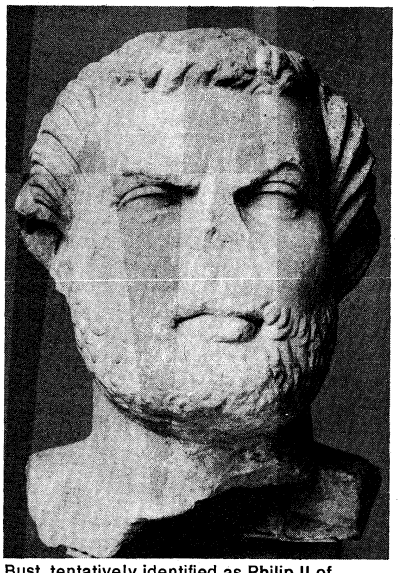
BIBLIOGRAPHY. ROBERT FAWTIER, *L'Europe occidentale de 1270 à 1380*, pt. 1, *De 1270 à 1328*, vol. 6 of GUSTAVE GLOTZ (ed.), *Histoire générale, histoire du moyen âge* (1940), a balanced, stimulating, comprehensive evaluation by a renowned scholar who devoted his life to studying the Capetian dynasty; JOSEPH R. STRAYER, *Medieval Statecraft and the Perspectives of History*, pp. 107–122, 195–247, 291–314 (1971), essays illuminating different aspects of Philip's reign and providing excellent bibliographical information, written by a U.S. medievalist who has been a pathfinder in Capetian scholarship.

(E.A.R.B.)

Philip II of Macedonia

Philip II, king of Macedonia, made his country the ruling power in Greece and laid the foundations for Macedonia's expansion under his son Alexander the Great. Philip was born in 382 BC, a son of Amyntas III. In his boyhood he saw the Macedonian kingdom disintegrating while his elder brothers Alexander II and Perdiccas III, who each reigned for a few years, strove unsuccessfully against insubordination of their regional vassal princes, intervention of the strong Greek city Thebes, and invasion by the Illyrians of the northwest frontier. Philip himself spent some time as a hostage at Thebes, the leading city (with Athens) of this decade (370–360 BC), where the great Epaminondas, the most inventive tactician of all Greek generals until then, was in charge of the best army in Greece. These were probably the most formative years of Philip's education. When he returned to Macedonia his brother Perdiccas soon found him ready for a command.

By courtesy of the Ny Carlsberg Glyptotek, Copenhagen



Bust, tentatively identified as Philip II of Macedonia, middle of the 4th century BC. In the Ny Carlsberg Glyptotek, Copenhagen.

Macedonian recovery and expansion. Philip came to the throne suddenly and unexpectedly in 359, when Perdiccas was killed meeting an Illyrian invasion. The Illyrians prepared to close in; the Paeonians were raiding from the north, and two claimants to the throne were supported by foreign powers. In this crisis Philip showed a good sense of priorities by buying off his dangerous neighbours and, with a treaty, ceding Amphipolis to Athens. He used the time gained in military preparations. The army that later conquered Persia was developed all through his reign; but the decisive innovations in arms—the *sarissa*, a pike nearly half as long again as the spear of the Greeks—tactics, and training belong probably to this first year. In 358 he invaded Paeonia, and then he defeated the Illyrians decisively, in a battle that already suggests a master of war. The next year his marriage with Olympias, the Molossian princess of Epirus (the mother of Alexander the Great), helped to stabilize his western frontier. Now he ventured to antagonize Athens by recapturing Amphipolis, the strategic key securing

the eastern frontier and giving access into Thrace; and in 356 he took the west Thracian Crenides (renamed by him Philippi), a place newly founded to exploit new finds of silver and gold in Mount Pangaeum. These successes frightened his neighbours into forming a coalition against him, which was joined by Athens; but it achieved nothing.

The ten-year "war for Amphipolis" with Athens showed that the Athenians, with all their naval power, were quite unable to damage the continental and military power of Macedonia or even to save their own allies from Philip's attacks. Meanwhile he twice penetrated deeply into Thrace. And in the south a Thessaly divided against itself gave him an entry into Greece. These same ten years saw central Greece immersed in the Sacred War to liberate Delphi from its occupation by the Phocians, enabling Philip to intervene as the ally of Thebes and the Thessalian League of city states. His only great defeat in the field came in Thessaly in 353, owing (it seems) to overconfidence and failure of reconnaissance. The next year he retrieved it with a spectacular victory, which forced the Athenians to occupy Thermopylae and bar his path to the south. Characteristically Philip declined the trial of strength, prepared to wait for six years till he could gain Thermopylae by negotiation and without striking a blow. Meanwhile his Thessalian victory earned him election as president (*archōn*) of the Thessalian League (probably 352), a position unique for a foreigner in a Greek confederation and one that was to bind Thessaly to the kings of Macedonia for 150 years and more.

President of the Thessalian League

Philip's capture of Olynthus and annexation of Chalcidice in 348, enslaving the Olynthians and others of the Chalcidians, was disquieting to many. The Greeks themselves occasionally were brutal to small cities, but Olynthus was a large city. His enemies could affect a high moral tone and contempt for a barbarous Macedonian; but even his friends might have wondered whether he ought to be allowed into the heart of Greece with an army. Yet there were many ways in which he could serve them. Particularly, he could finish the Sacred War, which the Thessalians, Thebans, and others still could not finish for themselves. Athens could not prevent this now and had reason to fear that Philip's next campaign in Thrace (346) might challenge its own control of the sea route to southern Russia, its main source for imported corn. Significantly, however, it was Philip, and not Athens, who made the first overtures for peace, though all the military initiatives lay in his own hand. His plans for the future, in Greece and farther afield, included Athens as a willing ally not as a defeated enemy.

Even before the peace with Athens was ratified (346), the Athenian publicist Isocrates was inviting Philip to reconcile the four leading cities of Greece and to lead a united Greek alliance in a war of expansion against Persia. A step in this direction was Philip's intervention now to end the Sacred War, in recognition of which he was admitted to membership of the Delphic Amphictyony—an association of neighbouring states. The votes of the Thessalians and their clients gave him a control of its Council, which could be used on occasion for political and diplomatic ends. He lost no opportunity in the next years (346–343) of penetrating Greece without war, by winning and buying friends among the politicians of the smaller cities, intervening occasionally with subsidies or a force of mercenaries in their local disturbances. This policy made him some enemies, too, and it played into the hands of the great orator Demosthenes and others at Athens. Demosthenes saw Philip now as a bar to Athenian greatness and a threat to its freedom and existence; he talked tirelessly to warn the Athenians of the danger and to convince the Greeks in general that it was their danger too. Philip in these years conciliated Athens in small ways even under provocation, but he came to see that Demosthenes and the anti-Macedonians were beyond conciliation (343–342). Meanwhile, he reasserted his suzerainty over the neighbouring Illyrians, tightened his grip on Thessaly, and in 342 began the series of campaigns in Thrace that enabled him in two years to annex great parts of it as a province, and finally—to demonstrate his power against the Scythians settled on the southern banks of the Danube Delta. The events in Thrace caused two of his Greek allies, the cities of Perinthus and Byzantium, to review their position, and his coercion of them led to the two great sieges, which show the development of his artillery and allied arms, of which his son Alexander was to make greater use in Asia.

The declaration of war by Athens in 340 enabled him to raise the two sieges without undue loss of face, though he had failed to establish a threat to the Athenian corn route to southern Russia. Athens was to be intimidated now by invasion of its territory through central Greece, where the key position was held by Thebes, his ally hitherto, but of late a dissatisfied and recalcitrant one. His services to it in the Sacred War had been more than offset by his new position as its successful rival for leadership in and through the Amphictyony, and his moves toward hegemony in Greece could be seen in Thebes as encroachments. When Philip swept south with his army in November 339, he hoped to rush the Thebans into honouring their alliance and letting him through into Attica. But the Thebans listened to Demosthenes and to their own instinct of self-preservation. The Greek alliance became something formidable with the accession of Thebes, and Philip was forced, as a contemporary orator put it with only a mild exaggeration, "to stake his all on the issue of one short day." Chaeronea was a famous victory, gained by decisive blows of Philip's cavalry. But his real skill as a general can be seen, though dimly, in a manoeuvre of controlled retreat aimed at dislocating the advancing Greeks and creating gaps for the cavalry to strike. By winning this battle he had won the war.

Victory of Chaeronea

In the various peace treaties with the Greek states, Thebes had to admit a Macedonian garrison and its democratic constitution was replaced by a pro-Macedonian government; but Athens suffered neither invasion of its territory nor interference with its democracy and was not disarmed by dismantling of the walls or surrender of the navy. For Philip, Athens was the one Greek state from which he needed not neutrality or unwilling alliance but active cooperation. All past experience had shown that wars against Persia succeeded only when the Persians were denied the use of the Aegean, and for this the great Athenian navy was the first need. In Greece (outside Thessaly), Philip could have had no illusions about his own unpopularity, except among those of the well-to-do who were attracted by his court and his patronage; some cities also (especially the neighbours of Sparta) were glad to lean on Macedonia for support against an ancient enemy. Philip intended to involve all the Greeks with the Persian war. So Isocrates had advised him eight years before; but on the details of the ways and means he had no advice to offer. Philip himself organized the Greeks now to keep the peace with him and with each other and to support him in the Persian war overseas. In the constitutional details of his settlement of Greece he may well have had the help of Aristotle, free from his recent duties as tutor of the young Alexander.

Philip's so-called League of Corinth, established in 337, was an organization designed to preserve and perpetuate a general peace (*koinē eirēnē*), inaugurated when the delegates of all the states of Greece (except Sparta) and the islands swore to abide by it and to recognize Philip as president (*hēgemōn*) for this purpose. The general peace was a political innovation of the Greeks themselves, used several times in the past 50 years in attempts to stabilize affairs while promoting this or that hegemony. The peace had never lasted long because the leading Greek states had neither the power nor the mutual trust to create an effective organization for collective action against aggressors. Philip designed a council of representatives from all the states (*synedrion*), which was empowered to deliberate and decide on action to be taken in the event of the Peace being broken or threatened. After the decisions were made, their execution lay with Philip as *hēgemōn*. The states were under obligation to supply troops or ships to the *hēgemōn* on demand, by quotas corresponding to their voting power on the council. Though neither Philip

The League of Corinth

nor Macedonia had representatives on the council, it was the knowledge that the *hēgemōn* had the power of Macedonia in his hand that made this organization effective. As it happened, Corinth, where the inaugural meeting was held, was one of only three Greek cities with a Macedonian garrison, a fact the significance of which can have been lost on nobody.

Last years. Philip was wise, no doubt, to build on the foundation of the earlier practice of the Greeks themselves and also to refrain from organizing them in any permanent alliance that would have recalled too much the unpalatable experiences of the past. He was not, however, a Greek politician or even a Greek, but king of the Macedonians; and he cannot possibly have seen the settlement of Greece—as most modern historians have seen it—as the culmination of his life's work. For him it culminated nothing and was not even an end in itself but only a means. Chaeronea had brought the Greeks to order, and his plans required that they should stay in order now. The *synedrion* at Corinth heard his program for a Persian war and duly acclaimed it early in 337. Early next year an advance force of the Macedonian army crossed to Asia Minor. Philip would lead the grand army into Asia presently, and the Greeks would be with him.

This meteor fizzled out. The subtle, pliant, patient, calculating diplomatist, master of timing in politics and war, ended his life in a tale of irresponsible incompetence. The historian Theopompus, who saw Philip at close quarters, made much of his vices, his love of drink and debauchery, and his wild extravagance with money. Allowance made for this notably faultfinding and puritanical writer, Philip's character did contain some real ambiguities, extending into his domestic life. His "political marriages" were mostly opportune symbols of goodwill toward princes or groups worth conciliating, but his last marriage, in 338, to the Macedonian Cleopatra, led to a final break with Olympias, his queen, who left the country accompanied by the crown prince Alexander. Though Olympias was unpopular at court and though Cleopatra's connections were powerful and important, it was not "politic" to put the succession in jeopardy. Philip showed that he had never intended this result, by taking trouble to be reconciled with Alexander. The tradition that makes him infatuated with Cleopatra is probably right. If so, he misjudged fatally the amount of harm that could be done by marrying her.

Philip's assassination

With the preparations far advanced for the crossing into Asia, at the grand celebration of his daughter Cleopatra's marriage to Alexander of Epirus (brother of Olympias), Philip was assassinated by Pausanias, a young Macedonian noble with a bitter grievance against the young queen's uncle Attalus and against Philip for denying him justice. This was the official explanation, and Pausanias himself could add nothing to it; he was killed on the spot. Suspicion fell on Olympias and Alexander, those with most to gain from Philip's death, and many modern interpreters have followed it. Aristotle, however, clearly did not believe it. In his *Politics* a few years later he used this incident as an example of a monarch murdered for private and personal motives—which would have been a puerile indiscretion if either he or the world in general had ever taken the canard seriously.

Assessment. So ended, unworthily, the first of the great Macedonians. Everything known about him comes from Greek sources, which concentrate on his impact upon the Greeks and their history. Yet even more impressive, in view of Macedonia's troubled and undistinguished past, would be the full story of his unification and expansion of his own kingdom; his control of its regional princes, nobles, and gentry and their retainers, to form a great Macedonian people, symbolized by the finest army the world had seen; then, the continuing attrition by warfare and diplomacy, which in some 20 years reduced much of the Balkan peninsula to subservience. The apparently untidy record of his campaigns into Illyria or Thrace or of his interventions with diplomacy or arms (or both) in Thessaly, Euboea, the Peloponnese, which might suggest that repetition is a sign of incompetence, seem better interpreted as the work of a strategist operating always on several fronts, often preferring diplomacy to war, limited objectives to the grandiose, the smaller risks to the greatest; especially never forgetting that there is always another day. His decisive day at Chaeronea came, in a sense, because his true policy in Greece had failed, thanks partly to Demosthenes. But probably to take control of Greece without a Chaeronea was a real impossibility at this date (or indeed later).

Though Philip certainly wanted to be acceptable in Greece and did attract many important Greeks to his court, his philhellenism has been overrated: Olynthus and other Greek cities knew better. Though he cultivated the Athenians for reasons of high policy, there is no evidence that he ever in his life set foot in Athens, a remarkable piece of insouciance at every level. Pella, his capital, had long been a resort or refuge of great men of letters, and under Philip the connection with Plato's Academy was preserved, Theopompus was entertained, and Isocrates was invited; the leading actors of the Athenian stage appeared in Macedonia, too. Aristotle, whose father had been physician to Amyntas, Philip's father, spent three or four important years as Alexander's tutor. Philip presumably was at home with these people, but tradition says nothing of him as a man of letters himself or as an intellectual, though as an orator he could impress a party of Athenians which included Demosthenes and Aeschines and other professionals. His charm was great; he was by nature convivial, hospitable, and *bon viveur*. Undoubtedly he drank too much and too often, with the saving grace that he was known to listen to home truths even when drunk. As a commander in the field he was unwearying, and in action he fought like a lion; in the end he was really disfigured with old wounds. He was a general perhaps not of genius but of a very high order, with the tactical skill to coordinate the cavalry and infantry arms which were largely of his own creating. Making and training over the years a great army, he was paradoxically sparing and even cautious in using it.

Personal attributes

If he had survived to invade Asia, it would not have been to overthrow the Persian empire. He might have established a Macedonian empire in Asia, perhaps, but it would have been a Mediterranean empire in character. The Greeks would have benefitted by colonization, but the problem of Greek freedom would have remained, with the political domination of the higher culture by the lower. Philip was aware of the problem, and the League of Corinth, with its facade of freedom, was his answer. It did not deceive the Greeks or satisfy them; but no later Macedonian king could improve on it. Philip had made Macedonia, and now Macedonia and its kings made world history.

BIBLIOGRAPHY. PAUL CLOCHE, *Un Fondateur d'Empire: Philippe II, roi de Macédoine* (1955), a sound and readable survey; FRITZ GEYER, "Philippos (7)," in *Pauly-Wissowa Real-encyclopädie*, vol. 38, col. 2266–2303 (1938), an admirable work of reference; D.G. HOGARTH, *Philip and Alexander of Macedon* (1897), still very good reading; A.D. MOMIGLIANO, *Filippo il Macedone* (1934), an interesting study, most illuminating for its insights into the political life and thought of contemporary Greece; F.R. WUST, *Philipp II von Makedonien und Griechenland in den Jahren von 346 bis 338* (1938), an excellent analysis of Philip's later years.

(G.T.G.)

Philip II of Spain

Champion of the Roman Catholic Counter-Reformation, Philip II was king of Spain from 1556 to 1598. During his reign the Spanish Empire attained its greatest power and widest geographical extent but also suffered setbacks in its relations with the Protestant nations of northern Europe.

Philip was born at Valladolid on May 21, 1527, the son of the Holy Roman emperor Charles V and Isabella of Portugal. From time to time, the Emperor wrote Philip secret memoranda, impressing on him the high duties to which God had called him and warning him against trusting any of his advisers too much. Philip, a very dutiful son, took this advice to heart. From 1543 Charles con-



Philip II, oil painting by Titian (c. 1487/90–1576). In the Corsini Gallery, Rome.
Alinari

ferred on his son the regency of Spain whenever he himself was abroad. From 1548 until 1551, Philip travelled in Italy, Germany, and the Netherlands; but his great reserve and his inability to speak fluently any language except Castilian made him unpopular with the German and Flemish nobility.

Marriages

Philip contracted four marriages. The first was with his cousin Maria of Portugal in 1543. She died in 1545, giving birth to the ill-fated Don Carlos. In 1554, Philip married Mary I of England and became joint sovereign of England until Mary's death, without issue, in 1558. Philip's third marriage, with Elizabeth of Valois, daughter of Henry II of France, in 1559, was the result of the Peace of Cateau-Cambrésis (1559), which, for a generation, ended the open wars between Spain and France. Elizabeth bore Philip two daughters, Isabella Clara Eugenia (1566–1633) and Catherine Micaela (1567–97). Elizabeth died in 1568, and in 1570 Philip married another cousin, Anna of Austria, daughter of the emperor Maximilian II. She died in 1580, her only surviving son being the later Philip III.

King of Spain. Philip had received the duchy of Milan from Charles V in 1540 and the kingdoms of Naples and Sicily in 1554 on the occasion of his marriage to Mary of England. On October 25, 1555, Charles resigned the Netherlands in Philip's favour, and on January 16, 1556, the kingdoms of Spain and the Spanish overseas empire. Shortly afterward Philip also received the Franche-Comté. The Habsburg dominions in Germany and the imperial title went to his uncle Ferdinand I. At this time Philip was in the Netherlands. After the victory over the French at St. Quentin (1557), the sight of the battlefield gave him a permanent distaste for war, though he did not shrink from it when he judged it necessary.

His Castilian character

After his return to Spain from the Netherlands in 1559 Philip never again left the Iberian Peninsula. From Madrid he ruled his empire through his personal control of official appointments and all forms of patronage. Philip's subjects outside Castile, thus, never saw him, and they gradually turned not only against his ministers but against himself. This happened particularly in the Netherlands, in Granada, and in Aragon.

Method of government. By sheer hard work Philip tried to overcome the defects of this system. His methods have become famous. All work was done on paper, on the basis of *consultas* (that is, memoranda, reports, and advice presented him by his ministers). In Madrid, or in the gloomy magnificence of his monastic palace of El Escorial, which he built (1563–84) on the slopes of the Sierra de Guadarrama, the King worked alone in his small office, giving his decisions or, as often, deferring them. Nothing is known of his order of work, but all his contemporaries agreed that his methods dangerously, and sometimes fatally, slowed down a system of government already notorious for its dilatoriness. Painstaking and conscientious, Philip's craving for ever more information hid an inability to distinguish between the important and the trivial and a temperamental unwillingness to make decisions.

This was coupled with an almost pathological suspicion of even his most able and faithful servants. Margaret of Austria, duchess of Parma; the duke of Alba; Don John of Austria; Antonio Pérez; and finally, Alessandro Farnese—to name only the most distinguished—suffered disgrace. "His smile and his dagger were very close," wrote his official court historian, Cabrera de Córdoba. It was no exaggeration; for, in the case of Juan de Escobedo, the secretary of Don John of Austria, Philip even consented to murder. As a result, Philip's court became notorious for the bitterness of its faction fights. The atmosphere of the Spanish court did much to poison the whole Spanish system of government, and this played no small part in causing the rebellions of the Netherlanders (1568–1609), of the Moriscos of Granada (1568–70), and of the Aragonese (1591–92).

The black legend

Yet the "black legend" that, in Protestant countries, represented Philip II as a monster of bigotry, ambition, lust, and cruelty is certainly false. Philip's spare and elegant appearance is known from the famous portraits by Titian and by Anthonis Mor (Sir Anthony More). He was a lover of books and pictures, and Spain's literary Golden Age began in his reign. An affectionate father to his daughters, he lived an austere and dedicated life. "You may assure His Holiness," Philip wrote to his ambassador in Rome, in 1566, "that rather than suffer the least damage to religion and the service of God, I would lose all my states and an hundred lives, if I had them; for I do not propose nor desire to be the ruler of heretics." This remark may be regarded as the motto of his reign. To accomplish the task set him by God of preserving his subjects in the true Catholic religion, Philip felt in duty bound to use his royal powers, if need be, to the point of the most ruthless political tyranny, as he did in the Netherlands. Even the popes found it sometimes difficult to distinguish between Philip's views as to what was the service of God and what the service of the Spanish monarchy.

Foreign policy. For the first 20 years of his reign, Philip sought to preserve peace with his neighbours in western Europe. He was fighting a major naval war with the Ottoman Empire in the Mediterranean and, from 1568, he was faced with rebellion and war in the Netherlands. From the late 1570s, his policy gradually changed. The death (August 1578) without heirs of his nephew, King Sebastian of Portugal, opened up the prospect of Philip's succession to Portugal. He had to conquer (1580) by force what he regarded as his just, hereditary rights, but the rest of Europe was alarmed at this growth in Spanish power. Both England and France gave increasing support to the rebellious provinces of the Netherlands. Gradually, in the 1580s, Philip became convinced that the Catholic religion in western Europe, and his own authority in the Netherlands, could be saved only by open intervention against England and France. To this end he fitted out the Armada that, with the help of the Spanish Army in the Netherlands, was intended to conquer England (1588). He sent money and troops to support the League, the ultra-Catholic party in France, against Henry of Navarre and the Huguenots. He even claimed the throne of France for his daughter, Isabella Clara Eugenia, after the murder of Henry III in 1589. Again, even his Catholic allies found it difficult to distinguish between Philip's championship of the Catholic Church and the interests of Spain.

The Spanish Armada

All these plans failed. Henry of Navarre became a Catholic (1593) and Philip had to accept (Peace of Vervins, 1598) his succession as Henry IV of France. England and the northern Netherlands remained Protestant and unconquered. Yet Philip's reign as a whole was not a failure. He had defeated the great Ottoman offensive in the

Mediterranean at the Battle of Lepanto (1571). In the Iberian Peninsula he had completed the work of unification begun by the "Catholic Kings," Ferdinand and Isabella. Most important of all, in his own eyes, he had won great victories for the Catholic Church. If England, Scotland, and the northern Netherlands were lost, the southern Netherlands (the modern Belgium) had been preserved. In Spain and Italy he had prevented the spread of heresy, and his intervention in France was one of the factors that forced Henry IV to become a Catholic. When Philip II died of cancer at El Escorial on September 13, 1598, Spain was still at the height of its power; it took almost 50 years before it was clear that the Counter-Reformation would make no further major conquests.

BIBLIOGRAPHY. There are many biographies of Philip II but, as yet, no generally accepted or standard one. The fullest account of Philip II's life was written in the early 17th century by his official court historian, L. CABRERA DE CORDOBA, *Felipe Segundo, Rey de España*, 4 vol. (1876–77). Cabrera was very well informed, especially about life and personalities at Philip's court. His work has proved indispensable to all later biographers. The best full-scale modern political biography is R.B. MERRIMAN, *The Rise of the Spanish Empire in the Old World and the New*, vol. 4 (1934). The best example of the detailed work done in recent years on aspects of Philip's career is G. MATTINGLY, *The Defeat of the Spanish Armada* (1959), a book distinguished both by its impeccable scholarship and by the brilliance of its writing. An attempt to assess Philip II's aims and methods in the light of modern research is an article by H.G. KOENIGSBERGER, "The Statecraft of Philip II," in *European Studies Review*, 1:1–21 (1971).

(H.G.K.)

Philip the Good, Duke of Burgundy

The most important, most representative, and longest reigning of the four Valois dukes of Burgundy, who were among the leading European rulers between 1384 and 1477, Philip the Good has been acclaimed by historians both as the true founder of the Burgundian state in the Low Countries, from which modern Holland and Belgium draw their origins, and as the greatest patron of the arts north of the Alps. From 1419 until his death, in 1467, and especially in the years after 1435, the Duke of Burgundy was, with the possible exception of the king of France, the most powerful and successful ruler in the West.

Cliche des Musees Nationaux, Paris



Philip the Good, portrait attributed to Rogier van der Weyden (1399/1400–1464). In the Louvre, Paris.

Philip was born at Dijon, France, on July 31, 1396, the son of John the Fearless and Margaret of Bavaria. When he became duke of Burgundy at the age of 23, his first aim was to extricate himself as expeditiously as possible from the French affairs in which his father, Duke John, had been embroiled and that had led to his assassination in 1419. Holding the dauphin Charles (later Charles VII of France) answerable for his father's murder, Philip signed the Treaty of Troyes with King Henry V of England in 1420, a treaty in which the Queen of France, Isabella of Bavaria, conferred succession to the French crown on Henry and partitioned France among England, Burgundy, and her disinherited son, the dauphin Charles. Philip paid little attention to potential conquests in France and preferred to remain uncommitted there. He maintained his alliance with England, apart from a break in 1435–39 when he tried but failed to conquer Calais, but seldom gave England serious help against France. On the other hand, especially after 1435, when he acknowledged Charles as king of France and accepted his disavowal of the murder of John the Fearless, he did his best to be on reasonably good terms with the King of France. His real interests lay not in France but in the development of his own territories.

Conquests

Behind an impressive, if bizarre, facade of courtly splendour and chivalrous festivity, Duke Philip the Good was an aggressive opportunist who, especially in the first half of his ducal reign, concentrated on the task of attacking and swallowing up his smaller neighbours. Namur was purchased in 1420; Hainaut fell to Burgundian arms in 1427; the rich duchy of Brabant was taken over in 1430; and the combined counties of Holland and Zeeland were conquered in a long series of personally led and bitterly contested campaigns between 1424 and 1433. The crowning achievement of Philip's policy of territorial expansion was his conquest of the duchy of Luxembourg in 1443.

Patron of the arts

It was under Philip that the richness and extravagance of court life in the Middle Ages reached its apogee. Philip, whose personal tastes in clothes were relatively simple, loved to surround himself with all the pomp and pageantry that the age could command. In 1430 he founded a new order of chivalry, a Burgundian version of the British Order of the Garter, called the Toison d'Or, or Golden Fleece, membership of which was limited to 24 noblemen of proved valour and wide renown. Court was held at Brussels or Bruges, in Brabant and Flanders, respectively; or at Hesdin or Lille, in northeast France; or at some other centre. The best artists of the day were employed by Philip to paint his banners and pennons, to decorate his palaces and carriages, and to illuminate what was probably the finest collection of picture books ever put together. The artist Jan van Eyck accompanied a ducal embassy to Portugal to paint the King's daughter Isabella, so that Philip could see her likeness before committing himself to marrying her. Sculptors worked on tombs at Philip's command, exquisite tapestries were embroidered under his personal supervision, and a host of musicians, jewellers, goldsmiths, and other craftsmen and artists were employed at his court. The bawdy stories exchanged by Philip and his courtiers after dinner were collected into *Les Cent Nouvelles Nouvelles*, or "The Hundred New Short Stories." Some of the more elaborate banquets, notably the Feast of the Pheasant in 1454, at Lille, were open to the public, who could admire the endless array of model ships and towers, pies with people inside them, peacocks, swans, and eagles (mock or real), and other paraphernalia that accompanied the various dishes. Other entertainment was held from time to time in the form of tournaments or passages of arms, and Duke Philip's courtiers roamed about Europe issuing challenges and doing battle with their colleagues from other lands.

Duke Philip was tall, handsome, and bony in figure; his face was long and lean, with a high forehead, prominent nose, and bushy eyebrows. Excellent in health, he enjoyed hunting, tennis, archery, and jousting in his youth, but he turned in his later years to making clogs, repairing broken glasses, and soldering broken knives. His many bastard children caused the Bishop of Tournai (himself a bastard) to criticize him for what the ecclesiastic called "the weakness of the flesh." Some were brought up at court; others went into the church. His mistresses were kept out of affairs of state, and it was mere geographical convenience and economy that caused him to maintain several at once in the different towns where he held court. Self-assured and flamboyant almost to the end, he died, possibly of pneumonia, at Bruges on June 15, 1467.

BIBLIOGRAPHY. O. CARTELLIERI, *Am Hofe der Herzöge von Burgund* (1926; Eng. trans., *The Court of Burgundy*, 1929), the best study available in English of the Burgundian court; P. BONENFANT, *Philippe le Bon*, 3rd ed. (1955); R. VAUGHAN, *Philip the Good: The Apogee of Burgundy* (1970), lists all relevant literature.

(R.Va.)

Philip the Magnanimous, Landgrave of Hesse

Philip the Magnanimous was the father of political Protestantism and champion of the independence of the German princes against the Catholic emperor Charles V. He contributed decisively to the establishment of conditions under which German Protestants could freely develop their churches in each state and, through the religious Peace of Augsburg in 1555, to achieving for Lutherans equality with Catholics.

By courtesy of the Wartburg Sammlungen, Eisenach; photograph, Deutsche Fotothek Dresden/Mobius



Philip the Magnanimous, as Landgrave of Hesse, portrait by Hans Krell, 1525. In the Wartburg Sammlungen, Eisenach, Germany.

Early years. Philip was born November 13, 1504, in Marburg. His father, Landgrave William II, was a cultivated, austere man and an experienced soldier. He died when Philip was barely five years old. Philip's mother, Duchess Anna of Mecklenburg, was a passionate, energetic, and ambitious woman; her relations with her son, however, were cool and were impaired by her second marriage in 1519 and by Philip's conversion to the doctrines of Martin Luther. As a victim of political forces he spent a gloomy youth, attaching himself increasingly to his older sister Elizabeth, later duchess of Saxony; throughout his life no one was closer to him than she.

In March 1518 the emperor Maximilian I declared Philip to be of age. While assuming the government in name, he retained his parents' capable counsellors, mostly lawyers by training, who imbued the young man with his parents' single-minded devotion to the welfare of the state. Philip eventually developed into a self-reliant politician. From his youth, he characteristically thought in terms of the territorial state, an attitude that was to determine his future relations with the Habsburg emperors. He ruled his territory in the spirit of an early enlightened despotism.

Formative influences

The landgraviate of Hesse had suffered greatly during his mother's regency, but Philip succeeded in bringing order to the confused administration of the state and, through skillful management of alliances, in freeing Hesse from its isolation in foreign affairs. In 1523 Philip, fighting alongside the powerful electors of the Palatinate and of Trier, defeated the rebellious imperial knight Franz von Sickingen. This victory also crushed the remaining opposition of the Hessian nobility and undoubtedly strengthened the young man's self-confidence. Two years later, in fact, he scored his first personal triumph when he suppressed the peasant revolt in the neighbouring imperial abbeys of Fulda and Hersfeld. He then advanced into Thuringia and, allied with the Elector of Saxony and the Dukes of Brunswick-Wolfenbüttel and Saxony, defeated Thomas Müntzer, a popular preacher from Mühlhausen/Thüringen, at a battle near Frankenhausen in May 1525. By this victory Philip saved the central German principalities from destruction.

Conversion to Protestantism. During these years, while the first storm of the Reformation was sweeping through Germany, Luther's teachings gained a firm hold in many Hessian localities and at the Landgrave's court. But it was not until the summer of 1524, after making a detailed study of the Bible, the writings of the Church Fathers, and the history of the church, that Philip himself joined Luther. At the same time, orthodox Catholic princes of south Germany were uniting to make common cause against the ecclesiastical innovations. After the defeat of the peasants this alliance was aimed to bring down the middle and north German princes as well, in order "to eradicate the damned Lutheran sect, the source of this revolt." Philip was deeply convinced that the religious question was at the same time a political one, and he was the first to recognize that freedom for the new faith would be secured only if Protestant sovereigns and towns united for defense. In May 1526 he won over the elector John of Saxony for a defensive alliance, which other north and east German princes then joined. Henceforward, he strove to unite the Protestant estates into a powerful alliance, which would render them unassailable and thus allow them to build their state churches.

Decree of the Diet of Speyer

Only when the decree of August 1526 of the imperial Diet of Speyer seemed to provide a legal basis for it and when a Hessian "synod" (partly church council and partly provincial diet) at Homberg had publicly discussed the religious question did Philip carry through the Reformation in his state. The outcome of the Homberg deliberations was the *Reformatio ecclesiarum Hassiae*, unique for its democratic–presbyterian church constitution and the ecclesiastical discipline of the congregations. On Luther's advice this reform, conceived by the former Franciscan Franz Lambert, was not carried out; instead Hesse became Protestant along the lines of the model provided by the Electorate of Saxony. Philip, however, continued to favour the teachings of the south German and Swiss reformers and consequently resolved to mediate between the Lutherans and the Zwinglians, adherents of the Swiss reformer Huldrych Zwingli. Within one year, 1527, Philip secularized all the monasteries. The new state church organization was now methodically built up. In order to ensure a new generation of clergy and officials, the Landgrave took the educational system in hand and founded the first Protestant university in 1527 at Marburg. On former monastic and church estates, he set up four hospitals for the insane, the first "psychiatric" hospitals known to medical history. Church and school were subordinated to the "common good." There arose the Protestant authoritarian state, which considered itself a taskmaster responsible for the religious and moral life of its subjects and dictated their beliefs. Like all his fellow princes, Philip was intolerant, but, in contrast to most, he respected the individual's freedom of conscience by permitting dissenters to emigrate. He soon showed this attitude in his dealings with the Anabaptists; his purpose was not to punish them as the imperial laws dictated but to convert them by clemency and instruction in doctrine; in this work Martin Bucer, the reformer from Strasbourg, was his helper. They introduced into Hesse in 1539 the rite of confirmation, which became a model for Protestant churches.

Leadership of the Protestant states. His agile and fertile mind, infectious energy, and fearlessness rapidly made Philip the leader of the Protestant estates. He continued to strive for a great Protestant alliance, for he clearly recognized that the situation of the Protestants would deteriorate if the Catholic emperor Charles V were to triumph in the struggle for European predominance. The way to change and reform began in 1529, when the second imperial Diet of Speyer repealed the first imperial Diet's decree of 1526. A group of princes and

towns led by Philip protested (hence Protestants) against the repeal on the epoch-making grounds that in questions of faith each imperial estate would have to justify itself before God alone.

Moreover, the Landgrave saw with growing apprehension that doctrinal differences between Protestants endangered the development of an all-embracing Protestant defensive alliance. His attempt at the Colloquy of Marburg, in October 1529, to bring about a theological settlement in personal discussion with the reformers, headed by Luther and Zwingli, essentially failed over the question of the communion. Thus, the ambitious plans he shared with Zwingli for a European alliance against the Catholic house of Habsburg remained only a dream.

Schmalkaldic League

Despite everything, Philip managed in 1531 to unite six princes and ten towns in the Schmalkaldic League, which was to serve as a defensive alliance. Although it had serious organizational shortcomings, the league gradually became the centre of Protestant politics. At the same time it became a rallying point for the enemies of the House of Habsburg, as well as for those Catholic princes who were fearful for their independence. The league, of which the Landgrave was the driving spirit, became, moreover, a factor in European politics. Philip reached the peak of his career in 1534, when he executed a campaign to restore Duke Ulrich of Württemberg, who had been driven from his state by the Habsburgs. As a result of his success, Württemberg was opened up to the Reformation, and Austria's power in south Germany was broken.

Decline. It was Philip's tragedy that he destroyed his own leading position a few years later by his extremely provocative marital transactions. In addition to his marriage to the duchess Christina of Saxony in 1523, he contracted a second marriage with Margarete von der Saale, a maid of honour of his sister Elizabeth. As a bigamist he fell under the judgment of the Emperor, with whom he now tried to come to terms; but while Philip and the Schmalkaldic League remained inactive and the other Protestant princes indulged in petty disputes, Charles V prepared to settle the religious problem once and for all by force of arms. In the summer of 1546 he attacked. The league's unwieldy organization, long deplored by Philip, now took its toll. Mistakes of leadership, lack of finance, and ultimately the attack by the Protestant duke Maurice of Saxony on the territory of his cousin the elector John Frederick of Saxony hastened the collapse. After the capture of the Elector,

Prisoner of the Emperor

Philip submitted himself to the mercy of the Emperor in June 1547. Deceived by the Emperor, he, too, was led away a prisoner. His long imprisonment in the Netherlands deeply affected his mental powers of resistance. In order to gain his freedom, he accepted the so-called Augsburg Interim, by which the Emperor attempted to restore the unity of the Catholic faith without interference from the princes. Philip, however, failed in his bid for freedom, for the Hessian population resisted conversion to Catholicism. Only after his son-in-law, the elector Maurice of Saxony, and Philip's eldest son, William, in alliance with other German princes and Henry II of France, unexpectedly rose against the Emperor in March 1552 were he and John Frederick released.

Aged and ailing, but also wiser, the Landgrave returned to his homeland. After the victory over the Emperor, the adherents of the Confession of Augsburg—the official Lutheran doctrine—succeeded in gaining a position of legal equality with the Catholics in the empire, in accordance with the Peace of Augsburg of 1555. In this respect Philip, who was far ahead of his time, expressed genuine tolerance for all Christian denominations. He continued to pursue his old plans for a Protestant union and strove on behalf of his embattled coreligionists in France and the Netherlands, but in his later years he cautiously remained in the background of the political stage. He devoted all his strength to the rebuilding of his state, which had been ravaged by war and by its occupation by foreign troops. By the time of his death in Cassel on March 31, 1567, his contemporaries were already referring to the warm-hearted and generous sovereign as the Magnanimous.

BIBLIOGRAPHY. There is as yet no modern biography of Philip. Valuable individual research is contained in the publications issued in 1904 to honour him: *Festschrift zum Gedächtnis Philipp des Grossmütigen*, published by the Verein für hessische Geschichte und Landeskunde (1904); *Philipp der Grossmütige*, published by the Historischen Verein für das Grossherzogtum Hessen (1904); and C.A. VON DRACH and G. KONNECKE, *Die Bildnisse Philipps des Grossmütigen* (1905). From the abundance of previously published works the following in particular should be mentioned: K. DULFER, *Die Packschen Händel: Darstellung und Quellen* (1958); W. HEINEMEYER, "Die Bildungspolitik des Landgrafen Philipp von Hessen," in *Hessisches Jahrbuch für Landesgeschichte 21* (1971); and "Landgraf Philipps des Grossmütigen Weg in die Politik," *ibid. 5* (1955); F.H. LITTELL, *Landgraf Philipp und die Toleranz* (1957); W.W. ROCKWELL, *Die Doppelehe des Landgrafen Philipp von Hessen* (1904); W. SOHM, *Territorium und Reformation in der hessischen Geschichte 1526–1555*, 2nd ed. by G. FRANZ (1957); and A.O. HANCOCK, "Philipp of Hesse's View of the Relationship of Prince and Church," *Church History*, 35:157–169 (1966). A bibliography of literature on Philip is given in K.E. DEMANDT, *Schrifttum zur Geschichte und geschichtlichen Landeskunde von Hessen*, vol. 1, pp. 203–212 (1965). General information is found in *The New Cambridge Modern History*, vol. 2, ch. 3 and 6 (1958). Philip's political correspondence is preserved in the Hessisches Staatsarchiv Marburg and has been calendared in F. KUCH and W. HEINEMEYER (eds.), *Politisches Archiv des Landgrafen Philipp*, 4 vol. (1904–59). All known sources for the history of the Reformation in Hesse are in W. SOHM and G. FRANZ (eds.), *Urkundliche Quellen zur hessischen Reformationsgeschichte*, 4 vol. (1915–55; 2nd ed. of vol. 1, 1957).

(W.He.)

Philippines

The Republic of the Philippines is an archipelago consisting of about 7,100 islands and islets, lying about 500 miles off the southeast coast of Asia. The total land area of the Philippines is 115,830 square miles (300,000 square kilometres); it had a population of about 38,000,000 in 1971. It is bounded on the east by the Philippine Sea (Pacific Ocean), on the south by the Celebes Sea, and on the west and north by the South China Sea. The Philippines takes its name from Philip II, king of Spain at the time of the Spanish colonization of the islands in the 16th century. Both the present capital, Quezon City, and the former one, Manila, still housing many governmental offices, are situated on Luzon, the largest island.

The archipelago spreads out in the form of a triangle, with the archipelago of Sulu and the island of Mindanao forming its southern base and the islands of Batanes, in the north, its apex. The islands stretch out for about 1,100 miles (1,800 km) from north to south and for about 650 miles from east to west at the archipelago's widest extent. Only 2,773 of the islands and islets have names, and only 462 have areas of one square mile or more; many are no more than reefs or small atolls. The large islands fall into three groups: (1) the Luzon group, consisting of Luzon (with a land area of 40,420 square miles) and Mindoro (3,759 square miles) in the north; (2) the Visayan group in the centre, consisting of Bohol (1,492 square miles), Cebu (1,703 square miles), Leyte (2,786 square miles), Masbate (1,262 square miles), Negros (4,906 square miles), Palawan (4,550 square miles), Panay (4,446 square miles), and Samar (5,050 square miles); and (3) Mindanao (36,537 square miles) in the south.

Because it was under Spanish rule for 333 years and under United States' tutelage for a further 50 years, the Philippines has many cultural affinities with the West. It is, for example, the only Roman Catholic country in Southeast Asia. Its peoples, however, most of whom were originally of Malay origin, are Asian both in consciousness and in aspiration. In many ways Filipino society is composed of paradoxes. While extremes of wealth and poverty exist, and while agricultural land is often owned by wealthy landowners, the Philippines is vigorously democratic. A country of rich resources, it is still far from developing its full potentialities. Economically it is primarily agricultural—its major exports being coconuts and coconut products, sugar, and abaca (Manila hemp). Educationally, however, it is among the most advanced

of Asian countries, maintaining more than 40 universities and 374 private colleges, and having a literacy rate of about 72 percent (for coverage of associated physical features, see CHINA SEA; PACIFIC OCEAN; for historical aspects, see PHILIPPINES, HISTORY OF THE).

THE NATURAL AND HUMAN LANDSCAPE

The physical environment. *Relief and drainage.* The irregular configuration of the archipelago, the coastline of more than 21,500 miles, the great extent of mountainous country with the mountains lying close to the sea in many places, the narrow and interrupted coastal plains, the generally northward trend of the river systems, and the lakes comprise the outstanding physical features of the Philippines. The islands have volcanic, coral, and all principal rock formations. The mountain ranges for the most part run in the same general direction as the islands themselves, approximately north to south. The Cordillera Central (central mountain chain) of Luzon, running north to the Pacific Ocean from the northern boundary of the central plain, is the most prominent chain. It consists of two, and in places three, parallel ranges, each of which averages 5,900 feet (1,800 metres) in height. The narrow Ilocos, or Malayan Range, lying close along the narrow west coast, rises in places to more than 5,000 feet and is seldom less than 3,500 feet in height; it is largely volcanic. This range and the Cordillera Central merge in Nueva Vizcaya Province to form the Caraballo Mountains. To the north of the latter, and between the two ranges, is the fertile Cagayan Valley, about 50 miles wide. In the southwestern part of northern Luzon are the rugged Zambales Mountains, consisting of more or less isolated old volcanic stocks (bodies of rock originally formed deep down under heat and pressure).

Mountain ranges

The central plain of Luzon, about 150 miles by 50 miles, is only about 100 feet above sea level except near its centre. The greater part of southern Luzon is occupied by isolated volcanoes and irregular masses of hills and mountains. The highest peak is Mayon (7,943 feet) near Legazpi in Albay Province.

Through the island of Palawan, about 25 miles wide, there extends a range with an average height of from 4,000 to 5,000 feet. Each of the Visayan Islands, except Samar and Bohol, is traversed longitudinally by a single range with occasional spurs. Several peaks in Panay reach an altitude of 6,000 feet or more. Mount Canlaon, on Negros is a volcano 8,087 feet high; other peaks reach an altitude of more than 6,000 feet.

There are several important ranges on Mindanao, the Diuata Range along the eastern coast being the most prominent. West of this lies the fertile Agusan Valley, 12 to 37 miles wide. Farther west and southwest is the valley of the Rio Grande de Mindanao, the largest river system on the island, and between its lower course and the southern coast lies a range trending northwest and southeast. On the southern border of the basin of Rio Grande de Mindanao River is Mount Apo, 9,691 feet high, the highest peak in the Philippines. Around Lake Lanao are a number of volcanic peaks. A low cordillera extends into Zamboanga Province.

The plains lying amid the mountains—for example the central plain of Luzon, the Cagayan plain of northern Luzon, the central plain of Panay, and the Agusan and Cotabato valleys on Mindanao—have the densest population of the islands, except in Cebu, where the people live mostly on the coastal plain because the interior is high and rugged. The most important rivers are the Cagayan, the Agno, the Pampanga, the Pasig, and the Bicol on Luzon and the Rio Grande de Mindanao and the Agusan-

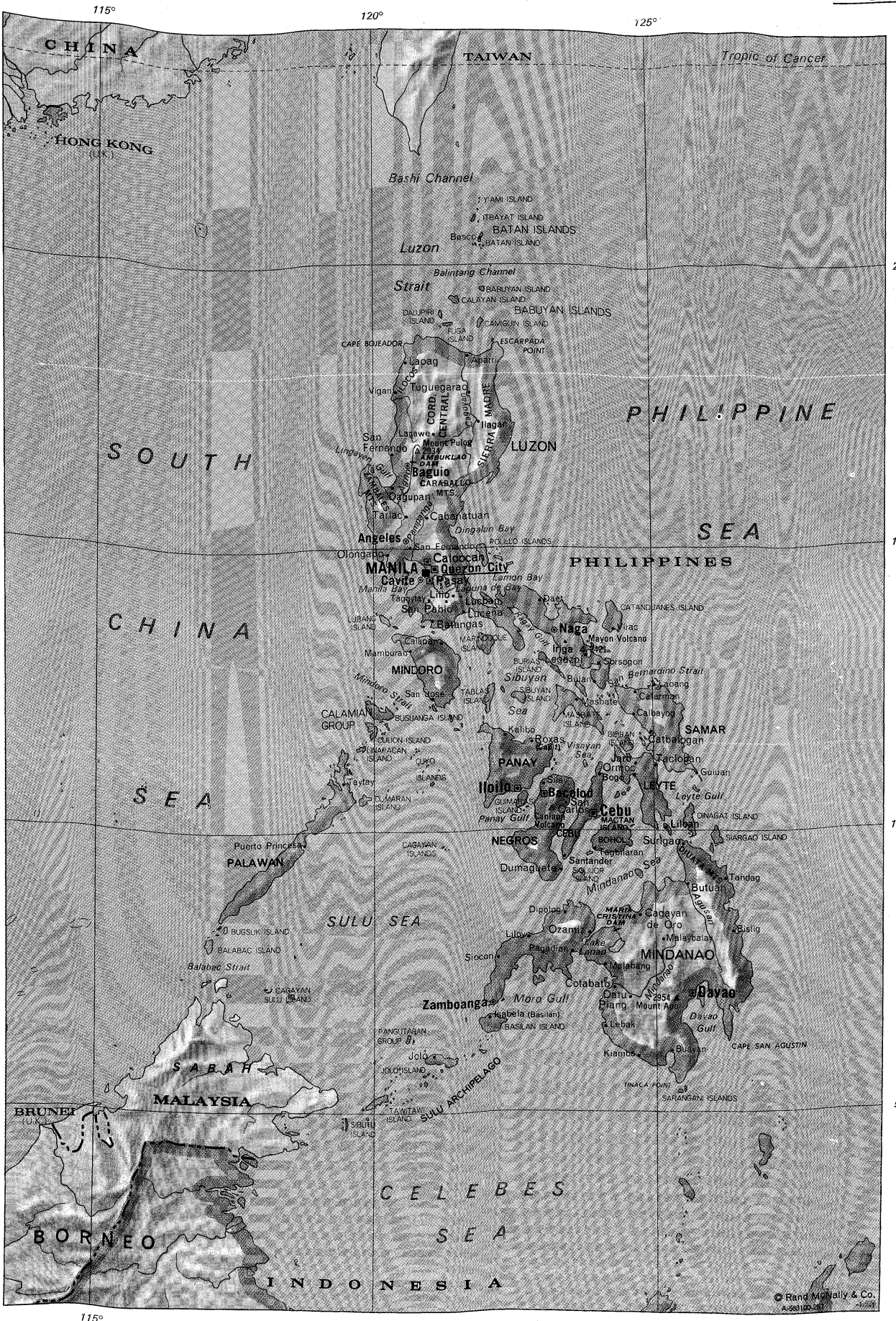
MAP INDEX

Cities and towns

Angeles 15·09n 120·35e
Aparri 18·22n 121·38e
Bacolod 10·40n 122·57e
Baguio 16·25n 120·36e
Basco 20·27n 121·58e
Basilan, see Isabela
Batangas 13·45n 121·03e
Bislig 8·13n 126·19e
Bogo 11·03n 124·00e
Buayan 6·06n 125·15e
Bulan 12·40n 123·52e
Butuan 8·57n 125·33e
Cabanatuan 15·29n 120·58e
Cagayan de Oro 8·29n 124·39e
Calapan 13·25n 121·11e
Calbayog 12·04n 124·36e
Caloocan 14·39n 120·59e
Capiz, see Roxas
Catarman 12·30n 124·38e
Catbalogan 11·47n 124·53e
Cavite 14·29n 120·55e
Cebu 10·18n 123·54e
Cotabato 7·14n 124·15e
Daet 14·07n 122·57e
Dagupan 16·03n 120·20e
Datu Piang 7·02n 124·29e
Davao 7·04n 125·36e
Dipolog 8·36n 123·20e
Dumaguete 9·18n 123·18e
Guiuan 11·02n 125·44e
Ilagan 17·09n 121·53e
Iloilo 10·43n 122·33e
Iriga 13·25n 123·25e
Isabela (Basilan) 6·42n 121·58e
Jaro 11·11n 124·46e
Jolo 6·03n 121·00e
Kalibo 11·43n 122·22e
Kiamba 5·59n 124·37e
Lagawe 16·45n 121·10e
Laoag 18·12n 120·36e
Laoang 12·34n 125·00e
Lebak 6·37n 124·03e
Legazpi 13·08n 123·44e
Lilio 14·08n 121·26e
Liloan 10·09n 125·07e
Liloy 8·08n 122·40e
Lucban 14·07n 121·33e
Lucena 13·56n 121·37e
Malabang 7·36n 124·04e
Malaybalay 8·09n 125·05e
Mamburao 13·13n 120·35e
Manila 14·35n 120·59e
Masbate 12·22n 123·37e
Naga 13·38n 123·11e
Olongapo 14·49n 120·17e
Ormoc 11·00n 124·37e
Ozamiz 8·09n 123·51e
Pagadian 7·50n 123·27e
Pasay 14·33n 121·00e
Puerto Princesa 9·44n 118·44e
Quezon City 14·38n 121·00e
Roxas (Capiz) 11·35n 122·45e
San Carlos 10·29n 123·25e
San Fernando 15·02n 120·41e
San Fernando 16·37n 120·18e
San Jose 12·27n 121·03e
San Pablo 14·04n 121·19e
Santanger 9·25n 123·20e
Silay 10·48n 122·58e
Siocon 7·42n 122·09e
Sorsogon 12·58n 124·00e
Surigao 9·48n 125·30e
Tacloban 11·15n 125·00e
Tagaytay 14·06n 120·56e
Tagbilaran 9·39n 123·51e
Tandag 9·05n 126·12e
Tarlac 15·29n 120·35e
Taytay 10·50n 119·31e
Tuguegarao 17·37n 121·44e
Vigan 17·35n 120·23e
Virac 13·35n 124·14e
Zamboanga 6·54n 122·05e

Physical features and points of interest

Agno, *river* 16·02n 120·13e
Agusan, *river* 9·00n 125·30e
Ambuklao Dam 16·28n 120·45e
Apo, Mount, *mountain* 7·00n 125·16e
Babuyan Island 19·32n 121·57e
Babuyan Islands 19·10n 121·40e
Balabac Island 7·55n 117·00e
Balabac Strait 7·35n 117·00e
Balintang Channel 19·45n 122·00e
Bashi Channel 22·00n 121·00e
Basilan Island 6·35n 122·00e
Batan Island 20·25n 121·57e
Batan Islands 20·35n 121·55e
Bay, Laguna de, *lake* 14·20n 121·11e
Biliran Island 11·35n 124·30e
Bohol, *island* 9·50n 124·10e
Bojeador, Cape 18·30n 120·34e
Bugsuk Island 8·15n 117·18e
Burias Island 12·55n 123·05e
Busuanga Island 12·05n 120·05e
Cagayan Islands 9·40n 121·25e
Cagayan Sulu Island 7·00n 118·28e
Cagoyan, *river* 18·25n 121·40e
Calamian Group, *islands* 12·00n 120·10e
Calayan Island 19·20n 121·25e
Camiguin Island 18·54n 121·55e
Canlaon Volcano 10·24n 123·08e
Caraballo Mountains 16·08n 121·05e
Catanduanes Island 13·45n 124·15e
Cebu, *island* 10·15n 123·40e
Celebes Sea 3·00n 122·00e
Central, Cordillera, *mountains* 17·30n 121·00e
Culion Island 11·50n 120·00e
Cuyo Islands 10·53n 121·00e
Dalupiri Island 19·05n 121·13e
Davao Gulf 6·40n 125·45e
Dinagat Island 10·10n 125·35e
Dingalan Bay 15·15n 121·25e
Diuata Mountains 9·10n 125·47e
Dumaran Island 10·33n 119·50e
Escarpada Point 18·31n 122·13e
Fuga Island 18·53n 121·22e
Guimaras Island 10·35n 122·37e
Ilocos, *mountains* 18·00n 120·45e
Itbayat Island 20·45n 121·50e
Jolo Island 6·00n 121·10e
Lamon Bay 14·33n 121·50e
Lanao, Lake 7·52n 124·15e
Leyte, *island* 10·55n 124·50e
Leyte Gulf 10·40n 125·25e
Linapacan Island 11·18n 119·47e
Lingayen Gulf 16·10n 120·15e
Lubang Island 13·45n 120·10e
Luzon, *island* 16·00n 121·00e
Luzon Strait 20·30n 121·00e
Mactan Island 10·18n 123·58e
Madre, Sierra, *mountains* 17·15n 122·00e
Manila Bay 14·35n 120·45e
Maria Cristina Dam 8·11n 124·09e
Marinduque Island 13·25n 122·00e
Masbate Island 12·10n 123·30e
Mayon Volcano 13·15n 123·42e
Mindanao, *island* 7·30n 125·00e
Mindanao, *river* 7·07n 124·24e
Mindanao Sea 9·10n 124·20e
Mindoro, *island* 13·00n 121·10e
Mindoro Strait 12·30n 120·30e
Moro Gulf 6·30n 123·10e
Negros, *island* 10·00n 123·00e
Palawan, *island* 9·30n 118·30e
Pampanga, *river* 14·46n 120·40e
Panay, *island* 11·10n 122·30e
Panay Gulf 10·15n 122·25e
Pangutaran Group, *islands* 6·15n 120·35e
Philippine Sea 17·00n 127·00e
Polillo Islands 14·50n 122·10e
Pulog, Mount, *mountain* 16·35n 120·54e
Ragay Gulf 13·40n 122·40e
Samar, *island* 12·00n 125·00e
San Agustin, Cape 6·16n 126·11e
San Bernardino Strait 12·37n 124·12e
Sarangani Islands 5·25n 125·25e
Sibutu Island 4·50n 119·30e
Sibuyan Island 12·25n 122·35e
Sibuyan Sea 12·50n 122·20e
Siargao Island 9·52n 126·03e
Siquijor Island 9·10n 123·33e
South China Sea 13·00n 117·00e
Sulu Archipelago 5·40n 121·13e
Sulu Sea 7·30n 120·00e
Tablas Island 12·25n 122·02e
Tawitawi Island 5·12n 120·00e
Tinaca Point 5·34n 125·20e
Visayan Sea 11·30n 123·35e
Y'Ami Island 21·07n 121·57e
Zambales Mountains 15·40n 120·15e



PHILIPPINES

Size of symbol indicates relative size of town • ◉ ▣ ▣ ■

Elevations in metres

0 100 200 400 600 km

0 100 200 400 mi

on Mindanao. The Pasig, flowing through the city of Manila, is commercially important. The largest lake in the archipelago is Laguna de Bay, located south of Manila.

Volcanic areas

Volcanoes are the most conspicuous feature of the landscape, but there is relatively little volcanic activity. There are altogether about 50 volcanoes, of which 10 are known to be active. All gradations of volcanoes can be seen, from the almost perfect cone of Mayon, which has been compared to Fujiyama in Japan, to old, worn-down volcanic stocks, the present forms of which give little indication of their origin. The several distinct volcanic areas are in southwestern Luzon, southeastern Luzon, the mountain region in Camarines Norte, and in Negros, Mindanao, Jolo, and elsewhere. Seismic disturbances are common.

Climate. The climate is tropical and there are two marked seasons—the dry and the wet. The dry season generally begins in March and ends in June, and the wet season covers the rest of the year. During the wet season, rainfall is heavy in all parts of the archipelago except in Cebu and Zamboanga. Rain is heaviest along the eastern shores facing the Pacific Ocean. From June to December typhoons frequently strike the archipelago; in some years the number of typhoons reaches 200 or more. When accompanied by floods or high winds these typhoons, which are heaviest in Samar, Leyte, eastern Quezon Province, eastern Mindanao provinces, and the Batan Islands, cause great loss of life and property. Mindanao, except for the eastern and northeastern part, is generally free from typhoons.

November through February constitutes the most agreeable season; the air is cool and invigorating at night, and the days are pleasant and sunny. The cities of Baguio and Tagaytay on Luzon, the town of Malaybalay on Mindanao, and some other places, such as Lilio and Majayjay in Laguna, and Lucban in Quezon, experience a pleasant climate throughout the year; at times the temperature dips as low as 60° F (16° C). During the hot or dry season in most places, and especially in the cities of Cebu, Davao, and Manila, the temperature sometimes rises as high as 100° F (38° C).

Temperature ranges

Vegetation. Almost 40 percent of the land area of the islands is covered with forest. Many of the mountains are covered with dense forest. On northern Luzon the principal mountain tree is pine. In many places where the forests have been burned away their place has been taken by *cogon*, a tall coarse grass, or similar grasses. There are about 10,000 known species of flowering plants and ferns. Much of the vegetation resembles that found in Malaysia; and the vegetation of the coastal areas, including the mangrove swamps, is practically identical with that of similar areas throughout the Malay Archipelago. Himalayan elements occur in the mountains of northern Luzon, while a few Australian types are found at various altitudes. A majority of the plant species—about 60 percent—are indigenous; these include the dominant forest species. Members of the lauan (Philippine mahogany) family of trees often comprise a high percentage of the forest. There are many other valuable hardwoods as well as many softwoods. About 800 species of orchids occur, some of them very rare.

Animal life. There are about 220 species of mammals, among which are carabaos (water buffalo), goats, horses, hogs, cats, dogs, monkeys, squirrels, lemurs, mice, pangolin anteaters, chevrotains (mouse deer), mongooses, civet cats, and red and brown deer. The tamarau, a very small species of carabao, is found only on the island of Mindoro. There are at least 56 species of bats, many of which are peculiar to the Philippines; large fruit bats live in colonies. Early records show that there were once elephants on the islands.

There are about 1,000 species and subspecies of known birds, some of which are migrants. The bird life includes jungle fowl, pigeons, peacocks, pheasants, doves, parrots, hornbills, kingfishers, sunbirds, tailorbirds, weaverbirds, herons, and quails. Many species of birds are unique to the island of Palawan. The rare monkey-eating eagle is found only on Mindanao, Luzon, Mindoro, and Palawan.

The landscape under human settlement. *Rural settlement.* Before the arrival of the Spanish in the 16th century, the people lived in small independent villages called *barangay*, each of which was ruled by a petty king called a *datu*. Settlement in rural areas today is focussed upon *poblaciones*, or small towns, originally founded by the Spanish that have agricultural populations numbering as many as 10,000 people. Roads or trails were built from the *población* in from four to six directions, like the spokes of a wheel. Along the roadsides, barrios or villages have grown up; these barrios are subdivided into smaller units called *sitios*. More people live in the barrios than in the *población*. A large barrio may be composed of as many as 1,000 households and have a population of 3,000 inhabitants or more. The cultivation of rice or corn, and fishing, provide the basic subsistence.

Towns and cities. While the Philippines has 1,369 municipalities, 30 municipal districts, and 58 chartered cities, almost all municipalities and all municipal districts are in fact rural communities.

Manila, Cebu, Jaro, Vigan, and Nueva Caceres (Naga) were granted city charters by the Spanish. More cities were founded under United States administration, and still more have been created since independence in 1946. Quezon City was built to be the capital of the Philippines.

Recently founded cities

In most of these cities, life has become increasingly complex as a result of the growth of population, business, and industry, and many of the problems of contemporary urban life have developed.

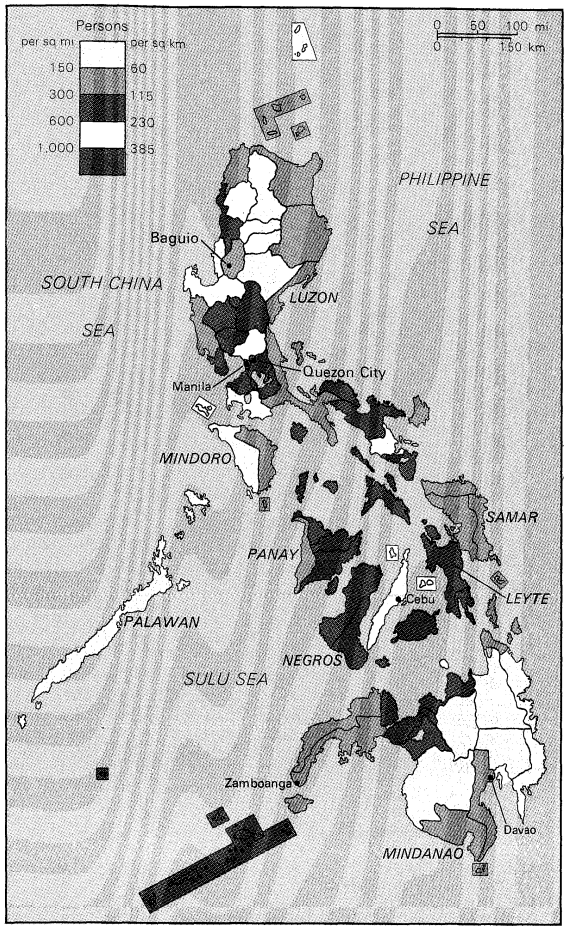
PEOPLE AND POPULATION

Groups historically associated with the contemporary country. *Ethnic groups.* The people of the Philippines are called Filipinos. Their ancestors, who were of the Malay or brown race, came to the Philippines from the southeastern Asia mainland as well as from what is now Indonesia. From the 10th century, contacts with China resulted in a group of mixed Filipino-Chinese descent who number about 10 percent of the entire population. An estimated 120,000 Chinese nationals also live in the country. Mixed Spanish-Filipino and Filipino-American types among the population may be distinguished by their fairer complexion, taller stature, and aquiline noses. While some emigrants have come from the Indian subcontinent, Indian-Filipino interracial mixture has been negligible, the Indian community numbering about 1,500. There are about 10,000 resident United States nationals (excluding military personnel). Spaniards number about 1,000. The aboriginal inhabitants of the islands were the Pygmies, or Negritos, also called Aetas or Balugas. From four foot five to four foot eight inches in height, and black-complexioned, they number about 30,000—one-twelfth of 1 percent of the total population.

Linguistic groups. Estimates of the total number of languages and dialects spoken in the Philippines differ; latest scholarly studies suggest that there are about 75 of them. Traditionally, seven major linguistic groups are identified. These are: (1) the Tagalog group, concentrated in Manila, the provinces of central and south central Luzon, and the island provinces of Mindoro, Marinduque, the most parts of Palawan; (2) Cebuano, in Cebu, Bohol, eastern Negros, western Leyte, and parts of Mindanao; (3) Ilongo, spoken in parts of Panay and western Negros; (4) Waray-Waray, spoken in Samar and Leyte; (5) Ilocano, dominant in many parts of northern Luzon; (6) Pampango, spoken in parts of central Luzon; and (7) Bikol, spoken in southern Luzon and on the islands of Catanduanes and Masbate. Other notable languages are Pangasinan and Ibanag, spoken in northern Luzon, and Aklan, spoken in Panay.

Pilipino

The national language of the Philippines, which is also the most widely spoken, is Pilipino, which is based on Tagalog; together with English, it is used as a medium of instruction in schools. Tagalog itself is the richest of all Philippine languages and has the most extensive written literature. Apart from Pilipino, there are two other official languages—English and Spanish. Only the United States and the United Kingdom have larger English-speaking populations than the Philippines.



Population density of the Philippines.

Religious groups. Out of an estimated population of almost 38,000,000, there are about 31,000,000 Roman Catholics; about 2,000,000 Aglipayans, adherents of a Filipinized form of Catholicism; almost 2,000,000 Muslims; about 1,500,000 followers of Iglesia ni Kristo, a form of Protestantism; and more than 500,000 Protestants. There are also some pagans and adherents of minor faiths.

Of late years Catholicism has been strengthened in the country by the increase in the number of Filipinos in the church hierarchy (cardinals, archbishops, and bishops), by the building of more seminaries, and by the increased involvement of the church in the social life of the country.

Demography. *Birth and mortality rates.* Because it is a Roman Catholic country and because most of its population is rural, the Philippines has a high birth rate. Birth control and family planning are rarely practiced.

Pestilence and epidemics may be said to be things of the past in the Philippines, but communicable diseases still constitute a problem. Thanks to education and good health conditions, life expectancy increased from 37.5 years in 1904, to about 52.5 in 1960, and to about 58 in 1968. The annual population increase of 3.3 percent is one of the highest in the world and represents a population increase of about 1,130,000 a year.

Immigration and emigration. There is a considerable amount of Filipino emigration, especially of manual workers and professional personnel. Many emigrants go to the United States, including Hawaii; to Okinawa in the Ryukyu Islands; to Guam in the Mariana Islands; to U.S. military bases in South Vietnam and Korea; and to Canada. There are more than 6,000 Filipino doctors and 8,000 Filipino nurses in the United States alone. Manual labourers usually go to Vietnam, Okinawa, Guam, and Hawaii, and to some parts of the U.S. mainland. Few Filipinos emigrate to other countries.

Distribution of population. In 1970 it was estimated that the population density numbered 327 people per square mile in the country as a whole. The population is unevenly distributed, with densities ranging between 41 and 3,962 per square mile. In Manila, there are about 89,000 persons per square mile. In Bukidnon Province on Mindanao the density is only 130, and in the Batanes Province on the Batan Islands it is only 136. About 3,295,000 people live in the northern Luzon region; about 5,099,000 in central Luzon; about 6,827,000 in southern Tagalog (excluding Manila); about 2,966,000 in the Bicol region of southeastern Luzon; about 5,413,000 in the east Visayan region; about 3,785,000 in the west Visayan region; and about 7,966,000 in Mindanao. The population of Quezon City is about 754,000.

Demographic trends. The population tends to move from rural areas to the towns and cities. Whereas at the beginning of the 20th century more than 80 percent of the population lived in rural areas, today only about 66 percent do so. Of those who have moved to the towns and cities, about 12 percent work in factories. The movement to the towns accelerated after World War II; Manila, for example, which today numbers 1,331,000, had a population of only 600,000 in 1931. Greater Manila today has a population of about 3,000,000. The farmlands, formerly to be found in its vicinity, have now been swallowed up by towns; two areas in particular—Pasay and Caloocan—have become cities in their own right, and Diliman now forms a part of Quezon City.

The movement to the towns

THE NATIONAL ECONOMY

Despite the government's optimistic view of progress made since independence, economic growth appears slow when compared to that of other Asian countries. The World Bank has estimated that between 1961 and 1968 the Philippines attained an average growth rate of less than 1 percent in its real per capita gross national product.

The extent and distribution of resources. *Mineral resources.* The Philippines is rich in mineral resources. There are deposits of gold in Mountain Province and Camarines on Luzon; iron ore in Surigao on Mindanao and Bulacan on Luzon; copper in Zambales on Luzon; lead and zinc in Marinduque; high-grade chromium in Zambales, Camarines Sur, and Pangasinan on Luzon, and Antique on Panay. Deposits of silver, nickel, mercury, molybdenum, (a silver-white element used as an alloy), cadmium (a white element resembling tin), and manganese are found in several other places. Nonmetallic minerals include cement, found in Cebu, Luzon, Romblon, Masbate, Panay, and Bohol; salt, found on Luzon; asbestos on Luzon; granite and marble on Romblon and Panay; asphalt in Leyte; mineral waters on Luzon and Mindanao; gypsum on Luzon; sulfur on Luzon; guano (sea bird manure used as fertilizer) and phosphate rock on Panay; and coal and silica in some other places.

Biological resources. About 40 percent of the Philippines' total land area is covered with rain forests. Of this, 12,000,000 hectares (46,000 square miles)—almost two-thirds of the forest area—abound with trees that have commercial value. Woods such as narra (a reddish wood used in cabinet-making), ipil (a very hard, durable wood), molave (a durable yellow hardwood), and kamagong (a dark wood with lighter grains running through it) are in particular demand. The lumbering industry, however, has tended to deplete the forest resources.

No fewer than 2,000 varieties of fish are in the seas surrounding the islands and in lakes, rivers, estuaries, and fish ponds. The most important commercial fishes are milkfish (a herringlike fish); *dilis* (anchovies); herring; sardines; mackerel; *alumahan* (mackerel); *tanginggi* (mackerel); lapu-lapu (grouper); snappers; *talakitok* (cavalla); *talilong* (mullet); barracuda; mudfish; and *dalagang bukid* (caesio). The people raise fish in ponds in some provinces of Luzon and Panay.

The Philippines' pearl fisheries constitute another biological resource. There are nine pearl beds in the Sulu Archipelago to the south. Pearl shells used for making buttons are exported to China, Hong Kong, and elsewhere. Tortoise shells and shells used for windows are exported to Singapore, Malaysia, and other countries.

Pearl fisheries

Philippines, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
Regions				
Region I (Manila area)				
Chartered City				
Manila	15	38	1,139,000	1,331,000
Region II (Ilocos and Mountain Province)				
Provinces				
Abra	1,535	3,976	115,000	146,000
Benguet	1,025	2,655	184,000	264,000
Ifugao	972	2,518	77,000	92,000
Ilocos Norte	1,312	3,399	287,000	343,000
Ilocos Sur	996	2,580	338,000	385,000
Kalinga-Apayao	2,721	7,048	90,000	136,000
La Union	576	1,493	293,000	374,000
Mountain Province	810	2,097	86,000	93,000
Region III (Cagayan Valley and Batanes)				
Provinces				
Batanes	81	209	10,000	11,000
Cagayan	3,476	9,003	445,000	581,000
Isabela	4,118	10,665	442,000	648,000
Nueva Viscaya	2,688	6,961	138,000	222,000
Region IV (central Luzon)				
Provinces				
Bataan	530	1,373	145,000	216,000
Bulacan	1,032	2,672	556,000	836,000
Nueva Ecija	2,040	5,284	608,000	851,000
Pampanga	842	2,181	617,000	907,000
Pangasinan	2,073	5,368	1,124,000	1,386,000
Tarlac	1,179	3,053	427,000	560,000
Zambales	1,434	3,714	213,000	343,000
Region V (southern Luzon and islands)				
Provinces				
Batangas	1,222	3,166	681,000	926,000
Cavite	497	1,288	378,000	520,000
Laguna	679	1,760	472,000	700,000
Marinduque	370	959	115,000	144,000
Occidental Mindoro	2,270	5,880	84,000	146,000
Oriental Mindoro	1,685	4,365	229,000	328,000
Palawan	5,751	14,896	163,000	237,000
Quezon	4,612	11,946	653,000	983,000
Rizal*	718	1,860	1,456,000	2,845,000
Region VI (Bicol and Masbate)				
Provinces				
Albay	986	2,553	515,000	674,000
Camarines Norte	816	2,112	188,000	262,000
Camarines Sur	2,034	5,267	820,000	948,000
Catanduanes	584	1,511	156,000	162,000
Masbate	1,563	4,048	336,000	493,000
Sorsogon	827	2,141	348,000	427,000
Region VII (western Visayas)				
Provinces				
Aklan	702	1,818	226,000	263,000
Antique	974	2,522	238,000	289,000
Capiz	1,017	2,633	315,000	394,000
Iloilo	2,056	5,324	966,000	1,168,000
Negros Occidental	3,060	7,926	1,332,000	1,504,000
Romblon	524	1,356	132,000	167,000
Region VIII (eastern Visayas)				
Provinces				
Bohol	1,590	4,117	592,000	683,000
Cebu	1,965	5,088	1,333,000	1,634,000
Leyte	2,420	6,268	963,000	1,111,000
Negros Oriental	2,218	5,746	598,000	715,000
Samar del Norte	1,343	3,480	261,000	306,000
Samar Occidental	2,166	5,609	369,000	442,000
Samar Oriental	1,676	4,340	238,000	271,000
Southern Leyte	670	1,735	210,000	251,000
Region IX (northern Mindanao)				
Provinces				
Agusan del Norte	4,462†	11,556†	271,000†	278,000
Agusan del Sur	4,462†	11,556†	271,000†	175,000
Bukidnon	3,202	8,294	194,000	415,000
Camiguin	89	230	45,000	54,000
Lanao del Norte	1,194	3,092	271,000	350,000
Lanao del Sur	1,495	3,873	378,000	456,000
Misamis Occidental	749	1,939	248,000	320,000
Misamis Oriental	1,378	3,570	344,000	473,000
Surigao del Norte	1,058	2,739	195,000	239,000
Surigao del Sur	1,758	4,552	165,000	259,000
Region X (southern Mindanao and Sulu)				
Provinces				
Cotabato	6,348	16,441	747,000	1,136,000
Davao del Norte	3,139	8,130	291,000	443,000
Davao del Sur	2,462	6,378	469,000	785,000
Davao Oriental	1,994	5,164	133,000	248,000
South Cotabato	2,840	7,356	282,000	466,000
Sulu	1,038	2,688	327,000	426,000
Zamboanga del Norte	2,346	6,075	281,000	409,000
Zamboanga del Sur	3,831	9,922	742,000	1,034,000
Total Philippines	115,830‡	300,000§	27,088,000‖	36,684,000‖

*The area and population of Manila are excluded from the corresponding totals for Rizal Province. †Agusan Province was subdivided before the 1970 census into Agusan del Norte and Agusan del Sur. No breakdown of area or adjusted 1960 census figures is available. Figures given are for the entire Agusan Province. ‡Converted area figures do not add to total given because of rounding. §The total area was reported as land area, but appears to include some inland water. ‖Figures do not add to total given because of rounding.
Source: Official government figures.

Hydroelectric and other power resources. Hydroelectric power in the past has supplied only a small proportion of the country's electrical output, with thermal plants, most of which burn imported oil, supplying the major proportion. But with the completion of the Binga and Ambuklao dam projects on Luzon and the expansion of the Maria Cristina project on Mindanao, Philippine hydroelectric potential is being put to use. Irrigation and flood control are additional benefits of the Upper Pampanga River Project. The new Bataan thermal plant will serve as a supplement to the Luzon hydroelectric plants, especially during the dry season.

Sources of national income. *Agricultural resources.* The Philippines is chiefly an agricultural country. Its soil is rich and fertile, and its products grow abundantly throughout the year. The country's potential farmland amounts to more than 16,000,000 hectares, but only about half of this area is under cultivation.

The principal farm products are rice, coconut, sugar, abaca (Manila hemp), tobacco, corn, maguey (a fleshy fibrous plant used for making ropes, etc.), and pineapples. Many tropical fruit trees and plants are also raised, the most important being bananas, mangoes, lanzones (a grape-like fruit with a thick yellow skin), oranges, papaya, avocado, star apples (a bland-tasting purple fruit), atis (a white pulpy fruit), pomelos (a species of grapefruit), nangca (jackfruit), apples, strawberries, guavas, santol (a reddish fruit used for preserves), and durian (an edible fruit with a prickly rind and an offensive odour). Vegetables such as beans, tomatoes, pechay (Chinese cabbage), cabbages, mustard, lettuce, camotes (sweet potatoes), sugar beet, cassava, and squash are raised on farms and in gardens for domestic consumption.

Rice, the principal staple crop, is grown everywhere, but especially in the provinces of Batangas, Laguna, Nueva Ecija, Pampanga, Pangasinan, and Tarlac on Luzon; in Cotabato on Mindanao; in the western province of Negros and in the provinces of Capiz and Iloilo on Panay. More than 3,000,000 hectares, or almost one-half the total area under cultivation, is used for rice growing. In 1968 the Philippines became a rice exporting country for the first time for many decades. The increase in rice production was due primarily to the efforts of the government and Filipino farmers, assisted by the U.S. financed International Rice Research Institute.

The Philippines is the world's largest producer of coconuts and the principal exporter of coconut products. The area devoted to coconut production—about 2,000,000 hectares—is second only to that used for rice, and about

12,000,000 Filipinos are dependent upon this crop for their livelihood. The Philippines also produces about one-third of the world's copra (dried coconut meat used for oil extraction and as animal feed) output, which earns for the government about 40 percent of its total income from exports.

Sugar is the country's third largest export and earns a substantial amount of foreign exchange for the government. Sugarcane is grown extensively in the Batangas, Laguna, Pampanga, and Tarlac provinces of Luzon; in western Negros; and in Iloilo Province of Panay. About one-seventh of the Filipino population depends on this industry for a living. Current production amounts to about 2,000,000 tons, 60 percent of which is exported, mostly to the United States.

Manila hemp

Abaca, or Manila hemp, is the best hemp fibre available. It ranks high among the country's chief exports, although only about 200,000 hectares, or approximately 2 percent of the country's cultivated area, are used for its production, which employs about 3,000,000 people. It is grown extensively in the Davao provinces of Mindanao, in the Bicol region of Luzon, in Leyte, and in Samar.

Timber, lumber, and veneer products are the leading exports. Because of their high quality these products are in great demand. Other forest products include dyewoods, rattan, tanbarks, bamboo, gutta-percha (a gum), beeswax, rubber, and edible birds' nests.

Fishing also constitutes one of the most important of Filipino industries. Canned tuna is the principal fish exported. Commercial fishing is carried on primarily off Palawan, Negros, Mindanao, and Panay. Fish provides an important proportion of the protein in the Filipino diet.

Mining and manufacturing. Gold, silver, iron, copper, manganese, zinc, lead, coal, and chromite ore are mined. Manufactured products include cement, plywood, chemicals, electrical equipment, medical supplies, radios and musical instruments, plastics, and rubber and steel goods.

Financial services and foreign trade. In addition to the Central Bank of the Philippines, there are more than 40 commercial banks, some of which have more than 100 branches and agencies. Banking operations are also conducted by such other institutions as the Postal Savings Bank, the Agricultural Credit Administration, and the Development Bank of the Philippines. The principal exports, as mentioned, are timber, coconuts and coconut products, copra, sugar, and abaca, while the principal imports are machinery, fuel, textiles, electrical equipment, and transport equipment. The United States accounts for about 35 percent of the total trade of the Philippines; other trading partners are Japan, West Germany, The Netherlands, the United Kingdom, and Australia.

Management of the economy. *The private sector.* The economy of the Philippines is based on free enterprise; individuals and nongovernmental entities are free to participate in its development and management, sometimes with the aid, in whole or in part, of governmental credit. As a result of this encouragement, many manufacturing industries have been established since 1950; new businesses have, indeed, sprung up like mushrooms. Manufacturing firms have risen in number from fewer than 1,000 in 1949 to 29,000 in 1969, a rate of expansion that has been reflected in an increase in the variety of products available.

New manufacturing industries

The public sector. The government has assisted the private sector of the economy by exempting certain new industries from taxation for a certain period. Only nominal taxes are imposed on selected industries, and as already mentioned, credit, or loans on favourable terms, are available to others. Government banks, such as the Central Bank of the Philippines, the Philippine National Bank, the Development Bank, the Agricultural Bank, and others were established to encourage business, agriculture, and industry. Half of the original capital investments of thousands of private rural banks was provided by the government. Private development banks have likewise received governmental assistance. The National Economic Council and the Board of Investments have been created to help both public and private sectors in undertaking studies and in planning further economic development.

The government has also to some extent been engaged in business in its own right. It either owns or operates the National Development Company, the National Shipyards and Steel Corporation, the Philippine National Railways Company, the National Rice and Corn Administration, and other companies. The Philippine government has, however, on the whole been a failure in business because of the interference of party politics in management.

Taxation. The government derives its revenue from three major sources—first, from taxation; second, from earnings and other credits; and third, from extraordinary income, including the transfer from special funds (funds derived mainly from unexpended balances in the budget which are deposited as savings accounts that draw interest). About 85 percent of revenue is collected through the Bureau of Internal Revenue and the Bureau of Customs. The remainder is collected through other offices and represents receipts from land surveying and similar sources.

Internal revenue taxes constitute the largest portion of government income; customs revenues come from imported goods and from taxes on shipping. Revenues from earnings and other credits include service income, prior year credits, income from rentals, interest on investments, and miscellaneous receipts. Local governments have only limited taxing powers.

The main items of national expenditure are, in order of importance, education, representing about 30 percent of the total; public works and communications; national defense; the executive department; agriculture and natural resources; health; repayment of national debts; the judiciary; revenue collection; the Congress; foreign affairs; aid to local governments; social welfare; gratuities; and pensions.

Trade unions and employer associations. The strength of the trade unions is greater today than at any time in the past. The trade union movement is well established in Manila and in most other towns and cities. Farmers' or tenants' groups are also organized, as are the teachers and government employees. In the 1960s student groups sympathetic to labour and to political and social reform became active. Management, for its part, has organized company unions. Relations between trade unions and the employers union have been generally untroubled. The Court of Industrial Relations and the Court of Agrarian Relations have been effective in settling disputes between labour and management.

Economic politics. The government, through the National Economic Commission, the National Planning Commission, and other agencies, plans national economic development. In so doing, it has sought to increase economic independence, for while there have been advantages in the preferential trade relations between the U.S. and the Philippines, this trade resulted in the development of those industries that would primarily benefit the U.S. rather than the Filipino economy. While the trade preference agreements remained in force until 1974, new markets, including those in Communist countries, were explored. Trade with Japan is rapidly increasing as is that with The Netherlands, Indonesia, Canada, and Taiwan. Internally, there is to be an attempt to develop the Philippines as both an agricultural and an industrial country. A foreign investments incentive law was passed in 1969 by Congress to encourage joint investments by Filipino and foreign capital and to permit the use of foreign managerial skills. An attempt is also being made to encourage Filipino residents abroad to invest in the Philippines.

New markets

Problems and prospects. It has frequently been observed, first, that the Philippines is rich but that the people are poor, and, second, that the progress attained has been made by an insignificant number of individuals. Average per capita income still remains less than 1,000 pesos ($255) annually. About 55 percent of the people are still engaged in agricultural food production and only about 11 percent in manufacturing; about 2,000,000 people either have no work or are underemployed. About 10

percent of the people—those in the highest income group—receive 40 percent of national earnings, with the remaining 90 percent receiving the remaining 60 percent of this income. The government itself remains poor; despite its responsibility for a territory of 115,830 square miles and a population of about 38,000,000 people, its budgetary outlay for the financial year 1972–73 was still little more than U.S. $780,000,000. Historically, the traditional cacique (landowning) system, the personal ambitions of many government officials, and 400 years of colonial rule have brought about this condition. Full development of the economy and equitable distribution of the national wealth are the two main economic problems. By the mid-1970s various plans for bolder solutions, including several five-year plans, are expected to show results. Some observers have predicted revolution unless radical reforms are instituted to develop agriculture, commerce, and industry, to distribute more equitably wealth among the people, and to end the abuses of caciquism (the landownership system) and of the political machine.

Transportation. *Road networks.* Thousands of miles of roads—gravel, asphalt, and cement—link the towns on the archipelago's many islands. About 30 miles of modern expressways run north and south from Manila. Elsewhere on Luzon, most of the road to Laoag, 250 miles north of Manila, and the greater part of the road to Sorsogon in the south, are cemented, as are roads branching out to the city of Baguio and to the provinces of Batangas, Camarines Norte, Cavite, Pangasinan, and Zambales. Thousands of miles of roads (of various types) have also been constructed in Mindanao, Mindoro, Palawan, and the Visayas. About 750 miles of road were cemented and over 1,000 miles asphalted between 1966 and 1969 alone. A system of roads, bridges, and ferries to connect the islands of Luzon, Samar, Leyte, and Mindanao was under construction in the 1970s. At the end of 1970 about 450,000 motor vehicles were in use on the highways; the average of one motor vehicle for every 82 inhabitants is the highest of any Asian country except Japan. In Manila and in most other cities taxi services are usually available at low rates.

Railways. Two principal railroads operate in Luzon and Panay. The Luzon line, with about 730 miles of track, extends from San Fernando in La Union Province in the north to the city of Legazpi in Albay Province in the south. The railroad in Panay, about 90 miles in length, connects the city of Iloilo, in Iloilo Province, with Roxas City, in Capiz Province. Both railroads suffer from competitive air services and bus transport.

The port of Manila

Port facilities. The most important port is Manila, which handles from 70 percent to 80 percent of all shipping. Manila North Harbor accommodates domestic trade, while the Manila South Harbor, which can accommodate as many as 30 ships, handles shipping from abroad. Other ports used by international shipping are Cebu (next in importance to Manila), Iloilo, Davao, Zamboanga, Cagayan de Oro, Legazpi, and Bacolod. There are more than a hundred smaller ports.

Air transport. The international airport at Manila, like those at Hong Kong and Singapore, is a focal point for air routes and is used by about 25 air carriers. A separate terminal is used by domestic airlines. The Mactan airport, on Mactan Island near Cebu, is also international. There are also 77 other airports for domestic flights, all connected to Manila. The domestic airlines—apart from Philippine Air Lines, which is also international—are Air Manila, Filipinas Orient Airways, and Delta Air Lines.

ADMINISTRATION AND SOCIAL CONDITIONS

The constitutional framework. Work on a new constitution began in 1971. The Filipino constitution, which remains in force, adopted in 1935 and amended in 1940 and 1946, provides for a republican system of government in which sovereignty resides in the people and is exercised by them through universal suffrage. The national government provided for in the constitution has three independent, separate, but coordinate branches—the executive branch, the legislative branch, and the judiciary. At the head of the executive is the president, who is elected directly by a national election for a term of four years; he is forbidden by the constitution to hold office for more than eight consecutive years. A vice president, whose qualifications and tenure are similar to those of the president, is elected at the same time. Only natural-born Filipino citizens, 40 years of age or over, resident in the Philippines for at least two years before an election and qualified to vote, are eligible for the presidency. To assist him in his official functions, the president appoints department secretaries (or ministers), who are subject to confirmation by the Commission on Appointments—an independent body of 25 members, composed of 12 senators and 12 congressmen meeting under the chairmanship of the president of the Senate. Governmental departments include those for foreign affairs, finance, defense, education, social welfare, health, and agriculture and natural resources.

Laws are enacted by Congress composed of the Senate and the House of Representatives. The Senate is composed of 24 members who are elected by nationwide election for a term of six years. The qualifications for senator are the same as those for the presidency, except that the minimum age requirement is 35 years. One-third of the senators end their terms of office every two years; they may, however, be re-elected for an indefinite number of terms. The Senate, together with the president, enjoys treaty-making powers.

The House of Representatives is composed of not more than 120 members, who are elected for a term of four years and who may be re-elected an indefinite number of times. Members of the House of Representatives must be at least 25 years old.

The judiciary

The judiciary is composed of the Supreme Court, consisting of 11 justices, and of minor courts, including a court of appeals (composed of 15 justices), courts of first instance, and city and municipal courts. In addition, there are special courts concerned with industrial relations, agrarian relations, and family relations. Justices and judges of these courts are appointed by the president of the Philippines, subject to confirmation by the Commission on Appointments.

Other agencies created by the constitution are the Commission on Elections, the Presidential Electoral Tribunal, the Senate Electoral Tribunal, the House Electoral Tribunal, the General Auditing Office, and the Civil Service Commission.

Local government. The constitution empowers Congress to enact laws establishing local governments. The country is divided into 66 provinces, 4 subprovinces, 58 chartered cities, 1,369 municipalities, 30 municipal districts, and 29,948 barrios (villages). Each local unit has its defined territory. Cities, and even provinces, are, however, sometimes created primarily for political ends. Each province, city, municipality, municipal district, and barrio has its own local government, and its own elected and appointed officials. Each province is headed by a governor, each subprovince by a deputy governor, and each city or municipality by a mayor. The smallest political subdivision is the barrio, administered by a council, which is headed by a barrio captain.

The political process. *Elections.* Elections are held every four years. In electing the principal officials of both national and local governments, male and female voters must be citizens of the Philippines, must be 21 years of age or over, must know how to read and write, and must not have a mental disability or a criminal culpability.

Parties. Many political parties have existed since the early years of the 20th century, most of them having nationalistic leanings, but there is essentially a two-party system: the Nacionalista Party and the Liberal Party. Both parties are nationalistic with little ideological difference. Party strength is largely dependent upon the attractiveness of personal leadership.

The participation of the citizen. Most Filipinos take a keen interest in politics and the percentage of registered voters (almost 50 percent of the total population) who exercise the right to vote in elections (often 80 percent) is higher than in many other countries, including the United

States. Political meetings and rallies are well attended. Critics have claimed that vote-buying has become a common practice, that the strict laws governing limitation of election expenses are not enforced, and that only rich men, or candidates supported by them, win the elections.

Political developments. Filipinos adhere strongly to the democratic system. While they prefer free enterprise and reject Communism, Socialism has made some impression upon them and there is a growing movement to extend more government assistance to the less fortunate. Before World War II Socialism found adherents among the rural population of central Luzon and Manila. During the war they formed a guerrilla unit called Hukbalahap, meaning "Peoples' Army Against Japan." After the war they maintained their organization against what they termed a government of oligarchs. It appears that they maintain some sort of intelligence with the Communist government of China. Their activities are confined to central Luzon and are well under government control. Personal leadership remains, however, a key factor in political life—a circumstance which sometimes unfortunately encourages a degree of demagoguery.

Justice. Equal justice under the law for all is guaranteed by the constitution. The administration of justice is generally fair, especially at the higher level of the judiciary. Justices and judges enjoy fixed tenure and reasonably high compensation. The judiciary has consequently been less criticized than other branches of the government, and from 1900 to the present, the Supreme Court has retained the trust of the people.

The armed forces. The Department of National Defense is divided into four services: the army, the navy, the air force, and the constabulary (the national police force). Each branch has a commanding officer whose rank is at least that of brigadier general or, in the navy, commodore. Altogether about 58,500 men are in active service and many more in the reserves. Several army battalions have served in both the Korean and Vietnam wars. The armed forces are responsible for containing the dissident activities of the rebellious Hukbalahap movement in central Luzon, for controlling smuggling, and for maintaining order. Civilian authority prevails over military authority. Thus the president of the Philippines, who is commander in chief of the armed forces of the Philippines, is a civilian, as is the Secretary of National Defense.

Role of the U.S. in defense

Under a series of agreements reached in 1947, the United States maintains several military, naval, and air force bases in the Philippines. The agreements are in effect until the year 2046. The United States has also agreed to provide military assistance in the form of equipment and training, while the Philippines offers financial support to American advisers and trainees in the country.

Administration. *Educational services.* The largest governmental department is that of education, which administers the bureaus of public schools, private schools, and vocational education, as well as the Institute of National Language, the national library, the National Museum, and the National Historical Commission. The largest is the Bureau of Public Schools, which employs about 280,000 teachers and is responsible for the instruction of about 11,500,000 pupils. The Bureau of Private Schools supervises the private or nongovernmental schools, colleges, and universities, consisting in all of about 2,800 institutions, with a total enrollment of about 800,000 and a teaching staff or about 20,000. Five of the universities each have an enrollment of more than 30,000.

There are more than 40 universities, of which a third are located in greater Manila, the oldest being the University of Santo Tomás, founded in 1611.

Elementary education lasts for six years, secondary education for four years, and college courses last from four to nine years. Vocational schools offer specialized training. The government also administers a number of teacher training schools.

Health and welfare services. Health and welfare services are the responsibility of the Department of Health and the Department of Social Welfare. There are several national government hospitals, and each province and city also maintains its own hospitals. There are altogether more than 770 hospitals, both government and private. There are also almost 1,500 rural health units, each of them serving an average of about 25,000 people. In addition to the department of social welfare, there are a number of nongovernmental social welfare agencies, such as the Philippine National Red Cross, the Mental Health Association, the Community Chest, the Philippine Band of Mercy, the Social Welfare Council, and the Youth Coordinating Council.

Housing. The movement of population to the cities, especially to greater Manila, has resulted in a housing shortage. Many poorer people have consequently built makeshift huts—locally called *barung-barong*—on land belonging either to the government or to private individuals. The government has attempted to relocate the inhabitants of these shanties in multistoried apartments, but only a small proportion have so far been accommodated. Housing remains a major problem—a problem compounded by the associated problems of sanitation, public health, the maintenance of order, and the security of the individual's person and property.

Police services. The constabulary has divided the national territory into four regional zones—two on the island of Luzon, one in the Visayas, and one on Mindanao. Each province, including the cities within its territorial jurisdiction, is assigned a constabulary unit under a provincial command. Local police units are also maintained by the government in each city and municipality under the administrative and financial supervision of the local government.

Social conditions. *Wages and cost of living.* The daily minimum for most workers in greater Manila is ₱12 ($3.05); elsewhere it is ₱5.50 ($1.41). While governmental employees and most employees in the larger commercial concerns enjoy the benefits of this law, many others —including many municipal government employees and most labourers in rural areas—do not, primarily because of the inability of their employers to pay the amount. The ordinary wage earner consequently receives very little; the average annual wage barely exceeds ₱1,000 ($255). During the 1960s, moreover, the rising cost of living—especially after the devaluation of the Filipino peso in early 1970—sharply affected the small wage earner.

The minimum wage law

Health condition. Public health conditions have much improved since 1940. The last epidemic of smallpox, for example, occurred in 1930 and the last outbreak of cholera in 1934. There is, however, still an incidence of *el tor* (a type of cholera), tuberculosis, pneumonia, influenza, and diarrhea. Tuberculosis is the primary cause of death; pneumonia, bronchitis, and such viral diseases as influenza, measles, and poliomyelitis are still widespread. Malaria and leprosy have, however, been brought under control. Health education among the people has had much to do with improvement in health conditions.

Social and economic strata. The Philippines is a democratic country and has no caste system. Existing social and economic divisions are between the rich, the middle income group, and the poor. The divisions, however, are not immutable. The few haciendas (large landed estates), owned by the church or by Spanish families, that existed in the days of Spanish rule have partly disappeared. Some haciendas, however, still remain, but are now mostly owned by Filipino families. There are large plantations of sugar, rice, and other crops in Tarlac, Nueva Ecija, Pampanga, Bulacan, Laguna, Batangas, Negros, Iloilo and other provinces. As the value of businesses, industries, and real estate rose since independence some families acquired great wealth, constituting a group that has half-humorously been classified into three groups —the idle rich, the profligate rich, and the working rich. Billionaires and millionaires, although constituting only 2 percent of the population, own most of the wealth of the Philippines. But the majority of the population—more than 75 percent—still remain poor. A third group is constituted by the professionals, bankers, lawyers, physicians, teachers, office managers, and technologists—some of whom have acquired some wealth.

Thus today, at one extreme, the wealthiest individuals have an annual income of ₱100,000 ($25,650) or more each, whereas the per capita income of a labourer averages only ₱650 ($165) a year. This somewhat striking disparity of wealth is the origin of much of the current popular discontent. Although measures for land reform, social legislation, and a revision of the income tax laws have been introduced, the wealthy retain control of the economic and political life of the country.

CULTURAL LIFE AND INSTITUTIONS

While geographically the Philippines belongs to the East, culturally it is of the West. Its Oriental background still remains in the barrio with its bamboo and nipa (palm thatch) dwellings, its close-knit family system, and its folklore. In most large towns and cities, however, the Oriental culture has been replaced by Spanish and American culture. The presence of church, convent, and patio, certain folk dances, the *moro-moro* plays (*i.e.*, plays depicting the wars between "Moors" and Christians), the *Ati-atihan* (a passion play performed during Lent), and the Spanish language spoken still by the *ilustrados* (an élite claiming Spanish descent), testify to the vitality of Spanish culture. The English language, the public school system, and many of the appurtenances of modern living, demonstrate the degree of American influence. The Spanish and English languages in particular have contributed to an understanding and an appreciation of much in the West and have also to some degree inculcated Western standards. More, indeed, is often known of Western countries than of many Asian countries that, geographically, are closer neighbours.

The state of the arts. Early Spanish chroniclers testified that the early Filipinos carved the images of their *anitos* (gods and goddesses) and ancestors in wood. They also used various musical instruments such as gongs, drums, cymbals, stringed instruments, flutes (including nose flutes), and guitars. They performed appropriate songs and dances to celebrate courtship, marriage, the harvest, war, and other occasions. Many of these songs and dances, accompanied by the same musical instruments, have been preserved or reinterpreted today, and are presented by various dance and song groups. Some of these groups—such as the Bayanihan, the Filipinescas, the Baranggay, the Hariraya groups, as well as the Karilagan Ensemble, and the groups associated with the guilds of the Manila and Fort Santiago theatres—have performed not only in the Philippines but also in other countries. Many Filipino musicians and composers have risen to prominence, such as Antonio J. Molina, a composer and conductor, and Felipe P. de Leon, a composer known for his nationalistic themes.

Early Filipino art

Filipino painters have included Juan Luna (dean of Filipino painters), Fernando Amorsolo (known for his traditional rural scenes), Carlos V. Francisco (a muralist), Victorio Edades (a painter in the modern style), Vicente Manansala (a mural painter), and Arturo Rogerio Luz (an exponent of structural painting), while among sculptors Guillermo Tolentino and Napoleon Abueva are prominent. Tribesmen from mountainous northern Luzon and craftsmen at Paete, in Laguna Province, and Pampanga Province are famous for woodcarving, while Romblon Province is noted for its marble sculptures. Among Filipino architects, the most notable include Juan F. and Angel Nakpil, Fernando Ocampo, Leandro Locsin, Juan Arellano, Carlos Arguelles, and Tomas Mapua.

Folk traditions and popular culture. Filipinos have a rich folklore tradition. Myths and legends deal with such subjects as the origin of the world, the first man and woman on Earth, why the sky is high, why the ocean water is salty, and why there are different races. Other tales are associated with the Spanish conquest, while Muslim Filipinos have an epic called *Indarapatra and Sulayman*, and the Ilocanos of northern Luzon have another called the *Biag ni Lam-ang*. Most folktales and popular cultural traits are still to be found in remote barrios and *sitios*. Some authors, such as Dean S. Fansler, Maximo Ramos, and Armando and Carolina Malay, have compiled collections of these folktales.

Cultural institutions. The Bureau of Science and the National Museum, both in Manila, have gathered artifacts connected with many of the folk tales and popular beliefs. Nongovernmental organizations, such as the Conservation Society and some historical groups, have also sought to preserve the folklore heritage. A cultural centre in Manila was inaugurated in 1970 to encourage interest in the heritage of the Philippines, with emphasis on painting, sculpture, and the performing arts.

The mass media. The constitution guarantees freedom of press, speech, and assembly; and the mass media in the Philippines enjoys more freedom than in any other country in Asia and is among the most untrammeled in the world. Newspapers are published in English, Pilipino, Spanish, and many vernacular dialects in Manila, Cebu, Davao, and elsewhere. Radio and television broadcasting stations also provide news coverage, reaching a wider public. Transistor radio receivers diffuse news even in places where public electricity is not yet available. The most widely circulated daily newspapers are the *Manila Times*, the *Manila Chronicle*, the *Manila Daily Bulletin*, the *Philippines Herald*, and the *Daily Star*, all published in English. The *Taliba* and the *La Vanguardia* are published in Pilipino and Spanish, respectively.

Prospects for the future. The Republic of the Philippines has maintained a system of popular, representative government despite the wide gap between the small rich population and the poor majority. Minor political disturbances are viewed as natural to a developing nation; the desire for social welfare legislation is growing, and the increasingly politically active poor are becoming impatient. As the national economy expands, however, the government will acquire the means to improve the standard of living, and political stability within the country does not seem to be much threatened.

BIBLIOGRAPHY. P.S. DE ACHUTEGUI and M.A. BERNAD, *Religious Revolution in the Philippines*, 2 vol. (1960–66), a general description of the religious schism in the Catholic Church in the Philippines which resulted in the founding of the Philippine Independent Church, more popularly called Aglipayanism; E. AGUINALDO and V. ALBANO PACIS, *A Second Look at America* (1957), a recount of Aguinaldo's exploits against the American occupation forces in 1899–1901; E.M. ALIP, *Philippine History: Political, Social, Economic*, rev. ed. (1969), a general history covering pre-European, Spanish, American and present era, *The Government of Our Republic*, rev. ed. (1969), a general description of the development, organization, and functions of the government of the Republic of the Philippines, *Political and Cultural History of the Philippines*, 2 vol. (1958–69), a more detailed history from the earliest times to 1969; A. CRAIG and C. BENITEZ (eds.), *Philippine Progress to 1898* (1916), useful material about the Philippines from pre-European times to 1898; M.M. KALAW, *The Development of Philippine Politics, 1872–1920* (1926), a historical study of the development of political parties in the Philippines; C.M. RECTO, *Three Years of Enemy Occupation* (1946), personal impressions of the Japanese occupation of the Philippines, 1942–45; *Yearbook of Philippine Statistics*, an official government publication issued at intervals.

(Eu.M.A.)

Philippines, History of the

In ancient times, the inhabitants of the Philippines were a diverse agglomeration of peoples who arrived in various waves of immigration from the Asian mainland and who maintained little contact with each other. The contemporary Filipino can boast of no great architectural reminders of vast kingdoms of the past, as in Java or Cambodia. The Philippines is the only nation in Southeast Asia that became subject to Western colonialism before it had developed either a centralized government ruling over a large territory or an advanced elite culture. Chinese traders had been resident since about AD 1000, and some cultural influences from South Asia, such as a Sanskrit-based writing system, were carried to the islands by the Indonesian empires of Śrivijaya and Majapahit; but in comparison with other parts of the region, the impact of both China and India on the Philippines was of little importance. The peoples of the Philippine archipelago, unlike most peoples of Southeast Asia, never adopted Hinduism or Buddhism.

Pre-Spanish history. According to what can be inferred from somewhat later accounts, the Filipinos of the 15th century must have been primarily shifting cultivators, hunters, and fishermen. Sedentary cultivation was the exception. Only in the mountains of northern Luzon, where elaborate rice terraces were built in the early centuries of the Christian era, were livelihood and social organization linked to a fixed territory. The lowland peoples lived in extended kinship groups known as *barangays* under the leadership of a *datu,* or chieftain. The *barangay,* usually numbering a few hundred persons, was usually the largest stable economic and political unit.

Barangay society

Within the *barangay,* the status system, though not rigid, appears to have consisted of five classes: the *datu* and his family; the nobility; the freeholder; the sharecropper; and, at the bottom, the debt peon or war captive, termed slaves by Spanish observers. The lowest status was inherited but, through manumission and interclass marriage, seldom extended over more than two generations. The fluidity of the social system was, in part, the consequence of a bilateral kinship system in which lineage was reckoned equally through the male and female line. Marriage was apparently stable, though divorce was socially acceptable under certain circumstances.

Animism was the religion of the early Filipino, a mixture of monotheism and polytheism in which the latter dominated. The propitiation of evil spirits required numerous rituals, but there was no religious hierarchy. In religion, as in social structure and economic activity, there was considerable variation between—and even within—islands.

This pattern, however, began to change in the 15th century, when Islām was introduced to Mindanao and the Sulu islands through Brunei. Along with changes in religious beliefs and practices came new political and social institutions. By the mid-16th century two sultanates had been established, bringing under their sway a number of *barangays.* A powerful *datu* as far north as Manila embraced Islām. It was in the midst of this wave of Islāmic proselytism that the Spaniards arrived. Had they come a century later or had their motives been strictly commercial, Filipinos today might be a predominantly Muslim people.

The Spanish period. Spanish colonialist motives were not, however, strictly commercial. The Spanish at first viewed the Philippines as a stepping stone to the riches of the Spice Islands, but, even after the Portuguese and Dutch had foreclosed that possibility, the Spaniards still maintained their presence in the archipelago.

Ferdinand Magellan headed the first Spanish foray to the Philippines in 1521, meeting an untimely death in Cebu. After three further expeditions had ended in disaster, Philip II, "the most Catholic of kings," after whom the islands are named, sent out Miguel López de Legazpi, who established the first permanent Spanish settlement, in Cebu, in 1565. The Spanish city of Manila was founded in 1571, and by the end of the 16th century most of the coastal and lowland areas from Luzon to northern Mindanao were under Spanish control. Friars marched with soldiers and soon accomplished the nominal conversion to Catholicism of all the natives under Spanish administration. But the Muslims of Mindanao and Sulu, whom the Spanish called Moros, were never completely subdued by Spain; a 20th-century U.S. army was to be required for that task.

Spanish rule for the first 100 years was exercised in most areas through a type of tax farming imported from the Americas known as the *encomienda.* But abusive treatment of the native tribute payers and neglect of religious instruction by *encomenderos,* as well as frequent withholding of revenues from the crown, led to the abandonment of the system by the end of the 17th century. The governor general, himself appointed by the king, began to appoint his own civil and military governors to rule directly.

The *encomienda*

Central government in Manila retained a medieval cast up to the 19th century, and the governor general was so powerful he was often likened to an independent monarch. He dominated the Audiencia, or high court, was captain general of the armed forces, and enjoyed the privilege of engaging in commerce for private profit.

Manila dominated the islands not only as the political capital. The galleon trade with Acapulco assured Manila's commercial primacy as well. This exchange of Chinese silks for Mexican silver kept in Manila those Spaniards seeking quick profit, as well as attracting a large Chinese community. The Chinese, despite several massacres at the hands of suspicious Spaniards, persisted and soon established a dominance of commerce that they have never relinquished.

Manila was also the ecclesiastical capital. The governor general was civil head of the church in the islands, but the archbishop vied with him for political supremacy. In the late 17th and 18th centuries, buttressed with the legal status of lieutenant governor, the archbishop frequently won. Augmenting their political power, religious orders, Catholic hospitals and schools, as well as bishops acquired great wealth, mostly in land. Royal grants and devises formed the core of their holdings, but many arbitrary extensions were made beyond the boundaries of the original grants.

Power of the church

The power of the church derived not simply from wealth and official status. The priests and friars had a command of native languages rare among lay Spaniards; and in the provinces they outnumbered civil officials. Thus, they were an invaluable source of information to the colonial government. The cultural goal of the Spanish clergy was nothing less than the full Christianization and Hispanization of the Filipino. In the first decades of missionary work, native animism was vigorously suppressed; old practices were not tolerated. But as the Christian laity grew in number and the zeal of the clergy waned, it became more and more difficult to prevent the preservation of ancient beliefs and customs under Catholic garb. Thus, even in the area of religion, pre-Spanish Filipino culture was not entirely destroyed.

Economic and political institutions were also altered under Spanish impact but perhaps less thoroughly than in the religious realm. The priests tried to move all the people into pueblos, or villages, surrounding the great stone churches. But the dispersed demographic patterns of the old *barangays* largely persisted. Nevertheless, the *datu*'s once hereditary position became subject to Spanish appointment.

Agricultural technology changed very slowly until the late 18th century, as shifting cultivation gradually gave way to more intensive sedentary farming, partly under the guidance of the friars. The socio-economic consequences of the Spanish policies that accompanied this shift reinforced class differences. The *datus* and other representatives of the old noble class took advantage of the introduction of the Western concept of absolute ownership of land to claim, as their own, fields cultivated by their various retainers, even though traditional land rights had been limited to usufruct. These heirs of pre-Spanish nobility were known as the *principalia* and played an important role in the friar-dominated local government.

The 19th century. By the late 18th century, political and economic changes in Europe were finally beginning to affect Spain and, thus, the Philippines. Important as a stimulus to trade was the gradual elimination of the monopoly enjoyed by the galleon to Acapulco. The last galleon sailed in 1811, and by the 1830s Manila was open to foreign merchants almost without restriction. The demand for Philippine sugar and abaca, or hemp, grew apace, and the volume of exports to Europe expanded even further after the completion of the Suez Canal in 1869.

The growth of commercial agriculture resulted in the appearance of a new class. Alongside the landholdings of the church and the rice estates of the pre-Spanish nobility there arose haciendas of coffee, hemp, and sugar, often the property of enterprising Chinese–Filipino mestizos. Some of the families that gained prominence in the 19th century still play an important part in Philippine economics and politics.

Not until 1863 was there public education in the Philip-

pines, and even then the church controlled the curriculum. Less than 20 percent of those who went to school could read and write Spanish, and far fewer could speak it properly. The limited higher education in the colony was entirely under clerical direction, but by the 1880s many sons of the wealthy were sent to Europe to study. There, nationalism and a passion for reform blossomed in the liberal atmosphere. Out of this talented group of overseas Filipino students arose the so-called Propaganda Movement. Magazines, poetry, and pamphleteering flourished. José Rizal, this movement's most brilliant figure, produced a political novel, *Noli me tangere* (Eng. trans., *The Social Cancer*), which had a wide impact in the Philippines. In 1892 Rizal returned home and formed the Liga Filipina, a modest reform-minded society, loyal to Spain, that breathed no word of independence. But Rizal was quickly arrested by the overly fearful Spaniards, exiled to a remote island in the south, and finally executed in 1896. Meanwhile, within the Philippines there had developed a firm commitment to independence among a somewhat less privileged class.

The Katipunan

Shocked by the arrest of Rizal, these activists quickly formed the Katipunan under the leadership of Andres Bonifacio, a self-educated warehouseman. The Katipunan was dedicated to the expulsion of the Spaniards from the islands, and preparations were made for armed revolt. Filipino rebels had been numerous in the history of Spanish rule, but now for the first time they were inspired by nationalist ambitions and possessed the education that made success a real possibility.

The Philippine Revolution. In late 1896, Spanish friars uncovered evidence of the Katipunan's plans, and its leaders were forced into premature action. Revolts broke out in several provinces around Manila. After months of fighting, severe Spanish retaliation forced the revolutionary armies to retreat to the hills. In December 1897 a truce was concluded with the Spaniards. Emilio Aguinaldo, son of a municipal mayor and commander of the rebel forces, was paid a large sum and was allowed to go to Hong Kong with other leaders; the Spaniards promised reforms as well. But reforms were very slow in coming, and small bands of rebels, distrustful of Spanish promises, kept their arms; clashes grew more frequent.

Meanwhile, war had broken out between Spain and the United States. After the U.S. naval victory at Manila Bay in May 1898, Aguinaldo and his entourage returned to the Philippines with the help of Admiral Dewey. Confident of U.S. support, he reorganized his forces and soon liberated several towns south of Manila. Independence was declared on June 12, which is now celebrated as Independence Day. In September a constitutional congress met in Malolos, Bulacan, drawing up a fundamental law derived from European and Latin American precedents. A government was formed on the basis of that constitution in January 1899, with Aguinaldo as president.

Meanwhile, U.S. troops had landed and, with important Filipino help, forced the capitulation in August 1898 of the Spanish commander in Manila. The Americans, however, would not let Filipino forces enter the city. It was soon apparent to Aguinaldo and his advisers that earlier expressions of sympathy for Filipino independence by Dewey and U.S. consular officials had little significance. They felt betrayed.

U.S. commissioners to the peace negotiations in Paris had been instructed to demand the cession of the Philippines to the U.S.; such cession was confirmed with the signing of the peace treaty on December 10, 1898. Ratification followed in the U.S. Senate in February 1899, but with only one vote more than the required two-thirds. Arguments of "manifest destiny" could not overwhelm a determined anti-imperialist minority.

War against the U.S.

By the time the treaty was ratified, hostilities had already broken out between U.S. and Filipino forces. Since Filipino leaders did not recognize U.S. sovereignty over the islands and U.S. commanders gave no weight to Filipino claims of independence, the conflict was inevitable. It took two years of counterinsurgency warfare—marked by several unfortunate similarities with the later U.S. military involvement in Vietnam—plus some wise conciliatory moves in the political arena to break the back of the nationalist resistance. Aguinaldo was captured in March 1901 and shortly thereafter appealed to his countrymen to accept U.S. rule.

The Filipino revolutionary movement had had two goals, national and social. The first goal, independence, though realized briefly, was frustrated by the U.S. decision to establish a new empire. The goal of fundamental social change, manifest in the nationalization of friar lands by the Malolos Republic, was ultimately frustrated by the power and resilience of entrenched institutions. Share tenants who had rallied to Aguinaldo's cause, partly for economic reasons, merely exchanged one landlord for another. In any case, the proclamation of a republic in 1898 had marked the Filipinos as the first Asian people to try to throw off European colonialism.

The period of U.S. influence. The juxtaposition of U.S. democracy and imperial rule over a subject people was sufficiently jarring to most Americans that, from the beginning, the training of Filipinos for self-government and ultimate independence—the Malolos Republic was conveniently ignored—was an essential rationalization for U.S. hegemony in the islands. Differences between Democratic and Republican policies focussed on the speed with which self-government should be extended and the date on which independence should be granted.

In 1899 President McKinley sent Cornell University president Jacob G. Schurman at the head of a five-man fact-finding commission to the Philippines. Schurman reported back that Filipinos wanted ultimate independence, but this had no immediate impact on policy. McKinley sent out the Second Philippine Commission in 1900, under William Howard Taft; by July 1901 it had established civil government.

In 1907 the Philippine Commission, which had been acting as both legislature and governor general's Cabinet, became the upper house of a bicameral body. The new 80-member Assembly was directly elected by a somewhat restricted electorate from single-member districts, making it the first elective legislative body in Southeast Asia. When Governor General Harrison appointed a Filipino majority to the Commission in 1913, the U.S. voice in the legislative process was further reduced.

Francis B. Harrison was the only governor general appointed by a Democratic president in the first 35 years of U.S. rule. He had been sent by Woodrow Wilson with specific instructions to prepare the Philippines for ultimate independence, a goal that he enthusiastically supported. During Harrison's term, a Democratic Congress in Washington hastened to fulfill long-standing campaign promises to the same end. The Jones Act, passed in 1916, would have fixed a definite date for the granting of independence if the Senate had had its way, but the House prevented such a move. In its final form the Act merely stated that it was the "purpose of the people of the U.S." to recognize Philippine independence "as soon as a stable government can be established therein." Its greater importance was as a milestone in the development of Philippine autonomy. Under Jones Act provisions, the Commission was abolished and was replaced by a 24-man Senate, almost wholly elected. The electorate was expanded to include all literate males.

Some substantial restrictions on Philippine autonomy remained, however. Defense and foreign affairs remained exclusive U.S. prerogatives. United States direction of Philippine domestic affairs was exercised primarily through the governor general and the executive branch of insular government. There was little more than one decade of thoroughly U.S. administration in the islands, however—too short a time in which to establish lasting patterns. Whereas Americans formed 51 percent of the civil service in 1903, they were only 29 percent in 1913 and 6 percent in 1923. By 1916 Filipino dominance in both the legislative and judicial branches of government also served to restrict the U.S. executive and administrative roles.

By 1925 the only American left in the governor general's Cabinet was the secretary of public instruction, who was also the lieutenant governor general. This is one

indication of the high priority given education in U.S. policy. In the first few years of U.S. rule, hundreds of schoolteachers came from the United States. But Filipino teachers were trained so rapidly that by 1927 they constituted nearly 99 percent of the 26,200 in public schools. The school population expanded 500 percent in a generation; education consumed one-half of governmental expenditures at all levels, and educational opportunity in the Philippines was greater than in any other colony in Asia.

U.S. educational policy

As a consequence of this pedagogical explosion, literacy doubled to nearly 50 percent in the 1930s, and educated Filipinos acquired a common language and a linguistic key to Western civilization. By 1939, 27 percent of the population could speak English, a larger percentage than for any of the native dialects. Perhaps more important was the new avenue of upward social mobility that education offered. In educational policy was found the only successful U.S. effort to establish a socio-cultural basis for political democracy.

United States attempts to create equality of economic opportunity were more modest and less successful. In a predominantly agricultural country the pattern of landownership is crucial. The trend toward greater concentration, which began in the 19th century, continued during the American period, despite some legal barriers. Vast U.S.-owned plantations were forestalled, but legal restrictions had little effect on those politically well-connected Filipinos who were intent on amassing fortunes. The percentage of farmers under share tenancy doubled between 1900 and 1935; and the frustration of the tenants erupted in three small rebellions in central Luzon during the 1920s and 1930s.

Nor was U.S. trade policy conducive to the diffusion of economic power. From 1909 the Payne-Aldrich Tariff Act allowed free entry of Philippine products into the U.S. market; at the same time U.S. products, mostly manufactured, were exempted from tariff in the Philippines. The free flow of U.S. imports was a powerful deterrent to industrial growth. Export agriculture, especially sugar, prospered in the protected U.S. market. Owners of mills and large plantations profited most, thus reinforcing the political dominance of the landed elite.

United States preparation of the Philippines for democratic self-government suffered from an inherent contradiction, perhaps not recognized at the time. Transferring governmental responsibility to those capable of undertaking it was not consistent with building a social and economic base for political democracy. Self-government meant, of necessity, assumption of power by those Filipinos who were already in positions of leadership in society. But these men came for the most part from the landed elite; preservation of their political and economic position was incompatible with equalization of opportunity. Even the expansion of an educated middle class did not necessarily result in a transformation of the pattern of power. Most middle class aspirants for political leadership adjusted to the values and the practices of the existing power elite.

Formation of political parties

Filipino leaders quickly and skillfully utilized the opportunities for self-government that the Americans opened to them. The Filipino political genius was best reflected in an extralegal institution—the political party. The first party, the Partido Federal, was U.S.-backed and stressed cooperation with the overlords, even to the point of statehood for the Philippines. But, when openly nationalist appeals were allowed in the 1907 election, the Partido Nacionalista, advocating independence, won overwhelmingly. The Federalists survived with a new name, Progressives, and a new platform, ultimate independence after social reform. But neither the Progressives nor their successors in the 1920s, the Democrats, ever gained more than one-third of the seats in the legislature. The Nacionalista Party under the leadership of Sergio Osmeña and Manuel Quezon dominated Philippine politics from 1907 until independence.

More significant than the competition between the Nacionalistas and their opposition was the continuing rivalry between Osmeña and Quezon. In fact, understanding this personality conflict provides more insight into the realities of prewar Philippine politics than any examination of policy or ideology.

In 1933 the U.S. Congress passed the Hare-Hawes-Cutting Act, which set a date for Philippine independence. The Act was a fulfillment of the vague pledge in the Jones Act; it was also responsive to the demands of a series of "independence missions" sent to Washington by the Philippine legislature. But this unprecedented transfer of sovereignty was decided upon in the dark days of the Great Depression—and with the help of some incongruous allies. The Depression had caused U.S. farm interests to look desperately for relief, and those who suffered real or imaginary hurt from the competition of Philippine products sought to exclude those products. They had already failed in a direct attempt to amend the tariff on Philippine imports but found that the respectable cloak of the advocacy of independence increased the effectiveness of their efforts. Tied to independence was the end of free entry into U.S. markets of Philippine sugar, coconut oil, rope, and other less important items. That these economic interests were able to accomplish what they did is partly explainable by the fact that their political clout was great compared to that of the small group of U.S. traders and investors in the Philippines.

The Philippine legislature rejected the Hare-Hawes-Cutting Act, apparently as a result of the Osmeña-Quezon feud, much to the displeasure of American officialdom. But, when Quezon came to Washington the following year to work for a new bill, the same alliance of forces in the U.S. Congress obliged by producing the almost identical Tydings-McDuffie Act. It was endorsed by Quezon and accepted with alacrity by the Manila legislature. It provided for a ten-year commonwealth during which the U.S. would retain jurisdiction over defense and foreign affairs. Filipinos were to draft their own constitution, subject to the approval of the president.

A constitutional convention was quickly elected, a constitution framed (which bears a strong resemblance to its U.S. model), and the document approved by plebiscite and by President Roosevelt. The last governor general, Frank Murphy, became the first high commissioner, with more of a diplomatic than a governing role. The commonwealth was inaugurated on November 15, 1935. The Nacionalista Party patched up its internal quarrels and nominated Quezon for president and Osmeña for vice president. They were elected overwhelmingly.

The Commonwealth

The commonwealth period was intended to be devoted to preparation for economic and political independence and perfection of democratic institutions. But even before the tragic events of 1941 the transition did not run smooth.

World War II. Japanese aggression in China prompted much attention to military preparedness. Nearly one-fourth of the national budget was devoted to defense. Gen. Douglas MacArthur, retiring as army chief of staff in Washington, was called by President Quezon to direct plans and preparations. Meanwhile, agrarian unrest festered, and leftist political activity grew. Quezon pushed significant reform legislation through the National Assembly, but implementation was feeble, despite the rapid accumulation of power in his hands.

The Japanese attack of December 8, 1941, came at a time when the U.S. military buildup had hardly begun. Before Christmas, Manila was declared an "open city," while Quezon and Osmeña were evacuated to MacArthur's headquarters on Corregidor. Despite a desire, at one point, to return to Manila in order to surrender, Quezon was persuaded to leave the Philippines in March on a U.S. submarine; he was never to return. Osmeña went too. Fil-American forces, under Gen. Jonathan M. Wainwright, surrendered in May. An Executive Commission, made up of more than 30 members of the old Filipino political elite, had been cooperating with Japanese military authorities in Manila since January.

The Executive Commission lasted until September 1943, when it was superseded by an "independent Philippine Republic." The president, chosen by the Japanese, was José Laurel, former associate justice of the Common-

The "independent Philippine Republic" under the Japanese

wealth Supreme Court and the only Filipino to hold an honorary degree from Tokyo Imperial University. Laurel was among a small group of Filipinos who had been longtime admirers of Japanese culture and were sympathetic to certain aspects of Japan's authoritarian politics. The scope of collaboration was much wider than this group, however. Over half of the Commonwealth Senate and more than one-third of the House served at one time in the Japanese-sponsored regime. Many Filipinos collaborated out of fear of retaliation, others in prospect of personal gain, and still others out of a patriotic desire to act as a buffer between the Japanese military and the Filipino people. Yet collaboration with Japan was neither as willing nor as widespread as elsewhere in Southeast Asia.

Nor should cooperation with the Japanese military by segments of the elite be equated with mass support for Japanese policies. Even before the fall of Bataan there were guerrilla units formed throughout the Philippines, some by U.S. officers but mostly by Filipinos. Most guerrilla units were middle class led and enthusiastically pro-U.S.; in central Luzon, however, a major force was the Hukbalahap, which, under Communist leadership, capitalized on earlier agrarian unrest. Though in a number of instances collaborators were secretly assisting guerrillas, many guerrillas in the hills were bitter against those who appeared to benefit from the Japanese occupation. The differences between the two groups became an important ingredient in early postwar politics. But the collaborators proved too influential to be punished, as some demanded.

The republic. Soon after the U.S. landings in Leyte in October 1944, civil government was returned to the Commonwealth, at least in name. Unlike the reactions elsewhere in Southeast Asia, Filipinos called the U.S. return liberation. Sergio Osmeña, who had become president on the death of Quezon in 1944, had few resources to deal with the problems at hand. Osmeña's role was complicated by the fact that he was not close to MacArthur; in fact, MacArthur chose to lionize a leading collaborator, who had also been in contact with U.S. military intelligence, Manuel Roxas. Other top figures in the Japanese puppet regime were jailed. By the time the Commonwealth Congress was reconvened, Senate president Roxas clearly had become MacArthur's candidate for president. He had the big money behind him, and his election was crucial for those collaborators who feared prosecution. Since Osmeña would not bow out, Roxas declared his candidacy in January 1946 and was nominated by the "liberal wing" of the Nacionalista Party. Thus was born the Philippines' second major political party, the Liberals. Osmeña, though he had the advantages of incumbency, was old and tired and did not fully use the political tools he possessed. In April, Roxas was elected by a narrow margin. The following month he was inaugurated as the last chief executive of the commonwealth and on July 4 became the first president of the Republic of the Philippines. In this first postwar election a defeated incumbent had quietly handed over the reins of power, thus strengthening constitutionalism in the Philippines.

Roxas, as expected, amnestied all Japanese collaborators except for those "small fry" who may also have committed felonies. In the 1949 election campaign there was an attempt to raise the collaboration issue against José Laurel, the Nacionalista presidential candidate, but it was not effective. In the fluidity of Philippine politics, "guerrillas" and "collaborators" were by that time to be found on both sides of all political fences.

The Philippines had gained independence in the "ashes of victory." Intense fighting, especially around Manila in the last days of the Japanese retreat, had nearly destroyed the capital. The economy generally was in shambles. Rehabilitation aid was obviously needed, and President Roxas was willing to accept some onerous conditions placed implicitly and explicitly by the U.S. Congress in order to get it. The conditions were such, in fact, that many Asians questioned Philippine "independence." The Bell Act in the U.S. extended free trade with the Philippines for eight years, to be followed by 20 years of gradually increasing tariffs. The U.S. demanded and received a 99-year lease on a number of Philippine military and naval bases in which U.S. authorities had virtually territorial rights. And finally, as a specific requirement for release of U.S. war-damage payments, the Philippines had to amend its constitution to give U.S. citizens equal rights with Filipinos in the exploitation of natural resources—the so-called Parity Amendment.

Philippine–U.S. relations. The changing character of Philippine–U.S. relations is a major theme in postwar Philippine history. The trend in recent years has been toward weakening of the link, achieved partly by a diversification of Philippine external ties and partly by more articulate anti-Americanism. Economic nationalism, though first directed against the local Chinese community's dominance of retail trade, by the 1950s was focussed on the special status of U.S. firms. Agitation against "parity" in the 1960s, however, has brought a commitment from both governments that it will not be continued beyond 1974, when the present agreement expires. Tariffs on U.S. goods—90 percent of the standard levy by 1968—and aggressive Japanese salesmanship allowed Japan to come abreast of the U.S., the Philippines' major trading partner, in 1970.

At independence the military ties with the United States were as strong as the economic. Manila was the site for the launching of the Southeast Asia Treaty Organization (SEATO) in 1954. The Philippines sent combat troops to Korea and engineers to South Vietnam. But the Military Bases Agreement of 1947 has been the greatest single cause of friction in U.S.–Philippine relations, triggering demonstrations in the front of the U.S. Embassy on several occasions. In the early 1970s, demands for the total withdrawal of the U.S. military by the 1980s had become increasingly vocal.

Reorientation of Philippine foreign relations in the 1960s took place in the midst of growing attention to the character of national identity. The easy acceptance of the idea of the Philippines as an Asian outpost of Christianity, characteristic of the early postwar period, has been increasingly rejected. Perhaps even more disturbing to Filipinos than questions raised by Asian neighbours about the nature of their sovereignty were doubts about their cultural identity. Asian ethnic origins were strong stimuli to the development of an Asian cultural identity. Artists, musicians, and writers have gone to pre-Spanish themes to find new inspiration. Historians, painfully aware that the nation's past has largely been chronicled by foreigners, have set about rewriting history. More important is the increasing trend toward seeking cultural identity through the national language. Along with nationalistic demands for the wider use of Pilipino (the national language, essentially Tagalog, the native tongue of Manila and its surrounding provinces) have come the deterioration of the quality of English teaching and the rapid expansion of Tagalog movies, magazines, and radio programs.

Given the size of the U.S. emotional and ideological investment in training for democratic self-government, the fate of the Philippines' democratic institutions is an important facet of Philippine–U.S. relations. For Filipinos themselves, who have adopted and adapted Western democracy with enthusiasm, it is a matter inextricably linked to the ability to maintain social order. Some observers have characterized the Philippines in the 1970s as "stable anarchy." Certainly, respect for law in the urban areas and among many rural officials is at a low ebb. Many laws are never enforced, and those that are enforced usually are implemented in a fashion reflecting day-to-day power relationships, not the intent of the legislation. In this context, it is not surprising that student violence—unknown in the Philippines before the late 1960s—and a now endemic agrarian violence have grown.

The "stable anarchy" of the 1970s

Social order did break down completely in many parts of the Philippines in late 1944 and early 1945, as two foreign armies fought their ways from island to island. In the aftermath of war, accumulated social injustice and diligent Communist organizing throughout central Luzon produced a rebellion that peaked in 1951; before the emer-

gence of new leadership in the military, under the then secretary of defense Ramon Magsaysay, and prior to certain steps toward social justice, the rebellion had been on the verge of catching fire around the Philippines. But for more than a decade thereafter, the noticeable trends were toward less violence and greater respect for democratic processes. Then, in the 1960s, deterioration set in.

Modernization proceeds: television networks, new factories, better roads, more automobiles and transistor radios, even more schoolhouses. But population growth is so great—at over 3 percent, one of the highest in the world—and the distribution of wealth so unequal that the average Filipino has seen no significant improvement in living standards in years. Meanwhile, foreign debts rise, and nationalism grows more strident. Christian–Muslim conflict, especially over land, has long simmered; in 1970 it escalated, and there was even talk of secession among Moro leaders. Communist-led insurgency in Luzon is expanding to new areas. Though the textbooks teach and the politicians espouse the right of the common man to choose his leaders and the leaders' duty to serve the nation, most observers claim that fraud, coercion, and lavish expenditure in the 1969 election were unprecedented. In September 1972 Pres. Ferdinand E. Marcos, charging a Communist plot to overthrow the government, placed the Philippines under martial law.

BIBLIOGRAPHY. The earliest centuries of recorded Philippine history are best analyzed in JOHN L. PHELAN, *The Hispanization of the Philippines* (1959), while massive documentation is provided in EMMA HELEN BLAIR and JAMES A. ROBERTSON, *The Philippine Islands, 1493–1898*, 55 vol. (1903–09). An important facet of 19th-century Philippines is described in EDGAR WICKBERG, *The Chinese in Philippine Life, 1850–1898* (1965). The noted Filipino historian, TEODORO A. AGONCILLO, covers the Philippine revolution in two volumes, *The Revolt of the Masses: The Story of Bonifacio and the Katipunan* (1956) and *Malolos: The Crisis of the Republic* (1960). Two major works on the American period are by a former governor general, W. CAMERON FORBES, *The Philippine Islands*, rev. ed. (1945), and a former vice governor general, J. RALSTON HAYDEN, *The Philippines: A Study in National Development* (1942), the more scholarly work. The Philippine independence movement and the Japanese threat to it is the focus of THEODORE FRIEND, *Between Two Empires: The Ordeal of the Philippines, 1929–1946* (1965). DAVID J. STEINBERG, *Philippine Collaboration in World War II* (1967), handles a sensitive subject with both frankness and tact. Post-World War II Philippine history has not yet been treated, however, in an authoritative scholarly volume.

(D.O.D.W.)

Philo of Alexandria

Philo of Alexandria, a Jewish contemporary of Jesus and the Apostle Paul and a key figure in understanding the spiritual crisis of the 1st century of the Christian Era, has a twofold importance for the history of thought. First, he occupies a unique position in the history of philosophy as the first to attempt, although unsystematically, to synthesize revealed faith and philosophic reason. And, second, he has an important place in the history of religion because he was regarded by Christians as a forerunner of Christian theology, especially in his view of the Logos as an intermediary between God and man. In a sense he may be regarded as the founder of medieval philosophy—Jewish, Arabic, and, above all, Christian.

Philo's family

Life. Little is known of the life of Philo. Josephus, the historian of the Jews who also lived in the 1st century, says that Philo's family surpassed all others in the nobility of its lineage. His father had apparently played a prominent role in Palestine before moving to Alexandria. Philo's brother Alexander Lysimachus, who was a general tax administrator in charge of customs in Alexandria, was the richest man in the city and indeed must have been one of the richest men in the Hellenistic world, because Josephus says that he gave a huge loan to the wife of the Jewish king Agrippa I and that he contributed the gold and silver with which nine huge gates of the Temple in Jerusalem were overlaid. Alexander was also extremely influential in Roman imperial circles, being an old friend of the emperor Claudius and having acted as guardian for the Emperor's mother.

Philo himself was born between 15 and 10 BCE ("Before the Common Era"). The community of Alexandria, to judge from the language of the Jewish papyri and inscriptions, had for nearly three centuries been almost exclusively Greek-speaking and indeed regarded the Septuagint (the 3rd-century-BCE translation of the Hebrew Bible into Greek) as divinely inspired. During the century and a half before Philo's birth, Alexandria had been the home of a number of Jewish writers whose works exist now only in fragments. These men were often influenced by the Greek culture in which they lived and wrote apologies for Judaism.

The Alexandrian Jews were eager to enroll their children of secondary school age in Greek gymnasiums, institutions with religious associations dedicated to the liberal arts and athletics; in them, Jews were certainly called upon to make compromises with their traditions. It may be assumed that Philo was a product of such an education: he mentions a wide range of Greek writers, especially the epic and dramatic poets; he was intimately acquainted with the techniques of the Greek rhetorical schools; and he praises the gymnasium. Philo's education, like that which he ascribes to Moses, most probably consisted of arithmetic, geometry, astronomy, harmonics, philosophy, grammar, rhetoric, and logic.

Like the cultured Greeks of his day, Philo often attended the theatre, though it had distinctly religious connotations, and he noted the different effects of the same music on various members of the audience and the enthusiasm of the audience for a tragedy of Euripides. He was a keen observer of boxing contests and attended chariot races as well. He also mentions the frequency with which he attended costly suppers with their lavish entertainment.

Jewish education

Philo says nothing of his own Jewish education. The only mention of Jewish education in his work indicates how relatively weak it must have been, because he speaks only of Jewish schools that met on the Sabbath for lectures on ethics. That he was far from the Palestinian Hellenizers and that he regarded himself as an observant Jew is clear, however, from his statement that one should not omit the observance of any of the Jewish customs that have been divinely ordained. Philo is critical both of those who took the Bible too literally and thus encountered theological difficulties, particularly anthropomorphisms (*i.e.*, describing God in terms of human characteristics), and those who went to excesses in their allegorical interpretation of the laws, with the resulting conclusion, anticipating Paul's antinomianism, that because the ceremonial laws were only a parable, they need no longer be obeyed. Philo says nothing of his own religious practices, except that he made a festival pilgrimage to Jerusalem, though he nowhere indicates whether he made more than one such visit.

In the eyes of the Palestinian rabbis the Alexandrian Jews were particularly known for their cleverness in posing puzzles and for their sharp replies. As the largest repository of Jewish law apart from the Talmud (compendium of Jewish law, lore, and legend) before the Middle Ages, Philo's work is of special importance to those who wish to discern the relationship of Palestine and the Diaspora (*i.e.*, Jewish communities living outside Palestine) in the realm of law *(halakah)* and ritual observance. Philo's exposition of the law may represent either an academic discussion giving an ideal description of Jewish law or the actual practice in the Jewish courts in Egypt. On the whole, Philo is in accord with the prevailing Palestinian point of view; nonetheless he differs from it in numerous details and is often dependent upon Greek and Roman law.

That Philo experienced some sort of identity crisis is indicated by a passage in his *On the Special Laws*. In this work, he describes his longing to escape from worldly cares to the contemplative life, his joy at having succeeded in doing so (perhaps with the Egyptian Jewish ascetic sect of the Therapeutae described in his treatise *On the Contemplative Life*), and his renewed pain at being forced once again to participate in civic turmoil. Philo appears to have been dissatisfied with his life in the bustling metropolis of Alexandria: He praises the Essenes—a Jewish sect

who lived in monastic communities in the Dead Sea area —for avoiding large cities because of the iniquities that had become inverterate among city dwellers, for living an agricultural life, and for disdaining wealth.

Embassy to Caligula

The one identifiable event in Philo's life occurred in the year 39 or 40, when, after a pogrom against the Jews in Alexandria, he headed an embassy to the emperor Caligula asking him to reassert Jewish rights granted by the Ptolemies (rulers of Egypt) and confirmed by the emperor Augustus. Philo was prepared to answer the charge of disloyalty levelled against the Jews by the notorious anti-Semite Apion, a Greek grammarian, when the Emperor cut him short. Thereupon Philo told his fellow delegates not to be discouraged because God would punish Caligula, who, shortly thereafter, was indeed assassinated.

Works. Philo's genuine works may be classified into three groups.

1. Scriptural essays and homilies based on specific verses or topics of the Pentateuch (the first five books of the Bible), especially Genesis. The most important of the 25 extant treatises in this group are *Allegories of the Laws*, a commentary on Genesis, and *On the Special Laws*, an exposition of the laws in the Pentateuch.

2. General philosophical and religious essays. These include *That Every Good Man Is Free*, proving the Stoic paradox that only the wise man is free; *On the Eternity of the World*, perhaps not genuine, proving, particularly in opposition to the Stoics, that the world is uncreated and indestructible; *On Providence*, extant in Armenian, a dialogue between Philo, who argues that God is providential in his concern for the world, and Alexander, presumably Philo's nephew Tiberius Julius Alexander, who raises doubts; and *On Alexander*, extant in Armenian, concerning the irrational souls of animals.

3. Essays on contemporary subjects. These include *On the Contemplative Life*, a eulogy of the Therapeutae sect; the fragmentary *Hypothetica* ("Suppositions"), actually a defense of the Jews against anti-Semitic charges to which Josephus' treatise *Against Apion* bears many similarities; *Against Flaccus*, on the crimes of Aulus Avillius Flaccus, the Roman governor of Egypt, against the Alexandrian Jews and on his punishment; and *On the Embassy to Gaius*, an attack on the Emperor Caligula (*i.e.*, Gaius) for his hostility toward the Alexandrian Jews and an account of the unsuccessful embassy to the Emperor headed by Philo.

A number of works ascribed to Philo are almost certainly spurious. Most important of these is *Biblical Antiquities*, an imaginative reconstruction of Jewish history from Adam to the death of Saul, the first king of Israel.

Philo's works are rambling, having little sense of form; repetitious; artificially rhetorical; and almost devoid of a sense of humour. His style is generally involved, allusive, strongly tinged with mysticism, and often obscure; this may be a result of a deliberate attempt on his part to discourage all but the initiated few.

Some of Philo's works are directed to Gentiles. Thus, the work *On the Life of Moses* contains much apologetic material. Such statements as that the philosophers Heracleitus and Zeno derived their theories from the Bible are obviously apologetic. Most of Philo's essays, however, are intended for sophisticated intellectual Jews who had found difficulties in accepting the biblical text literally. For them Philo developed the method of allegorical interpretation, used several centuries earlier by Greek scholars, especially the Stoics, in the interpretation of Homer, as well as by the Hellenistic Jewish philosopher Aristobulus a century before Philo. The fact that, according to Jewish tradition, the Bible need not be interpreted literally likewise facilitated Philo's choice of this method.

In his exposition of Scripture, Philo relied exclusively on the Septuagint translation, though his quotations do not always agree with the extant text of the Septuagint. The belief that the Septuagint was itself divinely inspired made it apparently unnecessary for him to refer to the original Hebrew. A debate has raged concerning whether or not Philo knew Hebrew; but his gross errors in etymologies of Hebrew names suggest that whatever slight knowledge of Hebrew he may have had was useless.

Thought. The key influences on Philo's philosophy were Plato, Aristotle, the Neo-Pythagoreans, the Cynics, and the Stoics. Philo's basic philosophic outlook is Platonic, so much so that Jerome and other Church Fathers quote the apparently widespread saying: "Either Plato philonizes or Philo platonizes." Philo's reverence for Plato, particularly for the *Symposium* and the *Timaeus*, is such that he never took open issue with him, as he did with the Stoics and other philosophers. But Philo is hardly a plagiarist; he made modifications in Plato's theories. To Aristotle he was indebted primarily in matters of cosmology and ethics. To the Neo-Pythagoreans, who had grown in importance during the century before Philo, he was particularly indebted for his views on the mystic significance of numbers, especially the number seven, and the scheme of a peculiar, self-disciplined way of life as a preparation for immortality. The Cynics, with their diatribes, influenced him in the form of his sermons. Though Philo more often employed the terminology of the Stoics than that of any other school, he was critical of their thoughts.

Influences on his thought

In the past, scholars attempted to diminish Philo's importance as a theological thinker and to present him merely as a preacher, but in the mid-20th century H.A. Wolfson, an American scholar, demonstrated Philo's originality as a thinker. In particular, Philo was the first to show the difference between the knowability of God's existence and the unknowability of his essence. Again, in his view of God, Philo was original in insisting on an individual Providence able to suspend the laws of nature in contrast to the prevailing Greek philosophical view of a universal Providence who is himself subject to the unchanging laws of nature. As a Creator, God made use of assistants: hence the plural "Let us make man" in Genesis, chapter 1. Philo did not reject the Platonic view of a preexistent matter but insisted that this matter too was created. Similarly, Philo reconciled his Jewish theology with Plato's theory of Ideas in an original way: he posited the Ideas as God's eternal thoughts, which God then created as real beings before he created the world.

Philo saw the cosmos as a great chain of being presided over by the Logos, a term going back to pre-Socratic philosophy, which is the mediator between God and the world, though at one point he identifies the Logos as a second God. Philo departed from Plato principally in using the term Logos for the Idea of Ideas and for the Ideas as a whole and in his statement that the Logos is the place of the intelligible world. In anticipation of Christian doctrine he called the Logos the first-begotten Son of God, the man of God, the image of God, and second to God.

His mysticism

Philo was also novel in his exposition of the mystic love of God that God has implanted in man and through which man becomes Godlike. According to some scholars, Philo used the terminology of the pagan religions and mystery cults, including the term *enthousiasmos* ("having God within one"), merely because it was part of the common speech of the day; but there is nothing inherently contradictory in Judaism in the combination of mysticism and legalism in the same thinker. The influence of the mystic notions of Platonism, especially of the *Symposium*, and of the popular mystery cults on Philo's attempt to present Judaism as the one true mystery is hardly superficial; indeed, Philo is a major source of knowledge of the doctrines of these mystery cults, notably that of rebirth. Perhaps, through his mystic presentation of Judaism, Philo hoped to enable Judaism in the Diaspora to compete with the mystery religions in its proselyting efforts, as well as in its attempts to hold on to its adherents. That he was essentially in the mainstream of Judaism, however, is indicated by his respect for the literal interpretation of the Bible, his denunciation of the extreme allegorists, and his failure to mention any specific rites of initiation for proselytes, as well as the lack of evidence that he was himself a devotee of a particular mystery cult.

The purpose of what Philo called mystic "sober intoxication" was to lead one out of the material into the eternal world. Like Plato, Philo regarded the body as the prison house of the soul, and in his dualism of body and

soul, as in his description of the flight from the self, the contrast between God and the world, and the yearning for a direct experience of God, he anticipated much of Gnosticism, a dualistic religion that became important in the 2nd century CE. But unlike all the Greek philosophers, with the exception of the Epicureans, who believed in limited freedom of will, Philo held that man is completely free to act against all the laws of his own nature.

Ethical and political theories

In his ethical theory Philo described two virtues, under the heading of justice, that are otherwise unknown in Greek philosophic literature—religious faith and humanity. Again, for him repentance was a virtue, whereas for other Greek philosophers it was a weakness. Perfect happiness comes, however, not through men's own efforts to achieve virtue but only through the grace of God.

In his political theory Philo often said that the best form of government is democracy; but for him democracy was far from mob rule, which he denounced as the worst of polities, perhaps because he saw the Alexandrian mob in action. For Philo democracy meant not a particular form of government but due order under any form of government in which all men are equal before the law. From this point of view, the Mosaic constitution, which embodies the best elements of all forms of government, is the ideal. Indeed, the ultimate goal of history is that the whole world be a single state under a democratic constitution.

Influence. Whether Philo was an original philosopher with a unified and consistent system or a mere preacher of sermons and whether his thought is a collateral branch of Pharisaic Judaism or a Hellenistic syncretism deeply influenced by the mystery cults, no one can dispute his influence. For the Jews his borrowings from the pagans and his philosophical and theological views apparently were considered too dangerous; and, except in the works of Josephus, his influence is not obvious until the Middle Ages, and then only indirectly. It is not until the 16th century that an explicit reference to Philo is found in Jewish literature, in the Hebrew work *Meor 'Enayim* of the Italian rabbi Azariah dei Rossi. The real importance of Philo lies in his influence on Neoplatonists, particularly his views on the mystic path to the knowledge of God, and, above all, on Christians. For the Christians his work provided a precedent for the synthesis of the Jewish and Greco-Roman traditions, and his view of the Logos provided a philosophical buttress for the doctrine of Jesus as mediator between God and man.

An indication of the high regard that the Christian tradition felt toward him is seen in legends that he met St. Peter and St. John. Although it is debatable whether Philo influenced the Gospel according to John, Paul's letters and, in particular, the Letter to the Hebrews show a strong similarity to Philo in both ideas and language. Philo was particularly influential on the Church Fathers Clement of Alexandria, Origen, and Ambrose. Further indications of Christian esteem for him are seen in the survival of the great majority of his works and the translation of some of his works into Latin and Armenian in the 4th and 5th centuries.

BIBLIOGRAPHY. H.L. GOODHART and E.R. GOODENOUGH, "A General Bibliography of Philo Judaeus," in Goodenough's *The Politics of Philo Judaeus, Practice and Theory*, pp. 125–348 (1938), an exhaustive classified listing of books and articles down to 1937; L.H. FELDMAN, *Scholarship on Philo and Josephus, 1937–1962* (1963), a classified critical bibliography; *Studia Philonica* (annual), a journal, exclusively devoted to Philo, published by the newly established (1971) Philo Institute at McCormick Theological Seminary, Chicago.

Greek text: LEOPOLD COHN and PAUL WENDLAND (eds.), *Philonis Alexandrini Opera quae supersunt*, 7 vol. (1896–1930), definitive, but lacks the fragments, the treatises extant only in Armenian, and Pseudo-Philo.

English translation: F.H. COLSON, G.H. WHITAKER, and R. MARCUS, *Philo* ("Loeb Classical Library," 10 vol. and 2 supplementary vol., 1929–62), an extremely careful translation, with brief notes and useful analytical introductions to each treatise.

Pseudo-Philo: M.R. JAMES, *The Biblical Antiquities of Philo*, prolegomenon by L.H. FELDMAN (1917, reprinted 1971), a clear, accurate translation. The prolegomenon is particularly concerned with the relationship to rabbinic and Hellenistic Jewish literature and with supplements to James's commentary.

Introductory works and surveys: E.R. GOODENOUGH, *An Introduction to Philo Judaeus*, 2nd ed. rev. (1963), a popular, lively, fair summary of Goodenough's previous work, with helpful suggestions on how to read Philo; H. CHADWICK, "Philo," in A.H. ARMSTRONG (ed.), *The Cambridge History of Later Greek and Early Medieval Philosophy*, pp. 137–157 (1967), the best-balanced, brief, comprehensive survey of Philo's thought.

Major works on Philo: E.R. GOODENOUGH, *By Light, Light: The Mystic Gospel of Hellenistic Judaism* (1935), a challenging, if one-sided, presentation of Philo as a mystic; SAMUEL BELKIN, *Philo and the Oral Law* (1940, reprinted 1968), a learned, occasionally forced, presentation of Philo as thoroughly acquainted with Palestinian Halakha; H.A. WOLFSON, *Philo*, 2 vol. (1947), a truly great, if controversial, work, stressing Philo's relationship to Palestinian Pharisaic Judaism and his importance for the history of philosophy; SAMUEL SANDMEL, *Philo's Place in Judaism* (1956), a well-balanced corrective to Wolfson, consisting of a case study of the account of Abraham in Philo as compared with rabbinic and other contemporary literature.

(L.H.F.)

Philosophy, History of Western

This article pursues three basic purposes: (1) it attempts to provide a synoptic overview of the history of philosophy in the West; (2) it strives to relate philosophic ideas and movements to their historical background and to the cultural situation of their time; and (3) it is specially mindful of philosophy's changing conception of its own definition, its function, and its unique task.

The main divisions of this article are as follows:

I. Nature, involvements, and writing of the history of philosophy
  Philosophy in the Western tradition
  General consideration of the history of philosophy
    The writing of the history of philosophy
    Shifts in the focus and concern of philosophy through history
II. Ancient Greek and Roman philosophy
  The Pre-Socratic philosophers
    Cosmology and the metaphysic of matter
    Epistemology of appearance
    Metaphysic of number
    Anthropology and relativism
  The seminal thinkers of Greek philosophy
    Socrates
    Plato
    Aristotle
  Hellenistic and Roman philosophy
    Stoics
    Epicureans
    Skeptics
    Neo-Pythagoreans and Neoplatonists
III. Medieval philosophy
  Early medieval philosophy
  Transition to Scholasticism
  The age of the schoolmen
  Philosophy in the late Middle Ages
IV. Modern philosophy
  The Renaissance and early-modern period
    Dominant strands of Renaissance philosophy
    Rise of Empiricism and Rationalism
    Literary forms and sociological conditions
  The Enlightenment
    Classical British Empiricism and its basic tasks
    Nonepistemological movements in the Enlightenment
    Critical examination of reason in Kant
    Literary forms
  The 19th century
    German Idealism of Fichte, Schelling, and Hegel
    Positivism and social theory in Comte, Mill, and Marx
    Independent and irrationalist movements
  The 20th century
    Individual philosophies of Bergson, Dewey, and Whitehead
    Marxist thought
    Analytic philosophy
    Continental philosophy
    Concluding comments

I. Nature, involvements, and writing of the history of philosophy

PHILOSOPHY IN THE WESTERN TRADITION

Various conceptions of what philosophy is

It is a paradox faced by all of those who attempt to write the history of philosophy that the "philosophy" whose history they write probably would not have been defined exactly alike by any two of the major figures whom they judge it fitting to include in their account. For throughout its long and varied history in the West, "philosophy" has meant many different things. Some of these have been a search for the wisdom of life (the meaning closest to the Greek words from which the term is derived); an attempt to understand the universe as a whole; an examination of man's moral responsibilities and his social obligations; an effort to fathom the divine intentions and man's place with reference to them; an effort to ground the enterprise of natural science; a rigorous examination of the origin, extent, and validity of men's ideas; an exploration of the place of will or consciousness in the universe; an examination of the values of truth, goodness, and beauty; an effort to codify the rules of human thought in order to promote rationality and the extension of clear thinking. Even these do not exhaust the meanings that have been attached to the philosophic enterprise, but they give some idea of its extreme complexity and manysidedness.

It is difficult to determine whether any common element can be found within this diversity and whether any core meaning can be discovered for philosophy that could serve as a universal and all-inclusive definition. But a first attempt in this direction might be to define philosophy either as "a reflection upon the varieties of human experience" or as "the rational, methodical, and systematic consideration of those topics that are of greatest concern to man." Vague and indefinite as such definitions are, they do suggest two important facts about philosophizing: (1) that it is a reflective, or meditative, activity and (2) that it has no explicitly designated subject matter of its own but is a method or type of mental operation (like science or like history) that can take any area or subject matter or type of experience as its object. Thus, although there are a few single-term divisions of philosophy of long standing—such as logic, ethics, epistemology (the theory of knowledge), or metaphysics (theory of the nature of Being)—its divisions are probably best expressed by phrases that contain the preposition "of"—such as philosophy of nature or philosophy of mind or philosophy of law or philosophy of art.

Part of what makes it difficult to find a consensus among philosophers about the definition of their discipline is precisely that they have frequently come to it from different fields, with different interests and concerns, and that they therefore have different areas of experience upon which they find it especially necessary or meaningful to reflect. Thomas Aquinas (a Dominican friar of the 13th century), George Berkeley (a bishop of the Irish Church in the 18th century), and Søren Kierkegaard (a Danish divinity student in the 19th century) all saw philosophy as a means to assert the truths of religion and to dispel the Materialistic or Rationalistic errors that, in their opinion, had led to its decline. Pythagoras in ancient south Italy, René Descartes in the late Renaissance, and Bertrand Russell in the 20th century have been primarily mathematicians whose views of the universe and of human knowledge have been vastly influenced by the concept of number and by the method of deductive thinking. Some philosophers, such as Plato or the British philosophers Thomas Hobbes and John Stuart Mill, have been obsessed by problems of political arrangement and social living, so that whatever else they have done in philosophy has been stimulated by a desire to understand and, ultimately, to change the social and political behaviour of men. And still others—such as the Milesians (the first philosophers of Greece); Francis Bacon, an Elizabethan philosopher; and, in the 20th century, Alfred North Whitehead, a process metaphysician—have begun with an interest in the physical composition of the natural world, so that their philosophies resemble more closely the generalizations of physical science than those of religion or sociology.

Characteristic oppositions in philosophy

The history of Western philosophy reveals in detail the concentrated activity of a multitude of serious and able men reflecting upon, reasoning about, and considering deeply the nature of their experience. But throughout this manifold diversity certain characteristic oppositions habitually recur, such as the division between monists, dualists, and pluralists in metaphysics; between Materialists and Idealists in cosmological theory; between Nominalists and Realists in the theory of signification; between Rationalists and Empiricists in the theory of knowledge; between Utilitarians, self-realizationists, and proponents of duty in moral theory; and between partisans of logic and partisans of emotion in the search for a responsible guide to the wisdom of life.

Many of these fundamental oppositions among philosophers will be treated in the article that follows. But if any single opposition is taken as central throughout the history of Western philosophy at every level and in every field, it is probably that between the critical and the speculative impulses. These two divergent motivations tend to express themselves in two divergent methods: that of analysis and that of synthesis. Plato's *Republic* is an example of the second. The *Principia Ethica* (1903) of G.E. Moore, a founder of linguistic philosophy, is an example of the first. Beginning with a simple question about justice, *The Republic* in its discursiveness slowly but progressively brings more and more areas into the discussion: first ethics, then politics, then educational theory, then theory of knowledge, and finally metaphysics. Starting with one rather specific question, Plato finally managed to make his discussion as broad as the world. *Principia Ethica* does just the opposite. Beginning with a rather general question—What is good?—it progressively breaks up this question into a whole series of subordinate questions, analyzing meanings ever more minutely, growing narrower and narrower but always with the utmost modesty and sincerity, striving for increasing simplicity and exactitude.

The analytic, or critical, impulse treats any subject matter or topic by concentrating upon the part, by taking it apart in the service of clarity and precision. It was essentially the method of Aristotle and of Peter Abelard, a medieval Scholastic; of David Hume, a Scottish Skeptic, and Rudolf Carnap, a contemporary semantic Positivist; and of Russell and Moore. The synthetic or speculative impulse operates by seeking to comprehend the whole, by putting it all together in the service of unity and completeness. It is essentially the method of Parmenides, a Pre-Socratic monist, and of Plato; of St. Thomas and of Benedict de Spinoza, a modern Jewish Rationalist; of G.W.F. Hegel, a German Idealist, and of Whitehead. Throughout philosophy's history, each of the two traditions has made its insistent claim.

There is one philosophical tradition—that of Positivism—which sees philosophy as originating in the obscure mists of religion and coming finally to rest in the pure sunshine of scientific clarity. This represents a necessary progress, since Positivism considers it a scandal when philosophers speak a language that is not accessible to "verification"; it holds that bold and adventuresome philosophical speculation is at best mere self-indulgence, a passing state occurring when philosophic problems are raised prematurely—that is, at a time when philosophy does not possess the logical means to solve them.

Though Positivism represents a partisan view that it is not necessary to hold, it does express indirectly a basic truth—that the philosophic enterprise has always hovered uncertainly between the lure of religious seriousness and that of scientific exactitude. In the earliest philosophers of Greece, it is impossible to separate ideas of divinity and the human soul from ideas about the mystery of being and the genesis of material change, and in the Middle Ages philosophy was acknowledged to be "the handmaiden of theology." But the increased secularization of modern culture has largely reversed this trend, and the Enlightenment emphasis upon the separation of nature from its divine creator has increasingly placed philosophic resources at the disposal of those interested in creating a philosophy of science.

Yet philosophy's continuing search for philosophic truth leads it to hope, but at the same time to profoundly doubt, that its problems are objectively solvable. With respect to a total description of Being or a definitive account of the nature of values, only individual solutions now seem possible; and the optimistic hope for objective answers that secure universal agreement must be given up.

Comparison with art: living classics

In this respect, philosophy seems less like science than like art and the philosopher more like an artist than a scientist, for his philosophic solutions bear the stamp of his own personality, and his choice of arguments reveals as much about himself as his chosen problem. As a work of art is a portion of the world seen through a temperament, so a philosophical system is almost a vision of the world subjectively assembled. Plato and Descartes, Immanuel Kant, the pivotal figure of modern philosophy, and John Dewey, a U.S. Pragmatist, have given to their systems many of the quaint trappings of their own personalities.

But if philosophy is not true in the same sense as science, it is not false in the same sense either; and this gives to the history of philosophy a living significance, which the history of science does not enjoy. In science, the present confronts the past as truth confronts error; thus, for science, the past, even when important at all, is important only out of historical interest. In philosophy it is different. Philosophical systems are never definitively proved false; they are simply discarded or put aside for future use. And this means that the history of philosophy consists not simply of dead museum pieces but of ever-living classics—comprising a permanent repository of ideas, doctrines, and arguments and a continuing source of philosophical inspiration and suggestiveness to those who philosophize in any succeeding age. It is for this reason that any attempt to separate philosophizing from the history of philosophy is both a provincial act and an unnecessary impoverishment of its rich natural resources.

GENERAL CONSIDERATION OF THE HISTORY OF PHILOSOPHY

The writing of the history of philosophy. The writing of the history of philosophy is itself controlled by a series of cultural habits or conventions.

Ways of ordering the history. The ensuing article on the history of Western philosophy is divided into three sections—ancient, medieval, and modern—and this division is so pervasive today that it is difficult to remember that the threefold distinction is only as old as the end of the 17th century. This distinction—first employed in the writing of European history proper by Georg Horn of Leiden in his *Arca Noae: Sive Historia Imperiorum et Regnorum a Condito Orbe ad Nostra Tempora* (1666) and a generation later by Christophorus Cellarius, a German historian, in 1696—slowly spread to historical writing in all fields and was given definitive influence in philosophical writing through the series of lectures on *Philosophiegeschichte* ("History of Philosophy") that Hegel delivered first at Jena, then at Heidelberg, and finally at Berlin between 1805 and 1830. Since Hegel, it has been taken for granted as standard practice, although a host of cultural assumptions is implied by its use.

Stress on ideas versus stress on men

Treatment of the total field of the history of philosophy has been traditionally subject to two types of ordering, according to whether it was conceived (1) as primarily a history of ideas or (2) as a history of the intellectual products of men. In the first ordering, certain ideas, or concepts, are viewed as archetypal (such as matter or mind or doubt); and the condensations occurring within the flow of thought tend to consist of basic types or schools. Such ordering has characterized works such as Friedrich Lange's *Geschichte des Materialismus . . .* (1866; Eng. trans., *The History of Materialism*, 3rd ed., 1925), A.C. Ewing's compilation *The Idealist Tradition: From Berkeley to Blanshard* (1957), or Richard H. Popkin's *History of Scepticism from Erasmus to Descartes* (1960). In the second type of ordering, the historian, impressed by the producers of ideas as much as by the ideas themselves—that is, with philosophers as agents—reviews the succession of great philosophic personalities in their rational achievement. Such ordering has produced the more customary histories such as Émile Bréhier's *Histoire de la philosophie* (1926–32), Bertrand Russell's *History of Western Philosophy* (1945), and Karl Jaspers' *Die grossen Philosophen* (1957; Eng. trans., 1962–66).

These two different types of ordering depend for their validity upon an appeal to two different principles about the nature of ideas, but their incidental use may also be influenced by social or cultural factors. Thus, the biographers and compilers of late antiquity (among them, Plutarch, Sextus Empiricus, Philostratus, and Clement of Alexandria), impressed by the religious pluralism of the age in which they lived, thought of philosophers, too, as falling into different sects and wrote histories of the Sophists, the Skeptics, the Epicureans, and other such schools; whereas almost 2,000 years later, Hegel—living in a period of romantic historiography dominated by the concept of the great man in history—deliberately described the history of philosophy as "a succession of noble minds, a gallery of heroes of thought."

As between these two ordering principles, the article below will be eclectic (as has come to be the custom), devoting chief attention to outstanding major figures, while joining more minor figures, wherever possible, into the schools or tendencies that they exemplify.

Factors in writing history of philosophy. The type of ordering suggested above also has some relationship to the more general problems of method in the writing of the history of philosophy. Here there are at least three factors that must be taken into account: (1) the historian must understand how (at least in part) any philosopher's doctrines depend upon those of his predecessors; (2) he must understand that a man's philosophy occurs at a certain point of time in history and, thus, how it expresses the effects of certain social and cultural circumstances; and (3) he must understand how in part it stems from the philosopher's own personality and situation in life. This is only to say that the history of philosophy, to be at all comprehensive and adequate, must deal with the mutual interplay of ideas, of agents, and of cultural contexts.

Logical, sociological, and individual factors

The first factor may be called logical because it is the intellectual response that a given philosophy makes to the doctrines of its forerunners in which the very problems taken as central have often been given by the current climate of controversy. Thus, many of the details of Aristotle's ethical, political, and metaphysical systems arise in arguments directed against statements and principles of Plato; much of *An Essay Concerning Human Understanding* (1690) by John Locke, an initiator of the Enlightenment, is directed against current Cartesian presuppositions; and the *Nouveaux essais sur l'entendement humain* (written 1704; Eng. trans., *New Essays Concerning Human Understanding*, 1916) by Gottfried Wilhelm Leibniz, a broadly learned Rationalist, is, in turn, specifically directed against Locke.

The second factor may be called sociological because it considers philosophy, at least in part, as a direct form of social expression, arising at a certain moment in history, dated and marked by the peculiar problems and crises of the society in which it flourishes. From this perspective, the philosophy of Plato may be viewed as the response of an aristocratic elitism to the immediate threat of democracy and the levelling of values in 5th-century Athens—its social theory and even its metaphysics servicing the thrust toward an aristocratic restoration in the Greek world. Thus, the philosophy of St. Thomas Aquinas may be viewed as an effort toward doctrinal clarification in support of the institution of the medieval Roman Church, as the Saint spent his life obediently fulfilling the philosophic tasks set for him by his superiors in the church and the Dominican order. Thus, the philosophy of Kant, with all of its technical vocabulary and rigid systematization, may be viewed as an expression of the new professionalism in philosophy, a clear product of the rebirth of the German universities during the 18th-century Enlightenment.

The third factor may be called biographical or individual because, with Hegel, it recognizes that philosophies

are generally produced by men of unusual or independent personality whose systems usually bear the mark of their creators. And what is meant here by the individuality of the philosopher lies less in the facts of his biography (such as the wealth or poverty, the married state or bachelorhood that he shares with other men) than in the essential form and style of his philosophizing. The cool intensity of Spinoza's geometric search for wisdom, the unswerving (if opaque) discursiveness of Hegel's quest for completeness or totality, the relentless and minute analytic search for distinctions and shades of meaning that marks Moore's master passion ("to be accurate—to get everything exactly right"), these qualities mark the philosophical writings of Spinoza, Hegel, and Moore with an unmistakably individual and original character.

Shifts in the focus and concern of philosophy through history. Any adequate treatment of individual figures in the history of philosophy tries to utilize this threefold division of logical, sociological, and individual factors; but in a synoptic view of the history of philosphy in the West, one is particularly aware of the various shifts of focus and concern that philosophy has sustained and, indeed, of the often profound differences in the way that it defines itself or visualizes its task from age to age or from generation to generation.

Philosophy among the Greeks slowly emerged out of religious awe into wonder about the principles and elements of the natural world. But as the Greek populations more and more left the land to become concentrated in their cities, interest shifted from nature to social living; questions of law and convention and civic values became paramount. Cosmological speculation partly gave way to moral and political theorizing, and the preliminary and somewhat fragmentary questionings of Socrates and the Sophists turned into the great positive constructions of Plato and Aristotle. With the political and social fragmentation of the succeeding centuries, however, philosophizing once again shifted from the norm of civic involvement to problems of salvation and survival in a chaotic world.

Medieval and modern history

The dawn of Christianity brought to philosophy new tasks. Augustine, the philosophic bishop of Hippo, and the Church Fathers used such resources of the Greek tradition as remained (chiefly Platonism) to deal with problems of the creation, of faith and reason, and of truth. New translations in the 12th century made much of Aristotle available and prepared the way for the great theological constructions of the 13th century, chiefly those of the Scholastic philosophers Bonaventure, Albertus Magnus, Aquinas, Roger Bacon, and Duns Scotus. The end of the Middle Ages saw a new flowering of the opposite tendencies in the nominalist William of Ockham and the mystic Meister Eckehart.

The waning of the Middle Ages became the Renaissance. Universalism was replaced by nationalism. Philosophy became secularized. The great new fact was that of the mystery and immensity of the natural world. The best philosophic minds of the 17th century turned to the task of exploring the foundations of physical science, and the symbol of their success—the great system of Newton's physics—turned the philosophers of the Enlightenment to epistemology and to the examination of the human mind that had produced so brilliant a scientific creation. The 19th century, a time of great philosophical diversity, discovered the irrational and in so doing prepared the way for the oppositions between Analysis and Phenomenology and between Positivism and Existentialism that characterize the present situation in philosophy.

Although the foregoing capsule presentation of the history of philosophy in the West follows a strict chronology, it does not do justice to the constant occurrence and recurrence of dominant strands in the history of thought. It would also be possible to write the philosophic history of the Middle Ages simply by noting the complicated occurrence of Platonic and Aristotelian doctrines, of the Renaissance according to the reappearance of Greek Materialism, Stoicism, and Skepticism, and of the 18th century in terms of the competing claims of Rationalist and Empiricist principles. Thus, chronology and the interweaving of philosophic systems cooperate in a completed history of philosophy. (A.W.L.)

II. Ancient Greek and Roman philosophy

THE PRE-SOCRATIC PHILOSOPHERS

Cosmology and the metaphysic of matter. Because the earliest Greek philosophers focussed their attention upon the origin and nature of the physical world, they are often called cosmologists or naturalists. Though monistic views prevailed at first, they were soon followed by several pluralistic theories.

Monistic cosmologies. There is a consensus, dating back at least as far as the 4th century BC and continuing to the present, that the first Greek philosopher was Thales of Miletus, who flourished in the first half of the 6th century BC. At that time, the word philosopher (lover of wisdom) had not yet been coined. Thales was counted, however, among the Seven Wise Men (Sophoi), whose name derives from a term that then designated inventiveness and practical wisdom rather than speculative insight. Thales showed these qualities by trying to give the mathematical knowledge that he derived from the Babylonians a more exact foundation and by using it for the solution of practical problems—such as the determination of the distance of a ship as seen from the shore or of the height of the Pyramids. Though he was also credited with predicting an eclipse of the Sun, it is likely that he merely gave a natural explanation of one on the basis of Babylonian astronomical knowledge.

Thales of Miletus

The reason that Thales was considered to be the first Greek philosopher is that he was the first to give a purely natural explanation of the origin of the world, free from all mythological ingredients, declaring that everything had come out of water—an explanation based on the discovery of fossil sea animals far inland. His tendency (and that of his immediate successors) to give nonmythological explanations of the origin of the world was undoubtedly prompted by the fact that all of them lived on the coast of Asia Minor surrounded by a number of nations with civilizations much farther advanced than that of the Greeks, whose mythological explanations differed greatly both among themselves and from those of the Greeks. It appeared necessary, therefore, to make a fresh start on the basis of what a person could observe and figure out by looking at the world as it presented itself. This procedure naturally resulted in a tendency to make sweeping generalizations on the basis of rather restricted, but carefully checked, observations.

Thales' disciple and successor, Anaximander of Miletus (flourished mid-6th century), tried to give a more elaborate account of the origin and development of the ordered world (the cosmos). According to him, it developed out of the *apeiron*, something both infinite and indefinite (without distinguishable qualities). Within this *apeiron* something arose to produce the opposites of hot and cold. These at once began to struggle with each other and produced the cosmos. The cold (and wet) partly dried up (becoming solid earth), partly remained (as water), and—by means of the hot—partly evaporated (becoming air and mist), its evaporating part (by expansion) splitting up the hot into fiery rings, which surround the whole cosmos. Because these rings are enveloped by mist, however, there remain only certain breathing holes that are visible to men, appearing to them as Sun, Moon and stars. Anaximander was the first to realize that upward and downward are not absolute directions, but that downward means toward the middle of the Earth and upward away from it, so that the Earth had no need to be supported (as Thales had believed) by anything. Starting from Thales' observations, Anaximander tried to reconstruct the development of life in more detail. Life, being closely bound up with moisture, originated in the sea. All land animals, he held, are descendants of sea animals; man, specifically, is the descendant of an oviparous sea animal of the family of the cetaceans, the *Galeus levis*, that carries its young in a pouch after they have come out of the egg or later lets them return to it until they can fend for themselves. Gradually, however, the moisture will be partly dried up, partly evaporated, until in the end

Anaximander of Miletus

all things will have returned into the undifferentiated *apeiron*, "in order to pay the penalty for their injustice" —that of having struggled against one another.

Anaximenes of Miletus

Anaximander's successor, Anaximenes of Miletus (flourished second half of the 6th century), taught that air was the origin of all things. His position was for a long time thought to have been a step backward because, like Thales, he placed a special kind of matter at the beginning of the development of the world. But this criticism missed the point. Neither Thales nor Anaximander appear to have specified the way in which the other things arose out of the water or *apeiron*. Anaximenes, however, declared that the other types of matter arose out of air by condensation and rarefaction. In this way, what to Thales had been merely a beginning became a fundamental principle that remained essentially the same through all of its transmutations. Thus, the term *archē*, which originally simply meant "beginning," acquired the new meaning of "principle," a term that henceforth played an enormous role in philosophy down to the present. This concept of a principle that remains the same through many transmutations is, furthermore, the presupposition of the idea that nothing can come out of nothing and that all of the comings to be and passings away that men observe are nothing but transmutations of something that essentially remains the same eternally. In this way it also lies at the bottom of all of the conservation laws—those of the conservation of matter, of force, and of energy—that have been basic in the development of physics. Though Anaximenes of course did not realize all of the implications of his idea, its importance can hardly be exaggerated.

The first three Greek philosophers have often been called hylozoists because they seemed to believe in a kind of living matter. But this is hardly an adequate characterization. It is, rather, characteristic of them that they did not clearly distinguish between kinds of matter, forces, and qualities nor between physical and emotional qualities. The same entity is sometimes called fire and sometimes the hot. Heat appears sometimes as a force and sometimes as a quality, and again there is no clear distinction between warm and cold as physical qualities and the warmth of love and the cold of hate. To realize these ambiguities is important to an understanding of certain later developments in Greek philosophy.

Xenophanes and Parmenides

Xenophanes of Colophon (b. *c.* 560 BC), rhapsode and philosophic thinker, who emigrated from Asia Minor to Elea in southern Italy, was the first to bring out more clearly what was implied in Anaximenes' philosophy. He criticized the popular notions of the gods, saying that men made their gods in their own image. But, more importantly, he argued that there could be only one God, the ruler of the universe, who must be eternal.

For, being the strongest of all beings, he could not have come out of something less strong, nor could he be overcome or superseded by something else, because nothing could arise that is stronger than the strongest. The argument clearly rested on the axiom that nothing can come out of nothing and that nothing that is can really vanish. The axiom was made more explicit and carried to its extreme consequences by Parmenides of Elea (flourished first half of the 5th century BC), the founder of the so-called school of Eleaticism (*q.v.*), of whom Xenophanes has been regarded as the teacher and forerunner. In a philosophical poem Parmenides insisted that "what is" cannot have come into being and cannot pass away because it would have to have come out of nothing or to become nothing, whereas nothing by its very nature does not exist. There can be no motion either; for it would have to be a motion into something that is—which is not possible since it would be blocked—or a motion into something that is not—which is equally impossible since what is not does not exist. Hence everything is solid immobile being. The familiar world, in which things move around, come into being, and pass away, is a world of mere belief (*doxa*). In a second part of the poem, however, Parmenides tried to give an analytical account of this world of belief, showing that it rested on constant distinctions between what is believed to be positive—*i.e.*, to have real being, such as light and warmth—and what is negative—*i.e.*, the absence of positive being, such as darkness and cold.

It is significant that Heracleitus of Ephesus, a contemporary of Parmenides, whose philosophy was later considered to be the very opposite of Parmenides' philosophy of immobile being, came, in some fragments of his work, near to what Parmenides tried to show: the positive and the negative, he said, are merely different views of the same thing; death and life, day and night, or light and darkness are really one.

Pluralistic cosmologies. Parmenides had an enormous influence on the further development of philosophy. Most of the philosophers of the following two generations tried to find a way to reconcile his thesis that nothing comes into being nor passes away with the evidence presented to men by their senses. Empedocles of Acragas (flourished mid-5th century) declared that there are four material elements (he called them roots of everything) and two forces, love and hate, that did not come into being and would never pass away or increase or diminish. But the elements are constantly mixed with one another by love and again separated by hate. Thus, through mixture and decomposition composite things come into being and pass away. Because he conceived of love and hate as blind forces, Empedocles had to explain how living beings could arise. This he achieved by means of a somewhat crude anticipation of the theory of the survival of the fittest. In the process of mixture and decomposition the limbs and parts of various animals would be formed by chance. But they could not survive. Only when by chance they had come together in such a way that they were able to support and reproduce themselves would they survive. It was in this way that the various species were produced and continued to exist.

Empedocles and Anaxagoras

Anaxagoras of Clazomenae, a 5th-century pluralist, believed that because nothing can really come into being, everything must be contained in everything but in the form of infinitely small parts. In the beginning all of these particles had been mixed in an even mixture, in which nothing could be distinguished, much like the indefinite *apeiron* of Anaximander. But then *nous*, or intelligence, began at one point to set these particles into a whirling motion foreseeing that in this way they would become separated from one another and then recombine in the most various ways so as to produce gradually the world in which men live. In contrast to the forces assumed by Empedocles, the *nous* of Anaxagoras is not blind but foresees and intends the production of the cosmos including living and intelligent beings; but it does not interfere with the process after having started the whirling motion —a strange combination of a mechanical and a nonmechanical explanation of the world.

Leucippus and Democritus

By far of greatest importance for the later development of philosophy and physical science was an attempt by the Atomists Leucippus (flourished mid-5th century) and (in the following generation) Democritus to solve the Parmenidean problem. Leucippus found the solution in the assumption that, contrary to Parmenides' argument, the nothing *does* in a way exist, viz., as empty space. There are then, however, only two fundamental principles of the physical world, empty space and filled space—the latter consisting of the atoms, which, in contrast to those of modern physics, are real atoms; *i.e.*, they are absolutely indivisible because nothing can penetrate to split them. On these foundations, laid by Leucippus, Democritus appears to have built a whole system, aiming at a complete explanation of the varied phenomena of the visible world by means of an analysis of its atomic structure. This system begins with elementary physical problems such as that of why a hard body can be lighter than a softer one. The explanation is: that although the heavier body contains more atoms, they are equally distributed and of round shape; the lighter body, however, has fewer atoms, most of which have hooks by which they form rigid gratings. The system ends with educational and ethical questions. A sound and cheerful man, useful to his fellowmen, is literally well composed. Although destructive passions involve violent long-distance atomic motions, education

can help to contain them, creating a better composure. Civilization is produced by the necessities and needs of life, which compel man to work and to make inventions. When life becomes too easy because all needs are met, there is a danger that civilization will decay as men become unruly and negligent.

Epistemology of appearance. All of the post-Parmenidean philosophers, like Parmenides himself, presupposed that the real world is different from the one that men perceive. Thus, the problems of epistemology, or theory of knowledge, arose. According to Anaxagoras, everything is contained in everything. But this is not what people perceive. He solved this problem, however, by assuming that, if there is a much greater amount of one kind of particle in a thing than of all other kinds the latter are not perceived at all. The observation was then made that sometimes different persons or kinds of animals have different perceptions of the same things. He explained this phenomenon by assuming that like is perceived by like. If, therefore, in the sense organ of someone, there is less of one kind of stuff than of another, he will perceive it less keenly than the latter. This reasoning was also used to explain why some animals see better by night and others by daylight. According to Democritus, the atoms have no sensual qualities such as tastes, smells, or colours at all. Thus, he tried to reduce all of them to tactile qualities (explaining a bright white colour, for instance, as sharp atoms hitting the eye like needles), and he made a most elaborate attempt to reconstruct the atomic structure of things on the basis of their apparent sensual qualities.

Zeno of Elea

Also of very great importance in the history of epistemology was Zeno of Elea (flourished mid-5th century), a younger friend of Parmenides. Parmenides had, of course, been severely criticized because of the strange consequences of his doctrine that in reality there is no motion and no plurality either because there is just one solid being. To support him, however, Zeno tried to show that the assumption that there is motion and plurality leads to consequences that are no less strange. This he did by means of his famous paradoxes, saying that the flying arrow rests since it can neither move in the place in which it is nor in a place in which it is not and that Achilles cannot outrun a turtle because when he has reached its starting point, the turtle will have moved to a further point, and so on ad infinitum—that, in fact, he cannot even start running, for, before traversing the stretch to the starting point of the turtle, he will have to traverse half of it and again half of that and so on ad infinitum. All of these paradoxes are derived from the problem of the continuum. Although they have often been dismissed as logical nonsense, many attempts have also been made to dispose of them by means of mathematical theorems, such as the theory of convergent series or the theory of sets. In the end, however, the difficulties inherent in his arguments have always come back with a vengeance, for the human mind is so constructed that it has two ways of looking at a continuum that are not quite reconcilable.

Metaphysic of number. All of the philosophies mentioned so far are in various ways historically akin to one another. Toward the end of the 6th century, however, there arose quite independently another kind of philosophy, which only later entered into interrelation with the developments just mentioned: the philosophy of Pythagoras of Samos.

Pythagoras of Samos

Pythagoras travelled extensively in the East and in Egypt and, after his return to Samos, emigrated to southern Italy because of his dislike of the tyranny of Polycrates. At Croton and Metapontum he founded a philosophical society with strict rules and soon gained considerable political influence. He appears to have brought his doctrine of the transmigration of souls from the East. Much more important for the history of philosophy and science, however, was his doctrine that "all things are numbers," which means that the essences and structures of all things can be determined by finding the numerical relations contained in them. Originally, this, too, was a very broad generalization made on the basis of comparatively few observations: for instance, that the same harmonies can be produced with different instruments—strings, pipes, disks, etc.—by means of the same numerical ratios—1:2, 2:3, 3:4—in one-dimensional extensions; the observation that certain regularities exist in the movements of the celestial bodies; and the discovery that the form of a triangle is determined by the ratio of the lengths of its sides. But, because the followers of Pythagoras tried to apply their principle everywhere with the greatest of accuracy, one of them—Hippasus of Metapontum—around 450 BC made one of the most fundamental discoveries of the entire history of science, that of incommensurability, viz., that the quantitative relation between the side and diagonal of such simple figures as the square and the regular pentagon cannot be expressed as a ratio of integers, however great. At first sight this discovery seemed to destroy the very basis of the Pythagorean philosophy, and the school thus split into two sections, one of which engaged in rather abstruse numerical speculations while the other succeeded in overcoming the difficulty by ingenious mathematical inventions and laid the foundations of all quantitative science. Pythagorean philosophy also exerted a great influence on the development of Plato's thought in his later years.

The speculations described so far comprise in many ways the most important part of the history of Greek philosophy because all of the most fundamental problems of Western philosophy turned up here for the first time and one finds here the formation of a great many concepts that have continued to dominate Western philosophy and science down to the present day.

Anthropology and relativism. In the middle of the 5th century BC, Greek thinking took a somewhat different turn through the advent of the Sophists, a designation derived from the verb *sophizesthai*, "making a profession of being inventive and clever," for the Sophists—in contrast to the philosophers mentioned so far—asked money for their instruction. Philosophically, they were, in a way, the leaders of a rebellion against the preceding development, which more and more had resulted in the belief that the real world is quite different from the phenomenal world. "What is the sense of such speculations?" they said, since men do not live in these so-called real worlds. This is the meaning of a pronouncement of Protagoras of Abdera (flourished around the mid-5th century): "Man is the measure of all things, of those which are that they are and of those which are not that they are not." For man, the world is what it appears to him to be, not something else; and, though he meant man in general, he illustrated it by pointing out that even in regard to an individual man it makes no sense to tell him that it is really warm when he is shivering with cold, because for him it is cold—for him, the cold exists, is there.

The Sophists

His younger contemporary Gorgias of Leontini, famous for his treatise on the art of oratory, made fun of the philosophers in a book *Peri tou mē ontos ē peri physeōs* ("On that which is not, or on Nature"), in which—referring to the "truly existing world," also called "the nature of things"—he tried to prove (1) that nothing exists, (2) that if something existed, man could have no knowledge of it, and (3) that if nevertheless somebody knew it, he could not communicate his knowledge to others.

The Sophists were not only skeptical of what had by then become a philosophical tradition but also of other traditions. On the basis of the observation that different nations have different rules of conduct even in regard to things considered most sacred—such as the relations between the sexes, marriage, and burial—they concluded that most rules of conduct are rather conventional. What is really important is to be successful in life and to gain influence on others. This they promised to teach. Gorgias was proud of the fact that, having no knowledge of medicine, he was more successful in persuading a patient to undergo a necessary operation than his brother, a physician, who knew when an operation was necessary. The older Sophists, however, were far from openly preaching immoralism. They nevertheless gradually came under suspicion because of their sly ways of arguing. One of the later Sophists, however, Thrasymachus of Chalcedon (late 5th century), was bold enough to declare openly

that "right is what is beneficial for the stronger or better one"; *i.e.*, for the one able to win the power to bend others to his will.

THE SEMINAL THINKERS OF GREEK PHILOSOPHY

Socrates. By many of his contemporaries, Socrates (flourished 5th century BC) was also considered to be a Sophist because of his tricky arguments, though he did not teach for money and his aims were entirely different from theirs. Although there is a late tradition according to which Pythagoras invented the word philosopher, it was certainly through Socrates—who insisted that he possessed no wisdom but was striving for it—that the term came into general use and was later applied to all earlier serious thinkers. In fact, all of the records of his life and activity left by his numerous adherents and disciples indicate that he never tried to teach anything directly. But he constantly engaged in conversations with everybody—old and young, high and low—trying to bring into the open by his questions the inconsistencies in their opinions and actions. But though he never taught directly, his whole activity rested on two unshakable assumptions: (1) the principle never to do wrong nor to participate, even indirectly, in any wrongdoing and (2) the conviction that nobody who really knows what is good and right could act against it. He demonstrated his unshakable adherence to the first principle on various occasions and under different regimes. When, after the Battle of Arginusae, the majority of the Athenian popular assembly demanded death without trial for the admirals, Socrates, who on that day happened to be president of the assembly (an office changing daily), refused to put the proposal to the vote because it was wrong to condemn anyone without a fair trial, though the people threatened him shouting that it would be terrible if the sovereign people could not do as they pleased. When, after the overthrow of democracy, the so-called Thirty Tyrants, who tried to involve everybody in their wrongdoings, ordered him to arrest an innocent citizen whose money they coveted, he simply disobeyed. This he did although at that time such disobedience was still more dangerous than disobeying the sovereign people had been at the time of unrestricted democracy. Likewise, in the time of the democracy he pointed out by his questions the inconsistency of allowing oneself to be swayed by the oratory of a good speaker instead of first inquiring into his capability as a statesman, whereas in private life a sensible citizen would not listen to the oratory of a quack but would try to find the best doctor. When, after the overthrow of democracy, the Thirty Tyrants had many people arbitrarily executed, he asked everybody whether a man was a good shepherd who diminished the number of the sheep instead of increasing it and did not cease doing so when Critias, the leader of the Thirty, warned him to take heed not to diminish the number of the sheep by his own, Socrates', person. But the most fundamental inconsistency that he tried to show up everywhere was that most people by their actions showed that they considered what they found to be good, wonderful, and beautiful in others—such as, for instance, doing right at great danger to oneself—not to be good for themselves, and considered to be good for themselves what they despised and condemned in others. Though all of these stands won him the most fervent admiration of many, especially among the young of all classes, it caused also great resentment among leading politicians, whose inconsistencies were shown up publicly by him and his adherents. Though Socrates had survived unharmed through the regime of the Thirty—partly because it did not last long, partly because he was supported by some close relatives of their leader Critias—it was under the restored democracy that he was accused of impiety and of corrupting the youth and finally condemned to death, largely also in consequence of his intransigent attitude during the trial.

After Socrates' death his influence became a dominating one through the greater part of Greek and Roman philosophy down to the end of antiquity and was more or less noticeable even in all of the rest. Many of his adherents —among them Xenophon, a military man and a historian, and Aeschines of Sphettus, one of those present at his death—tried to preserve his philosophical method by writing Socratic dialogues. Some founded schools or sects that perpetuated themselves over long periods of time, Eucleides of Megara emphasizing the theoretical aspects of Socrates' thought, and Antisthenes stressing the independence of the true philosopher from material wants. The latter, through his disciple Diogenes of Sinope, who carried voluntary poverty to the extreme and emphasized freedom from all conventions, became the founder of the sect of the Cynics. Aristippus of Cyrene, traditional founder of the Cyrenaic school, stressed man's independence from material needs in a somewhat different way, declaring that there is no reason why a philosopher should not enjoy material goods as long as he is completely indifferent to their loss. Though Aristippus renounced his son because he led a dissolute life, the school that he founded (through his daughter and his grandson) was hedonistic, holding pleasure to be the good.

Founding of schools and sects

Plato. By far the most important disciple of Socrates, however, was Plato, a scion of one of the most noble Athenian families, who could trace his ancestry back to the last king of Athens and to Solon, the great social and political reformer.

Life. As a very young man Plato became a fervent admirer of Socrates in spite of the latter's plebeian origin. Contrary to his master, however, who always concerned himself with the attitudes of individuals, he believed in the importance of political institutions. In his early youth he had observed that the Athenian masses, listening to the glorious projects of ambitious politicians, had engaged in foolhardy adventures of conquest, which led in the end to total defeat in the Peloponnesian War. When, in consequence of the disaster, democracy was abolished, Plato at first set great hopes in the new oligarchy—especially since its leader, Critias, was a close relative. But he soon discovered, that—to use his own words—the despised democracy had been gold in comparison with the new terror. When the oligarchy was overthrown and the restored democracy, in 399, adopted a new law code—in fact, a kind of written constitution containing safeguards against rash political decisions—Plato again had considerable hope and was even inclined to view the execution of Socrates as an unfortunate incident rather than a logical consequence of the new regime. It was only some years later, when demagogy appeared to raise its head again, that he "despaired and was forced to say that things would not become better in politics unless the philosophers would become rulers or the rulers philosophers." He wrote a dialogue, the *Gorgias*, violently denouncing political oratory and propaganda and then travelled to southern Italy in order to study political conditions there. Again, however, he found the much vaunted *dolce vita* of the Greeks there, in which the rich lived in luxury exploiting the poor, much worse than the democracy at Athens. But at Syracuse he met a young man, Dion, brother-in-law of the ruling tyrant, Dionysius I, who listened eagerly to his political ideas and promised to work for their realization if any occasion should arise. On his return to Athens, Plato founded the so-called Academy, an institution for the education of philosophers, and in the following years elaborated, besides other dialogues, his great work, *The Republic*, in which he drew the outlines of an ideal state. Because it is the passions and desires of men that cause all disturbances in society, the state must be ruled by an elite who are governed exclusively by reason and supported by a class of warriors entirely obedient to them. Both ruling classes must have no individual possessions and no families and lead an extremely austere life, receiving the necessities of life from the working population, who alone are permitted to own private property. The elite receive a rigid education to fit them for their task. At the death of Dionysius I, Dion induced Plato to come to Syracuse again to try to persuade Dionysius' successor to renounce his power in favour of a realization of Plato's ideals. But the attempt failed, and in his later political works, the *Statesman* and the *Laws*, Plato tried to show that only a god could be entrusted with the absolute powers of the philos-

Visit to Syracuse

opher-rulers of his *Republic*. Human rulers must be controlled by rigid laws, he held—though all laws are inevitably imperfect because life is too varied to be governed adequately by general rules. But the *Laws* still placed strict restrictions on the ownership of property.

Philosophy. In the field of theoretical philosophy, Plato's most influential contribution was undoubtedly his theory of Ideas, which he derived from Socrates' method in the following way: Socrates, in trying to bring out the inconsistencies in his interlocutors' opinions and actions, had often asked what it is that makes men say that a certain thing or action is good or beautiful or pious or brave; and he had asked what people are looking at when they make such statements. Plato sometimes made Socrates ask what is the Eidos or the Idea—*i.e.*, the image —that a person has before him when he calls something "good." A definite answer was never given, however, because no abstract definition would be adequate, the purpose being rather to make the interlocutor aware of the fact that he somehow does look at something undefinable when making such statements.

Theory of Ideas

What was at first simply a way of somehow expressing something that is difficult to express developed into a definite theory of Ideas when Plato made the discovery that something similar could be observed in the field of mathematics. No two things in the visible world are perfectly equal, just as there is nothing that is perfectly good or perfectly beautiful. Yet equality is one of the most fundamental concepts not only in mathematics but also in everyday life—the foundation of all measurement. Hence, like the notion of the good and the beautiful, it appears to come from a different world, a world beyond that of the senses, a world that Plato then called the world of Ideas. All of the things, then, that men perceive with their senses appear to be but very imperfect copies of the eternal Ideas. The most important and fundamental of all, however, is the Idea of the Good. It is "beyond being and knowledge," being the foundation of both. Though this doctrine may seem fantastic to some moderns, it contains an important element of truth. "Being" in this connection does not mean existence, but being something—a man, a lion, or a house—being recognizable by its quality or shape. Without such shapes, which are copies of Ideas, there would be complete chaos, in which nothing would be recognizable. Furthermore, it is impossible to build up any system of knowledge without constantly preferring what is formed and, thus, recognizable to what is not, what is true to what is false, a sound logical conclusion to a logical fallacy, even an elegant scientific demonstration to a clumsy one; *i.e.*, without considering the former as good and the latter as bad. In this sense at least, there is no such thing as value-free science.

Myth of the cave

In the famous myth of the cave in the seventh book of *The Republic*, Plato likened the ordinary person to a man sitting in a cave looking at a wall on which he sees nothing but the shadows of the real things that are behind his back, and he likened the philosopher to a man who has got out in the open and seen the real world of the Ideas. Coming back, he may be less able to distinguish the shades because he has been blinded by the light outside; but he is the only one who knows real reality, and he conducts his life accordingly. In his *Theaetetus* Plato later criticized a sensualist theory of knowledge, anticipating the explanations of the 17th-century English sensualists. In the *Timaeus*, he tried to build up a complete system of physics, partly employing Pythagorean ideas. Most modern Positivists do not take Plato seriously any longer. But one of the greatest physicists of the 20th century, Werner Heisenberg, has insisted that the modern physicist still has to learn a good deal from Plato concerning the foundation of his science.

Aristotle. After Plato's death the Academy continued to exist for many centuries under various heads.

When Plato's nephew, Speusippus, was elected as his successor, his greatest disciple, Aristotle, left to go first to Assus and then to the island of Lesbos, where he met Theophrastus, who became his most gifted disciple. But soon thereafter he was called to the Macedonian court at Pella to become the educator of the crown prince, who was later to become Alexander the Great. After the latter had become the king, Aristotle returned to Athens and opened there a school of his own, whose members became known as the Peripatetics.

Philosophy. Aristotle had become a member of the Academy at the age of 17, in the year 367 (during Plato's absence in Sicily) under the acting chairmanship of Eudoxus of Cnidus, a great mathematician and geographer. It is a very controversial question as to how far Aristotle, during the 20 years of his membership in the Academy, developed a philosophy of his own differing from that of his master. But two things can be considered as certain: (1) that he soon raised certain objections to Plato's theory of the Ideas, for one of the arguments against it attributed to him is discussed in Plato's dialogue *Parmenides*, which Plato must have written soon after his return from Sicily, and (2) that it was during his membership in the Academy that Aristotle began and (to a considerable extent) elaborated his theoretical and formal analysis of the arguments used in various Socratic discussions—an enterprise that (when completed) resulted in the corpus of his works on logic, a new science, which Aristotle himself claimed to have originated and about which (until rather recent times) it used to be said that he completed it in such a way that hardly anything could be added.

Certainly, quite some time before his return to Athens to open a school of his own, Aristotle declared that it is not necessary to assume the existence of a separate realm of transcendent Ideas of which the individual things that men perceive with their senses are but imperfect copies; that the world of perceived things is the real world; and that it is necessary merely to be able to say that something is generally true of certain types or groups of things in order to build up a system of knowledge about them. Thus, it would be wrong to say that, having abandoned the theory of Ideas, Aristotle was left with a completely contingent world. The last chapters of his *Posterior Analytica* show, on the contrary, that he merely replaced Plato's transcendent Ideas with something (*katholou*) corresponding to them that the human mind can grasp in the individual things.

Teleology and perfection

Aristotle retained another very important element of the theory of Ideas in his teleology, or doctrine of purposiveness. According to Plato, individual things are imperfect copies of the perfect Ideas. Aristotle pointed out, however, that all living beings develop from an imperfect state (from the seed, the semen, through the germinating plant, or embryo, to the child and adolescent), to the more perfect state of the fully developed plant or the full grown mature animal or man—after which they again decay and finally die, having reproduced themselves. But not all individuals reach the same degree of relative perfection. Many of them die before reaching it; others are retarded or crippled or maimed in various ways in the process. It is, therefore, of the utmost importance for man to find out what the best conditions are for reaching the most perfect state possible. This is what the gardener tries to do for the plants; but it is even more important for man to do it in regard to himself. The first question, then, is that in which one asks what kind of perfection a human being as human can reach. In answering this question, Aristotle observed that man, being the social animal par excellence, can reach as an individual only some of the perfections possible for man as such. Cats are more or less all alike in their functions; they can fend for themselves. With bees and termites, however, it is different. They are by nature divided into worker bees, drones, and queen bees or worker termites, soldier termites, and queens. With human beings the differentiation of functions is much more subtle and varied. Men can lead satisfactory lives only on the basis of a division of labour and distribution of functions. Some human individuals are born with very great talents and inclinations for special kinds of activity. They will be happy and will make their best possible contribution to the life of the community only if they are permitted to follow this inclination. Others are less one-sidedly gifted and more easily adaptable to different functions. These people can also be

happy in shifting from one activity to another. That this is so is an enormous advantage that the human species has over all other animals because it enables it to adapt to all sorts of circumstances. But the advantage is paid for by the fact that no human individual is able to develop all of the perfections that are possible for the race as a whole.

There is another possible and—in its factual consequences—very real disadvantage; the other animals, tightly enclosed within the limits set by nature, are crippled almost exclusively by external factors; but man, in consequence of the freedom of choice granted to him through the variety of his gifts, can and very often does cripple and harm himself. All human activities are directed toward the end of a good and satisfactory life. But there are many subordinate aims that are sensible ends only as far as they serve a superior end. There is no sense in producing or acquiring more shoes than can possibly be worn. This is self-evident. With regard to money, however, which has become exchangeable against everything, the illusion arises that it is good to accumulate it without limit. By doing so, however, man harms both the community and himself because, concentrating on such a narrow aim, he deprives his soul and spirit of larger and more rewarding experiences. Similarly, an individual especially gifted for large-scale planning needs power to give orders to those capable of executing his plans. Used for such purposes, power is good. But coveted for its own sake it becomes oppressive to those subdued by it and harmful to the oppressor because he thus incurs the hatred of the oppressed. Because of his imperfection, man is not able to engage in serious and fruitful activities without interruption. He needs relaxation and play, or amusement. Because the necessities of life frequently force man to work beyond the limit within which working is pleasant the illusion arises that a life of constant amusement would be the most pleasant and joyful. In reality, nothing is more tedious. This is the true meaning of Aristotle's teleology. It seems to be based entirely on empirical observation. It has nothing to do with a belief in divine providence and is not, as some modern critics believe, at variance with the law of causality. It forms the foundation, however, of Aristotle's ethics and political theory. Aristotle was an avid collector of empirical evidences. He induced his students, for instance, to make collections of the laws and political institutions (and their historical developments) of all known cities and nations in order to find out how they worked and at what points their initiators had been mistaken regarding the way in which they would work. In later times, Aristotle came to be considered (and by many is still considered) a dogmatic philosopher because the results of his inquiries were accepted as absolutely authoritative. In reality, however, he was one of the greatest Empiricists of all times.

Disciples and commentators. After Aristotle's death his immediate disciples carried on the same kind of work, especially in the historical field: Theophrastus wrote a history of philosophy and works on botany and on mineralogy, Eudemus of Rhodes wrote histories of mathematics and of astronomy, Meno a history of medicine, and Dicaearchus of Messene a history of civilization and a book on types of political constitutions. The next two generations of Peripatetics spread out in two different directions: literary history, in the form of histories of types of poetry, epic, tragedy, and comedy, and of biographies of famous writers, and physical science, Straton of Lampsacus creating a new kind of physics based on experiments, and the great astronomer Aristarchus of Samos inventing the heliocentric system. The school then went for some time into eclipse until, in the 1st century AD, after the rediscovery of Aristotle's lecture manuscripts, there arose a great school of commentators on his works, which had an enormous influence on medieval philosophy.

HELLENISTIC AND ROMAN PHILOSOPHY

The period after the death of Aristotle was characterized by the decay of the Greek city-states, which then became pawns in the power game of the Hellenistic kings who succeeded Alexander. Life became troubled and insecure. It was in this environment that two dogmatic philosophical systems came into being, the Stoic and the Epicurean, which were destined to give their adherents something to hold onto and to make them independent of the external world.

Stoics. The Stoic system was created by a Syrian, Zeno of Citium (flourished around the turn of the 3rd century BC), who went to Athens as a merchant but lost his fortune at sea. Zeno was consoled by the Cynic philosopher Crates, who taught him that material possessions were of no importance whatsoever for a man's happiness. He therefore stayed at Athens, heard the lectures of various philosophers, and—after he had elaborated his own philosophy—began to teach in a public hall, the Stoa Poikile (hence the name Stoicism). Zeno of Citium

Zeno's thought comprised, essentially, a dogmatized Socratic philosophy, with added ingredients derived from Heracleitus. The basis of human happiness, he said, is to live "in agreement" (with oneself), a statement which was later replaced by the formula "to live in agreement with nature." The only real good for man is the possession of virtue; everything else (wealth or poverty, health or illness, life or death) is completely indifferent. All virtues are based exclusively on right knowledge—self-control (*sōphrosynē*) being the knowledge of the right choice, fortitude the knowledge of what must be endured and what must not, and justice the right knowledge "in distribution." The passions, which are the cause of all evil, are the result of error in judging what is a real good and what is not. Because it is difficult to see, however, why murder, fraud, and theft should be considered evil if life and possessions are of no value, the doctrine was later modified by making among the "indifferent things" distinctions between "preferable" things, such as having the necessities of life and health; completely indifferent things; and "anti-preferable things," such as lacking the necessities of life or being ill—while insisting still that the happiness of the truly wise man could not be impaired by illness, pain, hunger, or any deprivation of external goods. In the beginning, Zeno also insisted that a man is either completely wise, in which case he would never do anything wrong and would be completely happy, or he is a fool. Later he made the concession, however, that there are men not completely wise but who are progressing toward wisdom. Though the latter might even have true insight, they are not certain that they have it, whereas the truly wise man is also certain of having true insight. The world is governed by divine *logos*—a word originally meaning "word" or "speech," then (with Heracleitus) also a speech that expresses the laws of the universe, and, finally, "reason." This *logos* keeps the world in perfect order. Man can deviate from or rebel against this order, but by doing so he cannot disturb it but can only do harm to himself.

Development of Zeno's philosophy

Zeno's philosophy was further developed by Cleanthes, the second head of the school, and by Chrysippus, its third head. Chrysippus elaborated a new kind of logic, which did not receive much attention, however, outside the Stoic school, until, in recent times (under the name of the "propositional logic"), it has been hailed by some logicians as superior to the "conceptual logic" of Aristotle. In the mid-2nd century BC, Panaetius of Rhodes adapted Stoic philosophy to the needs of the Roman aristocracy (whose members were then governing the known world) and made a great impression on some of the leading men of the time, who tried to follow his moral precepts. In the following century, in the time of the decay of the Roman republic, of civil war, and of slave rebellions, Poseidonius of Apamea, who was also one of the most brilliant historians of all times, taught that the Stoic takes a position above the rest of mankind, looking down on their struggles as on a spectacle. In the periods of the rising monarchy and of its established rule, Stoicism became the religion of the republican opposition, the most famous example of which was the younger Cato, who committed suicide after the victory of Julius Caesar. It was also the religion of Seneca, the educator and (for a long time) the adviser of Nero, who tried to keep Nero

on the path of virtue but failed and finally had to commit suicide on the orders of the Emperor. In spite of the oddities of Zeno's original doctrine, Stoicism gave consolation, composure, and fortitude in times of trouble to many proud men to the end of antiquity and beyond.

Epicurus and Lucretius Carus

Epicureans. The thought of Zeno's contemporary Epicurus also comprised a philosophy of defense in a troubled world. It has been (and still is) considered—in many respects justly—the opposite of Zeno's. Whereas Zeno had proclaimed that the wise man would try to learn from everybody and would always acknowledge his debt to earlier philosophers, Epicurus insisted that everything he taught was the result of his own thinking, though it is obvious that his physical explanation of the universe is a simplification of Democritus' Atomism. And whereas the Stoics had taught that pleasure and pain are of no importance for a man's happiness, Epicurus made pleasure the very essence of a happy life. Moreover, the Stoics from the beginning had acted as advisers of kings and statesmen. Epicurus, on the other hand, lived in the retirement of his famous Garden, cultivating intimate friendships with his adherents but warning against participation in public life. The Stoics believed in divine providence; Epicurus taught that the gods pay no attention whatsoever to human beings. Yet in spite of these contrasts, the two philosophies had some essential factors in common. Though Epicurus made pleasure the criterion of a good life, he was far from advocating a dissolute life and debauchery; he insisted that it was the simple pleasures that made life happy. When in his old age he suffered terrible pains from kidney trouble (or prostatitis?) he asserted that philosophizing and the memory and love of his distant friends made pleasure prevail even in the grips of such pain. Nor was Epicurus an atheist. His Roman admirer, the poet Lucretius Carus (*c.* 95–55 BC), in his poem *De rerum natura* ("On the Nature of Things") praised Epicurus enthusiastically as the liberator of mankind from all religious fears; and Epicurus himself had affirmed that this had been one of the aims of his philosophy. But although he taught that the gods are much too superior to trouble themselves with paying attention to mortals, he said—and, as his language clearly shows, sincerely believed—that it is important for human beings to look at the gods as perfect beings, since only in this way could they approach perfection. It was only in Roman times that people began to misunderstand Epicureanism, holding it to be an atheistic philosophy justifying a dissolute life, so that a man could be called "a swine from the herd of Epicurus." Seneca recognized the true nature of Epicureanism, however, and in his *Epistulae morales* ("Moral Letters") deliberately interspersed through his Stoic exhortations maxims from Epicurus.

Pyrrho of Elis and Sextus Empiricus

Skeptics. There was still another Hellenistic school of philosophy, the Skeptic school initiated by another of Zeno's contemporaries—Pyrrho of Elis—a school that was destined to become of great importance for the preservation of a detailed knowledge of Hellenistic philosophy in general. Pyrrho had come to the conviction that no man can know anything for certain nor ever be certain that the things he perceives with his senses are real and not illusory. He is said to have carried the practical consequences of his conviction so far that, when walking in the streets, he paid no attention to the vehicles and other obstacles, so that his faithful disciples always had to accompany him to see that he came to no harm. Pyrrho's importance for the history of philosophy lies in the fact that one of the later adherents of his doctrine, Sextus Empiricus (2nd–3rd century AD), wrote a large work *Pros dogmatikous* ("Against the Dogmatists"), in which he tried to refute all of the philosophers who had a more positive philosophy and that in so doing he quoted extensively from their works, thus preserving much that would otherwise have been lost. It is a noteworthy fact that the British sensualists of the 18th century and also Immanuel Kant derived most of their knowledge of ancient philosophy from Sextus.

Neo-Pythagoreans and Neoplatonists. All of the philosophical schools and sects of Athens that originated in the 4th century BC continued into late antiquity, most of them until the emperor Justinian I, in AD 529, ordered all of them closed because of their pagan character. Within that whole period of nearly 1,000 years only two new schools were added; even these, however—the Neo-Pythagorean and the Neoplatonic schools—drew their inspiration from early Greek philosophy, though only the latter was of greater importance for the history of philosophy. Neoplatonism began with Ammonius Saccas (first half of the 3rd century AD), who had been brought up as a Christian but had abandoned his religion for the study of Plato and developed his own kind of Platonic philosophy. Because he wrote nothing, his philosophy is known only through his famous disciple, Plotinus. Plotinus, however, did not publish anything either. But his philosophy is known through the publication of notes, discussions, and explanations following his lectures made by his disciple Porphyry, who also wrote a biography of Plotinus.

Plotinus and Iamblichus

The philosophy of Plotinus (and Ammonius) was derived from a study of Plato. Moreover, it used many philosophical terms first coined by Aristotle, and adopted some elements of Stoic philosophy as well. Yet it is essentially a new philosophy, agreeing with the religious and mystical tendencies of its time. Plotinus assumed the existence of several levels of Being, the highest being that of the One, or the Good, which are identical but indescribable and indefinable in human language. The next lower level is that of the *nous*, or pure intellect or reason; the third is that of the soul or souls. There then follows the world perceivable by the senses and, finally, at the lowest level there is matter, which is the cause of all evil. The highest bliss for man is union with the One, or Good, attained by contemplation and purification. That this is not a lasting state attained once for all—like the status of the Stoic wise man, who was supposed never to lose his wisdom again—is shown by the fact that Porphyry, in his *Vita Plotini*, said that Plotinus had experienced this supreme bliss seven times in his life, whereas he, Porphyry, had experienced it only once.

The further history of Neoplatonism is extremely complicated. While Porphyry had emphasized the ethical element in Plotinus' philosophy, his disciple Iamblichus of Chalcis in Syria (died *c.* AD 325), founder of a Syrian branch of the sect, mingled Neoplatonism with Neo-Pythagoreanism, writing on the Pythagorean way of life and on number theory. Above all, he multiplied the levels of being or the emanations from the One, which enabled him to incorporate the traditional Greek gods into his system. Another branch of the school was founded in Pergamum, in western Asia Minor, by his disciple Aedesius. He and his disciple Maximus tried to revive the ancient Greek mystery religions, such as Orphism. All of these developments became of great importance for the attempt made in the 4th century by the emperor Julian to revive paganism. In the following century the Athenian school reached a new high point when Proclus combined ideas of his predecessors into one comprehensive system. When in 529 Justinian closed all of the philosophical schools in Athens, however, a branch continued to exist in Alexandria. The Athenian Neoplatonists found refuge at the court of the Persian king Khosrow, and in 535 they were permitted to return to Athens. But gradually pagan philosophy as such died out, though it continued to exist as an influence in the development of Christian philosophy and theology. (K.v.Fr.)

III. Medieval philosophy

Medieval philosophy designates the philosophical speculation that occurred in western Europe during the Middle Ages; *i.e.*, from the fall of the Roman Empire in the 4th and 5th centuries AD to the Renaissance of the 15th century. Greek philosophy ceased to be creative after Plotinus in the 3rd century AD. A century later, Christian thinkers, such as Ambrose, Victorinus, and Augustine, began to assimilate Neoplatonism into Christian doctrine in order to give a rational interpretation of Christian faith. Thus, medieval philosophy was born of the confluence of Greek (and to a lesser extent of Roman) philosophy and Christianity. Plotinus' philosophy was already deeply religious,

having come under the influence of Near Eastern religion. Medieval philosophy continued this religious orientation, for it remained in close conjunction with Christian faith. Its methods were philosophical and rational, at first those of Plotinus and later those of Aristotle. But it developed within faith as a means of throwing light on the truths and mysteries of faith. Thus, religion and philosophy fruitfully cooperated in the Middle Ages. Philosophy, as the handmaid of theology, made possible a rational understanding of faith. Faith, for its part, inspired Christian thinkers to develop new philosophical ideas, some of which became part of the philosophical heritage of the West.

Confluence of Greek philosophy and Christian faith

Toward the end of the Middle Ages, this helpful interplay of faith and reason started to break down. Philosophy began to be cultivated for its own sake, apart from, and even in contradiction to, Christian religion. This divorce of reason from faith, made definitive by Francis Bacon and René Descartes, marked the birth of modern philosophy.

EARLY MEDIEVAL PHILOSOPHY

The early medieval period, which extended to the 12th century, saw the barbarian invasions of the Roman Empire, the collapse of its civilization, and the gradual building of a new, Christian culture in western Europe. Philosophy in these troubled and darkened times was cultivated by late Roman thinkers, such as Augustine (354–430) and Boethius (*c.* 480–*c.* 525), then by monks such as Anselm (1033–1109). The monasteries became the main centres of learning and education and retained their preeminence until the founding of the cathedral schools and universities.

Augustine. During these centuries philosophy was heavily influenced by Neoplatonism; Stoicism and Aristotelianism played only a minor role. Augustine was awakened to the philosophical life by reading Cicero, but the Neoplatonists most decisively shaped his philosophical methods and ideas. To them he owed his conviction that beyond the world of the senses there is a spiritual, eternal realm of truth that is the object of the human mind and the goal of all man's striving. This truth he identified with the God of Christianity. Man encounters this divine world of truth and beauty not through his senses but by turning inward to his mind, and above his mind to the intelligible light, in which he sees the truth. The Augustinian demonstration of the existence of God coincides with the proof of the existence of necessary, immutable Truth. Augustine considered the truths of both mathematics and ethics to be necessary, immutable, and eternal. These truths cannot come from the world of contingent, changing, and temporal things, nor from the mind itself, which is also contingent, mutable, and temporal. They are due to the illumining presence in man's mind of eternal and immutable Truth, or God. Any doubt that man knows the truth with certainty was dispelled for Augustine by the certitude that even if he is deceived in many cases, man cannot doubt that he exists and knows and loves.

Augustine conceived of man as a composite of two substances, body and soul, of which the soul is by far the superior. Nevertheless, the body is not to be excluded from human nature, and its eventual resurrection from the dead is assured by Christian faith. The soul's immortality is proved by its possession of eternal and unchangeable Truth.

Augustine's *Confessions* (*c.* 400) and *De Trinitate* (400–416; *On the Trinity*) abound with penetrating psychological analyses of knowledge, perception, memory, and love. His *De civitate Dei* (413–426; *The City of God*) presents the whole drama of human history as a progressive movement of humanity, redeemed by God, to its final repose in its Creator.

Boethius. One of the most important channels by which Greek philosophy was transmitted to the Middle Ages was Boethius. He began to translate into Latin all the philosophical works of the Greeks, but his imprisonment and death by order of Theodoric, king of the Ostrogoths, cut short this project. He translated only the logical writings of Porphyry (a 3rd-century Neoplatonist) and Aristotle. These translations and his commentaries on them brought to the thinkers of the Middle Ages the rudiments of Aristotelian logic. They also raised important philosophical questions, such as the nature of universals (terms that can be applied to more than one particular thing). Are universals real or only mental concepts? If real, are they corporeal or incorporeal; if incorporeal, do they exist in the sensible world or apart from it? Medieval philosophers debated at length these and other problems relating to universals. In his logical works Boethius presents the Aristotelian doctrine of universals, that they are only mental abstractions. In his *De consolatione philosophiae* (*c.* 525; *Concerning the Consolation of Philosophy*), however, he adopts the Platonic notion that they are innate ideas and their origin is in the remembering of knowledge in a previous existence. This book was extremely popular and influential in the Middle Ages. It contains not only a Platonic view of knowledge and reality but also a lively treatment of providence, divine foreknowledge, chance, fate, and human happiness.

The question of the nature of universals

Greek Fathers of the Church and Erigena. Another stream from which Greek philosophy, especially Neoplatonic thought, flowed into the Middle Ages was the Greek Fathers of the Church, notably Origen (*c.* 185–*c.* 254), Gregory of Nyssa (*c.* 335–*c.* 394), Nemesius (*c.* 400), Pseudo-Dionysius the Areopagite (*c.* 500), and Maximus the Confessor (*c.* 580–662). In the 9th century, John Scotus, called Erigena ("belonging to the people of Erin") because he was born in Ireland, a master at the Carolingian court of Charles the Bald, translated into Latin some of the writings of these Greek theologians, and his own major work, *De divisione naturae* (862–866; "On the Division of Nature"), is a vast synthesis of Christian thought organized along Neoplatonic lines. For him, God is the primal unity, unknowable and unnameable in himself, from which the multiplicity of creatures flows. He so far transcends his creatures that he is most appropriately called super-real and super-good. Creation is the process of division whereby the many derives from the One. The One descends into the manifold of creation and reveals himself in it. By the reverse process of analysis the multiplicity of creatures will return to its unitary source at the end of time, when everything will be absorbed in God.

Anselm. After the breakdown of the Carolingian Empire in the 10th century, intellectual speculation was at a low ebb in western Europe. In the next century some political stability was achieved by Otto I, who re-established the empire, and Benedictine monasteries were revitalized by reformers, such as Peter Damian. Like Tertullian, a Christian writer of the 2nd and 3rd centuries, Damian mistrusted secular learning and philosophy as harmful to the faith. Other monks showed a keen interest in dialectics and philosophy. Among the latter was Anselm, an Italian who became abbot of the French monastery of Bec and, later, archbishop of Canterbury.

Like Augustine, Anselm used both faith and reason in his search for truth. Faith comes first, in his view, but reason should follow, giving reasons for what men believe. Anselm's monks asked him to write a model meditation on God in which everything would be proved by reason and nothing on the authority of Scripture. He replied with his *Monologium* (1077; "Soliloquy"), the original title of which was "A Meditation on the Reasonableness of Faith." It contains three proofs of the existence of God, all of which are based on Neoplatonic thought. The first proof moves from the awareness of a multiplicity of good things to the recognition that they all share or participate more or less in one and the same Good, who is supremely Good in itself, and this is God. The second and third proofs are similar: beginning with their awareness of a multiplicity of beings that have more or less of being, and more or less of perfection, men recognize that they share in One who is supremely Being and perfect.

Anselm's proofs of the existence of God

Anselm's later work, the *Proslogium* (1077/78; "Allocution" or "Address"), also entitled "Faith Seeking Understanding," contains his most famous proof of the exis-

tence of God. This begins with a datum of faith: men believe God to be the being than which none greater can be thought. Some, like the fool in the Psalms, say there is no God; but even the fool, on hearing these words, understands them, and what he understands exists in his intellect, even though he does not grant that such a being exists in reality. But it is greater to exist in reality and in the understanding than to exist in the understanding alone. Therefore, it is contradictory to hold that God exists only in the intellect, for then the being than which none greater can be thought is one than which a greater can be thought, namely, one that exists both in reality and in the understanding. Philosophers still debate the meaning and value of this so-called ontological argument for God's existence.

Bernard of Clairvaux and Abelard. Anselm's inquiry into the existence and nature of God, as also his discussion of truth, love, and human liberty, aimed at fostering monastic contemplation. Other monks, such as the Cistercian Bernard of Clairvaux, were suspicious of the use of secular learning and philosophy in matters of faith. Bernard complained of the excessive indulgence in dialectic displayed by contemporaries, such as Peter Abelard and Gilbert de la Porrée. He himself developed a doctrine of mystical love, the influence of which lasted through the centuries. The monks of the Parisian Abbey of St. Victor were no less intent on fostering mystical contemplation, but they cultivated the liberal arts and philosophy as an aid to it. In this spirit, Hugh of Saint-Victor wrote his *Didascalion* (*c.* 1127; "Teaching"), a monumental treatise on the theoretical and practical sciences and the *trivium* (grammar, rhetoric, dialectic) and the *quadrivium* (arithmetic, music, geometry, astromony). During this same period the School of Chartres, attached to the famous cathedral near Paris, was the focus of Christian Platonism and Humanism.

Urban development in the 12th century shifted the centre of learning and education from the monasteries to the towns. Abelard founded several urban schools near Paris and taught in them. A passionate logician, he pioneered a method in theology that contributed to the later Scholastic method. His *Sic et non* (1115–17; *Yes and No*) cites the best authorities on both sides of theological questions in order to reach their correct solution. In philosophy his main interest was logic. On the question of universals he agreed with neither the Nominalists nor the Realists of his day. His Nominalist teacher Roscelin held that universals, such as "man" and "animal," are nothing but words or names (*flatus vocis*). Abelard argued that this does not take into account the fact that names have meaning. His Realist teacher William of Champeaux taught that universals are realities apart from the mind. For Abelard, only individuals are real; universals are indeed names or mental concepts, but they have meaning because they refer to individuals. They do not signify an essence common to individuals, as the Realists maintained (*e.g.*, the essence "humanity" shared by all men), but the individuals in their common condition or status of being in a certain species, which results from God's creating them according to the same divine idea.

TRANSITION TO SCHOLASTICISM

In the 12th century a cultural revolution took place that influenced the whole subsequent history of Western philosophy. The old style of education, based on the liberal arts and emphasizing grammar and the reading of the Latin classics, was replaced by new methods stressing logic, dialectic, and all the scientific disciplines known at the time. John of Salisbury, of the School of Chartres, witnessed this radical change:

> Behold, everything was being renovated: grammar was being made over, logic was being remodeled, rhetoric was being despised. Discarding the rules of their predecessors, [the masters] were teaching the *quadrivium* with new methods taken from the very depths of philosophy.

Growing interest in Aristotelianism

In philosophy itself, there was a decline in Platonism and a growing interest in Aristotelianism. This change was occasioned by the translation into Latin of the works of Aristotle in the late 12th and early 13th centuries. Until then, only a few of his minor logical treatises were known. Now his *Topics, Analytica priora,* and *Analytica posteriora* were rendered into Latin, giving the schoolmen (the teachers of Western Christian philosophy in the 13th and 14th centuries) access to the Aristotelian methods of disputation and science, which became their own techniques of discussion and inquiry. Many other philosophical and scientific works of Greek and Arabian origin were translated at this time, creating a "knowledge explosion" in western Europe.

Arabic thought. Among the translations from Arabic were some of the writings of Avicenna (980–1037). This Islāmic philosopher had an extraordinary impact on the medieval schoolmen. His interpretation of Aristotle's notion of metaphysics as the science of *ens qua ens* ("being as being"), his analysis of many metaphysical terms, such as "being," "essence," "existence," and his metaphysical proof of the existence of God were often quoted, with approval or disapproval, in Christian circles. Also influential were his psychology, logic, and natural philosophy. His *al-Qānūn fī aṭ-ṭibb* ("Canon of Medicine") was an authority on the subject until modern times. The *Maqāṣid al-Falāsifah* (1094; "The Aims of the Philosophers") of the Arabian theologian al-Ghazālī, known in Latin as Algazel (d. *c.* 1111), an exposition of Avicenna's philosophy written in order to criticize it, was read as a complement of Avicenna's works. The anonymous *Liber de causis* ("Book of Causes") was also translated into Latin from Arabic. This work, excerpted from Proclus' *Stiocheiōsis theologikē* (*Elements of Theology*), was often ascribed to Aristotle, and it gave a Neoplatonic cast to his philosophy until its true origin was discovered by Thomas Aquinas.

The commentaries of the Arabian philosopher Averroës were translated along with Aristotle's works. As Aristotle was called "the Philosopher" by the medieval philosophers, Averroës was dubbed "the Commentator." These two, more than any others, taught the Scholastics philosophy as a purely rational discipline, divorced from revealed religion. The Christian schoolmen often attacked Averroës as the archenemy of Christianity for his rationalism and for his doctrines of the eternity of the world and the unity of the intellect for all men.

Jewish thought. Of considerably less influence on the Scholastics was medieval Jewish thought. Ibn Gabirol, known to the Scholastics as Avicebron or Avencebrol, was thought to be an Arab or Christian, though in fact he was a Spanish Jew. His chief philosophical work, written in Arabic and preserved only in a Latin translation, entitled *Fons vitae* (*Fountain of Life*), stresses the unity and simplicity of God. All creatures are composed of form and matter, either the gross corporeal matter of the sensible world or the spiritual matter of angels and human souls. Some of the schoolmen were attracted to the notion of spiritual matter and also to Gabirol's analysis of a plurality of forms in creatures, according to which every corporeal being receives a variety of forms by which it is given its place in the hierarchy of being—*e.g.*, a dog has the forms of a corporeal thing, a living thing, an animal, a dog.

Maimonides, or Moses ben Maimon, was known to Christians of the Middle Ages as Rabbi Moses. His *Dalālat al-hā'irīn* (*c.* 1190; *Guide of the Perplexed*) helped them to reconcile Greek philosophy with revealed religion. For Maimonides, there can be no conflict between reason and faith because both come from God; an apparent contradiction is due to a misinterpretation of either the Bible or the philosophers. Thus, he showed that creation is reconcilable with philosophical principles and that the Aristotelian arguments for an eternal world are not conclusive because they ignore the omnipotence of God, who can create a world of either finite or infinite duration.

The universities and Scholasticism

While Western scholars were assimilating the new treasures of Greek, Islāmic, and Jewish thought, universities were being founded, which became the centres of Scholasticism. Of these the most important were Paris and Oxford. Scholasticism is the name given to the theological and philosophical teachings of the schoolmen in the

universities. There was no one Scholastic doctrine; each of the Scholastics developed his own, which was often in disagreement with that of his fellow teachers. They had in common a respect for the great writers of old, such as the Fathers of the Church, Aristotle, Plato, Boethius, the Pseudo-Dionysius, and Avicenna. These they called "authorities." Their interpretation and evaluation of the authorities, however, frequently differed. They also shared a common style and method that developed out of the teaching practices in the universities. Teaching was done by lecture and disputation (a formal debate). A lecture consisted of the reading of a prescribed text followed by the teacher's commentary on it. Masters also held disputations in which the affirmative and negative sides of a question were thoroughly argued by students and teacher, before the latter resolved the problem.

THE AGE OF THE SCHOOLMEN

Robert Grosseteste and Roger Bacon. The newly translated Greek and Arabian treatises had an immediate effect on Oxford University. Its first chancellor, Robert Grosseteste (*c.* 1168–1253), commented on some of Aristotle's works and translated the *Nicomachean Ethics* from Greek to Latin. He was deeply interested in scientific method, which he described as both inductive and deductive. By the observation of individual events in nature, man advances to a general law, called a "universal experimental principle," that accounts for these events. Experimentation either verifies or falsifies a theory by testing its empirical consequences. For Grosseteste, the study of nature is impossible without mathematics. He cultivated the science of optics (*perspectiva*), which measures the behaviour of light by mathematical means. His studies of the rainbow and comets employ both observation and mathematics. His treatise *De luce* (1215–1220; "On Light") embodies a metaphysics of light, presenting light as the basic form of all things and God as the primal uncreated light.

Grosseteste's pupil, Roger Bacon (*c.* 1220–*c.* 1292), made the mathematical and experimental methods the key to natural science. The term experimental science was popularized in the West through his writings. For him, man acquires knowledge through reasoning and experience, but without the latter he can have no certitude. Man gains experience through the senses and also through an interior divine illumination that culminates in mystical experience. Bacon was critical of the methods of the Parisian theologians, such as Albertus Magnus and Thomas Aquinas. He strove to create a universal wisdom embracing all the sciences and organized by theology. He also proposed the formation of a single worldwide society or "Christian republic" that would unite all men under the leadership of the pope.

William of Auvergne. At the University of Paris, William of Auvergne was one of the first to feel the impact of the philosophies of Aristotle and Avicenna. As a teacher, and then as bishop of Paris, he was concerned with the threat of pagan and Islāmic thought to the Christian faith. He opposed the Aristotelian doctrine of the eternity of the world as contrary to the Christian notion of creation. His critique of Avicenna centred around the latter's conception of God and creation. The God of Avicenna, who creates the universe eternally and necessarily, through the mediation of ten Intelligences, was opposed by William of Auvergne with the Christian notion of a God who creates the world freely and directly. Creatures are radically contingent and dependent on God's creative will. Unlike God, they do not exist necessarily; indeed, their existence is distinct from their essence and accidental to it. God has no essence distinct from his existence; he is pure existence. In stressing the essential instability and noneternity of the world, he attributed true existence and causality to God alone. William of Auvergne was a follower of Augustine, but, like others at the time, he was compelled to rethink the older Augustinian notions in terms of the newer Aristotelian and Avicennian philosophies.

Opposition to Aristotelianism

Bonaventure. The Franciscan friar Bonaventure reacted similarly to the growing popularity of Aristotle and his Arabian commentators. He admired Aristotle as a natural scientist, but he preferred Plato and Plotinus, and, above all, Augustine, as metaphysicians. His main criticism of Aristotle and his followers was that they denied the existence of divine ideas. As a result, Aristotle was ignorant of exemplarism (that is, God's creation of the world according to ideas in his mind) and also of divine providence and government of the world. This involved Aristotle in a threefold blindness: he taught that the world is eternal, that all men share one agent intellect (the active principle of understanding in man), and that there are no rewards or punishments after death. Plato and Plotinus avoided these mistakes, but, because they lacked Christian faith, they could not see the whole truth. For Bonaventure, faith alone enables men to avoid error in these important matters.

Bonaventure did not confuse philosophy with theology. Philosophy is the knowledge of the things of nature and the soul innate in man or acquired by his own efforts, whereas theology is the knowledge of heavenly things based on faith and divine revelation. Bonaventure, however, rejects the practical separation of philosophy from theology. Philosophy needs the guidance of faith; far from being self-sufficient, it is but a stage toward the higher knowledge that culminates in the vision of God.

For Bonaventure, every creature to some degree bears the mark of its Creator. The soul has been made in the very image of God. Thus, the universe is like a book in which the triune God is revealed. His *Itinerarium mentis in Deum* (1259; "Journey of the Mind to God") follows Augustine's path to God, from the external world to the interior world of the mind, and then above the mind from the temporal to the eternal. Throughout this journey, men are aided by a moral and intellectual divine illumination. The mind has been created with an innate idea of God, so that, as Anselm pointed out, man cannot think that God does not exist. In a terse reformulation of the Anselmian argument for God's existence, Bonaventure states that if God is God, he exists.

Albertus Magnus. The achievement of the Dominican friar Albertus Magnus was of vital importance for the development of medieval philosophy. A man of immense erudition and intellectual curiosity, he was one of the first to recognize the true value of the newly translated Greco-Arabian scientific and philosophic literature. Everything he considered valuable in it, he included in his encyclopaedic writings. He set out to teach this literature to his contemporaries and in particular to make the philosophy of Aristotle, whom he considered to be the greatest philosopher, understandable to them. He also proposed to write original works in order to complete what was lacking in the Aristotelian system. In no small measure, the triumph of Aristotelianism in the 13th century can be attributed to him.

Triumph of Aristotelianism

Albertus' observations and discoveries in the natural sciences advanced botany, zoology, and mineralogy. In philosophy he was less original and creative than his famous pupil Thomas Aquinas. He produced a synthesis of Aristotelianism and Neoplatonism, blending together the philosophies of Aristotle, Avicenna, and Gabirol, and, among Christians, Augustine and Pseudo-Dionysius.

Thomas Aquinas. Albertus Magnus' Dominican confrere and pupil Thomas Aquinas shared his master's great esteem for the ancient philosophers, especially Aristotle, and also for the more recent Arabian and Jewish thinkers. He welcomed truth wherever he found it and used it for the enrichment of Christian thought. For him, reason and faith cannot contradict each other because they come from the same divine source. In his day conservative theologians and philosophers regarded Aristotle with suspicion and leaned toward the more traditional Christian Platonism. Aquinas realized that their suspicion was due, in part, to the fact that Aristotle's philosophy had been distorted by his Arabian commentators; so he wrote his own commentaries on Aristotle to show the essential soundness of his system and to convince contemporaries of its value for Christian theology.

Aquinas' own philosophical views are best expressed in his theological works, especially his *Summa theologiae*

(1265/66–1273) and *Summa contra gentiles* (1258–1264). In these works he clearly distinguishes between the domains and methods of philosophy and theology. The philosopher seeks the first causes of things, beginning with data furnished by the senses; the subject of the theologian's inquiry is God as revealed in sacred Scripture. In theology, appeal to authority carries most weight; in philosophy, it carries least.

Aquinas found Aristotelianism and, to a lesser extent, Platonism useful instruments for Christian thought and communication; but he transformed and deepened everything he borrowed from them. For example, he took over Aristotle's proof of the existence of a primary unmoved mover, but the primary mover at which Aquinas arrives is very different from that of Aristotle; it is in fact the God of Judaism and Christianity. He also adopted Aristotle's teaching that the soul is man's form and the body is his matter, but for Aquinas this does not entail, as it does for the Aristotelians, the denial of the immortality of the soul or the ultimate value of the human person. Aquinas never compromised Christian doctrine by bringing it into line with the current Aristotelianism; rather, he modified and corrected the latter whenever it clashed with Christian belief. The harmony he established between Aristotelianism and Christianity was not forced but achieved by a new understanding of philosophical principles, especially the notion of being, which he conceived as the act of existing (*esse*). For him, God is pure being, or the act of existing. Creatures participate in being according to their essence; for example, man participates in being, or the act of existing, to the extent that his humanity, or essence, permits. The fundamental distinction between God and creatures is that creatures have a real composition of essence and existence, whereas God's essence is his existence.

Averroists. A group of masters in the Faculty of Arts at Paris welcomed Aristotle's philosophy and taught it in disregard of its possible opposition to the Christian faith. They wanted to be philosophers, not theologians, and to them this meant following the Aristotelian system. Because Averroës was the recognized commentator on Aristotle, they generally interpreted his thought in an Averroistic way. Hence, in their own day they were known as "Averroists"; today they are often called "Latin Averroists" because they taught in Latin. Their leader, Siger of Brabant, taught as rationally demonstrated certain Aristotelian doctrines that contradicted the faith, such as the eternity of the world and the oneness of the intellect for all men. They were accused of holding a "double truth" —of maintaining the existence of two contradictory truths: one commanded by faith, the other taught by reason. Although Siger never proposed as true philosophical conclusions contrary to faith, other members of this group upheld the right and duty of the philosopher to follow human reason to its natural conclusions, even when they contradicted the truths of faith.

This growing rationalism confirmed the belief of theologians of a traditionalist cast that the pagan and Muslim philosophies would destroy the Christian faith. They attacked these philosophies in treatises such as Giles of Rome's *Errores philosophorum* (1270; Eng. trans., *The Errors of the Philosophers*). In 1277 the bishop of Paris condemned 219 propositions based on the new trend toward rationalism and naturalism. These included even some of Aquinas' Aristotelian doctrines. The same year, the archbishop of Canterbury made a similar condemnation at Oxford. These reactions to the novel trends in philosophy did not prevent the Averroists from treating philosophical questions apart from religious considerations. Theologians, on their part, were increasingly suspicious of the philosophers and less optimistic about the ultimate reconciliation of philosophy and theology.

Condemnation of Aristotelian doctrines

PHILOSOPHY IN THE LATE MIDDLE AGES

In the late Middle Ages, earlier ways of philosophizing were continued and formalized into definite schools of thought. In the Dominican order, Thomism (theology and philosophy of Thomas Aquinas) was made the official teaching, though the Dominicans did not always adhere to it rigorously. Averroism, cultivated by philosophers such as John of Jandun (d. *c.* 1328), remained a live, though sterile, movement into the Renaissance. In the Franciscan order, the Englishmen John Duns Scotus and William of Ockham developed new styles of theology and philosophy that vied with Thomism throughout the late Middle Ages.

Scotus. John Duns Scotus (*c.* 1265–1308) opposed the rationalists' contention that philosophy is self-sufficient and adequate to satisfy man's desire for knowledge. In fact, he claimed, a pure philosopher such as Aristotle could not truly understand the human condition because he was ignorant of the Fall of man and his need for grace and redemption. Unenlightened by Christian revelation, Aristotle mistook man's present fallen state, in which all his knowledge comes through the senses, for his natural condition, in which the object of his knowledge would be coextensive with all being, including the being of God. The limitation of Aristotle's philosophy is apparent to Scotus in the Aristotelian proof of the existence of God as the primary mover of the universe. More adequate than this physical proof, he contended, is his own very intricate metaphysical demonstration of the existence of God as the absolutely primary, unique, and infinite being. He incorporated the Anselmian argument into this demonstration. For Scotus, the notion of infinite being, not that of primary mover or being itself, is man's most perfect concept of God.

In opposition to the Greco-Arabian view of the government of the universe from above by necessary causes, Scotus stressed the contingency of the universe and its total dependence on God's infinite creative will. He adopted the traditional Franciscan voluntarism, elevating the will above the intellect in man.

Defense of the doctrine of God's will

Scotus' doctrine of universals justly earned him the title "Doctor subtilis." Universals, in his view, exist only as abstract concepts, but they are based on common natures, such as humanity, which exist, or can exist, in many individuals. Common natures are real, and they have a real unity of their own distinct from the unity of the individuals in which they exist. The individuality of each individual is due to an added positive reality that makes the common nature to be *this* individual; for example, humanity to be Socrates. Scotus calls this reality an "individual difference," or "thisness" (haecceitas). This is an original development of the earlier medieval realism of universals.

William of Ockham. In the late 14th century, Thomism and Scotism were called the "old way" (*via antiqua*) of philosophizing in contrast to the "modern way" (*via moderna*) begun by such men as William of Ockham (*c.* 1285–1349). Ockham, no less than Scotus, wanted to defend the Christian doctrine of the freedom and omnipotence of God and the contingency of creatures against the necessitarianism of Greco-Arabian philosophy. But for him the freedom of God is incompatible with the existence of divine ideas as positive models of creation. God does not use preconceived ideas when he creates, as Scotus maintained, but he fashions the universe as he wishes. As a result, creatures have no natures or essences in common. There are no realities but individual things, and these have nothing in common. They are more or less like each other, however, and on this basis men can form universal concepts of them and talk about them in general terms.

The absolute freedom of God was often used by Ockham as a principle of philosophical and theological explanation. Because the order of nature has been freely created by God, it could have been different: fire, for example, could cool, as it now heats. If he wishes, he can give us the sight, or "intuitive knowledge," of a star without the reality of the star. The moral order could also have been different. God could have made hating him meritorious instead of loving him. It was typical of Ockham not to put too much trust in the power of human reason to reach the truth. For him, philosophy must often be content with probable arguments, for example, in establishing the existence of the Christian God. Faith alone gives certitude in this and in other vital matters. Another

principle invoked by Ockham is that a plurality is not to be posited without necessity. This principle of the economy of thought, later stated as Beings are not to be multiplied without necessity, is called "Ockham's razor."

Ockhamism was censured by a papal commission at Avignon, France, in 1326, and in 1474 it was forbidden to be taught at Paris: nevertheless, it spread widely in the late Middle Ages and rivalled Thomism and Scotism in popularity. Other Scholastics in the 14th century shared Ockham's basic principles and contributed with him to skepticism and probabilism in philosophy. John of Mirecourt (fl. *c.* 1345) stressed the absolute power of God and the divine will to the point of making him the cause of man's sin. Nicholas of Autrecourt (fl. *c.* 1347) adopted a skeptical attitude regarding such matters as man's ability to prove the existence of God and the reality of substance and causality. Rejecting Aristotelianism as inimical to the Christian faith, he advocated a return to the Atomism of the ancient Greeks as a more adequate explanation of the universe.

Trend away from Aristotelianism

Meister Eckehart. The trend away from Aristotelianism was accentuated by the German Dominican Meister Eckehart, who developed a speculative mysticism of both Christian and Neoplatonic inspiration. Eckehart depicts the ascent of the soul to God in Neoplatonic terms: by gradually purifying itself from the body, the soul transcends being and knowledge until it is absorbed in the One. The soul is then united with God at its highest point or "citadel." God himself transcends being and knowledge. Sometimes Eckehart describes God as the being of all things. This language, which was also used by Erigena and other Christian Neoplatonists, leaves him open to the charge of pantheism (the doctrine that the being of creatures is identical with that of God); but for Eckehart there is an infinite gulf between creatures and God. Eckehart means that creatures have no existence of their own but are given existence by God, as the body is made to exist and is contained by the soul. Eckehart's profound influence can be seen in the flowering of mysticism in the German Rhineland in the late Middle Ages.

Nicholas of Cusa. Nicholas of Cusa also preferred the Neoplatonists to the Aristotelians. To him, the philosophy of Aristotle is an obstacle to the mind in its ascent to God because its primary rule is the principle of contradiction, which denies the compatibility of contradictories. But God is the "coincidence of opposites." Because he is infinite, he embraces all things in perfect unity; he is at once the maximum and the minimum. Nicholas uses mathematical symbols to illustrate how, in infinity, contradictories coincide. If a circle is enlarged, the curve of its circumference becomes less; if a circle is infinite, its circumference is a straight line. As for man's knowledge of the infinite God, he must be content with conjecture or approximation to the truth. The absolute truth escapes man; his proper attitude is "learned ignorance."

For Nicholas, God alone is absolutely infinite. The universe reflects this divine perfection and is relatively infinite. It has no circumference, for it is limited by nothing outside of itself. Neither has it a centre; the earth is neither at the centre of the universe nor is it completely at rest. Place and motion are not absolute but relative to the observer. This new, non-Aristotelian conception of the universe anticipated some of the features of modern theories of the universe.

Thus, at the end of the Middle Ages, some of the most creative minds were abandoning Aristotelianism and turning to newer ways of thought. The philosophy of Aristotle, in its various interpretations, continued to be taught in the universities, but it had lost its vitality and creativity. Christian philosophers were once again finding inspiration in Neoplatonism. The Platonism of the Renaissance was in direct continuity with the Platonism of the Middle Ages. (A.A.Ma.)

IV. Modern philosophy

THE RENAISSANCE AND EARLY-MODERN PERIOD

The philosophy of a period arises as a response to social need, and the development of philosophy in the history of Western civilization since the Renaissance has, thus, reflected the process in which creative philosophers have responded to the unique challenge of each stage in the development of Western culture itself.

The career of philosophy—how it views its tasks and functions, how it defines itself, the special methods it invents for the achievement of philosophic knowledge, the literary forms it adopts and utilizes, its conception of the scope of its subject matter, and its changing criteria of meaning and truth—hinges on the mode of its successive responses to the challenges of the social structure within which it arises. Thus, Western philosophy in the Middle Ages was primarily a Christian philosophy, clarifying the divine revelation, reflecting the feudal order in its cosmology, devoting itself in no small measure to the institutional tasks of the Roman Catholic Church. It was no accident that the major philosophical achievements of the 13th and 14th centuries were the work of churchmen who also happened to be professors of theology at the universities of Oxford and Paris.

Cultural impact of new inventions and nations

The Renaissance of the late 15th and 16th centuries presented a different set of problems and therefore suggested different lines of philosophical endeavour. What is called the European Renaissance followed upon the introduction of three novel mechanical inventions from the East: gunpowder, block printing from movable types, and the compass. The first was used to explode the massive fortifications of the feudal order and thus became an agent of the new spirit of nationalism that threatened the rule of churchmen—and, indeed, the universalist emphasis of the church itself—with a competing secular power. The second, printing, made the propagation of knowledge widespread, secularized learning, reduced the intellectual monopoly of an ecclesiastical elite, and restored the literary and philosophical classics of Greece and Rome. The third, the compass, increased the safety and scope of navigation, produced the voyages of discovery that opened up the Western Hemisphere, and symbolized a new spirit of physical adventure and a new scientific interest in the structure of the natural world.

Each of these inventions with its wider cultural consequences presented new intellectual problems and novel philosophical tasks within a changed political and social environment. For, as the power of a single religious authority was slowly eroded under the influence of the Protestant Reformation, and as the prestige of the universal Latin language gave way to vernacular tongues, philosophers became less and less identified with their positions in the ecclesiastical hierarchy and more and more identified with their national origins. The works of Albertus Magnus, St. Thomas Aquinas, St. Bonaventure, and John Duns Scotus had been basically unrelated to the countries of their birth; but the philosophy of Machiavelli was directly related to Italian experience, that of Sir Francis Bacon and Thomas Hobbes was English to the core, and that of René Descartes set the standard and tone of French intellectual life for 200 years.

Dominant strands of Renaissance philosophy. The conventional division of pure knowledge in the contemporary world is drawn between the natural sciences, the social sciences, and the humanities. In the Renaissance, however, fields of learning had not yet become so sharply departmentalized: thus each of these divisions arose in the comprehensive and broadly inclusive area of Renaissance philosophy. For, as the Renaissance mounted its revolt against the reign of religion and therefore reacted against the church, against authority, against Scholasticism, and against Aristotle, there was a sudden blossoming of interest in problems centring about man, about nature, and about civil society. These three interests found exact representation in the three dominant strands of Renaissance philosophy: (1) political theory, (2) Humanism, and (3) the philosophy of nature.

Political theory. As secular authority replaced ecclesiastical, and as the dominant interest of the age shifted from religion to politics, it was natural that the rivalries of the national states and their persistent crises of internal order should raise with renewed urgency philosophical problems about the nature and the moral status of politi-

cal power, which had been practically dormant since pre-Christian times. And this new preoccupation with national unity, internal security, state power, and international justice stimulated the growth of political philosophy in Italy, in France, in England, and in Holland.

Machiavelli and power

In early-16th-century Italy, Niccolò Machiavelli, sometime state secretary of the Florentine republic, explored in *Il Principe* (written 1512–13; Eng. trans., *The Prince*, 1955) and in his *Discorsi sopra la prima deca di Tito Livio* (completed by 1517; Eng. trans., *Discourses*, 1950) the techniques for the seizure and retention of power in ways that seemed to exalt "reasons of state" above morality and codified the actual practices of Renaissance diplomacy for the next 100 years. In fact, Machiavelli was motivated by the most patriotic hopes for the ultimate unification of Italy and by the conviction that the low estate of Italian Renaissance morality needed to be elevated by restoring the ancient Roman virtues. Over half a century later in France, Jean Bodin, magistrate of Laon and a member of the States General, insisted that the state must possess a single, unified, and absolute power; he thus developed in detail the doctrine of national sovereignty in all of its administrative consequences and in its role as the source of all legal legitimacy.

In the 17th century in England, Thomas Hobbes, who was to become tutor to the future Charles II, elaborated the fiction that, in the "state of nature" that preceded civilization, life was "nasty, brutish, and short" with "every man's hand raised against every other," and that a "social contract" was thus agreed upon to convey all private rights to a single sovereign in return for general protection and for the institution of a reign of law. Because law is simply "the command of the sovereign," Hobbes at once turned justice into a by-product of power and denied any right of rebellion except when the sovereign becomes too weak to protect the commonwealth and hold it united.

In Holland, a prosperous and tolerant commercial republic in the 17th century, the issues of political philosophy took a different form. Thus, when the Dutch East India Company commissioned a great jurist, Grotius, to provide a defense of their trade rights and of their free access to the seas, the resulting two treatises, *Mare Liberum* (1609; "Freedom of the Seas") and *De Jure Belli ac Pacis* (1625; Eng. trans., *On the Law of War and Peace*, 1925), were the first significant codifications of international law. Their philosophical originality lay, however, in the fact that, in defending the rights of a small, militarily weak nation against the powerful absolutisms of England, France, and Spain, Grotius was led to a preliminary investigation of the sources and validity of the concept of "natural law"—the notion that inherent in human reason and immutable even against the willfulness of sovereign states are imperative considerations of natural justice and moral responsibility, which must serve as a check against the arbitrary exercise of vast political power.

Ambivalence of political philosophers

In general, the political philosophy of the Renaissance was dualistic: it was haunted, even confused, by the conflict between political necessity and general moral responsibility. Machiavelli, Bodin, and Hobbes asserted claims that justified the actions of Italian despotism and the absolutisms of the Bourbon and Stuart dynasties. Yet Machiavelli was obsessed with the problem of human virtue; Bodin insisted that even the sovereign ought to obey the law of nature—that is, to govern in accordance with the dictates of natural justice; and Hobbes himself found in natural law the rational motivation that causes a man to seek for security and peace. In the end, though Renaissance political necessity required that the philosophical doctrines of Thrasymachus (who held that right is what is in the interest of the strong) be implemented, it could never finally escape a twinge of Socratic conscience.

Humanism. The Renaissance was characterized by the renewed study of mathematics, medicine, and classical literature. The first two sparked the scientific revolution of the 16th and 17th centuries; the last became the foundation of the philosophy of Renaissance Humanism. From its origin, Humanism—suspicious of science and generally indifferent to religion—emphasized anew the centrality of man in the universe, his supreme value and importance. Characteristic of this emphasis was the famous *Oratio de hominis dignitate* (written 1486; Eng. trans., *Oration on the Dignity of Man*, 1956) of a late-15th-century Platonist, Pico della Mirandola, a leading member of Lorenzo de' Medici's Florentine Academy. But the new emphasis upon man's personal responsibility and on the possibility of his self-creation as a work of art was in no small part a consequence of the rediscovery of a series of crucial classical texts, which served to reverse the trends of medieval learning. Renaissance Humanism was predicated upon the victory of rhetoric over dialectic and of Plato over Aristotle, as Quintilian and Cicero had triumphed over Abelard (see above *Early medieval philosophy: Bernard of Clairvaux* and *Abelard*) and as the cramped format of Scholastic philosophic method gave way to a Platonic discursiveness.

Much of this had been prepared by Italian scholarly initiative in the early 15th century. The recently discovered manuscript of Quintilian was used by Lorenzo Valla, an anti-authoritarian Humanist, for the creation of modern rhetoric and the principles of textual criticism. But even more important was the rebirth of an enthusiasm for the philosophy of Plato in Medicean Florence and at the cultivated court of Urbino, a dukedom southeast of Florence. Precisely to service this enthusiasm, Marsilio Ficino, head of the Florentine Academy, had translated the entire Platonic corpus into Latin by the end of the century.

Enthusiasm for Plato

Except for Pico and Giordano Bruno, a late-16th-century Italian philosopher, the direct influence of Platonism upon Renaissance metaphysics is difficult to trace. The Platonic account of the moral virtues, however, was admirably adapted to the requirements of Renaissance education and gave new support to the Renaissance ideal of the courtier and the gentleman. But Plato also represented the philosophical importance of mathematics and the Pythagorean attempt to discover the secrets of the heavens, the earth, and the world of nature in terms of number and exact calculation; and this aspect of Platonism spilled over from Humanism into the domain of Renaissance science. The scientists Nicolaus Copernicus, Johannes Kepler, and Galileo owe more to the general climate of Pythagorean confidence in the explanatory power of number than does Renaissance metaphysics.

But Platonism also had the effect of influencing the literary form in which Renaissance philosophy was written. Although the very early medieval Platonists St. Augustine and John Scotus Erigena occasionally had used the dialogue form, later Scholasticism had abandoned it in favour of the formal treatise, of which the great "Summas" of Alexander of Hales and St. Thomas Aquinas are pristine examples. The Renaissance rediscovery of the Platonic dialogues suggested the literary charm of this conversational method to Humanists, scientists, and political theorists alike. The Humanist philosopher Bruno put forth his central insights in a dialogue, *De la causa, principio e uno* (1584); Galileo presented his novel mechanics in his *Dialogo sopra i due massimi sistemi del mondo* (1632); and even the politician Machiavelli wrote *Dell'arte della guerra* (1521; "On the Art of War") as a genteel conversation taking place in a quiet Florentine garden.

Of course, Renaissance Humanism was primarily a moral and a literary, rather than a narrowly philosophical, movement. And it flowered in figures with broadly philosophical interests such as Erasmus of Rotterdam, erudite citizen of the world; Sir Thomas More, the learned but unfortunate chancellor of Henry VIII; and, in the next generation, in the great French essayist and mayor of Bordeaux, Michel de Montaigne. But the recovery of the Greek and Latin classics, which was the work of Humanism, had profound effects upon the entire field of Renaissance philosophy and science through the ancient schools of philosophy to which it once more directed attention. In addition to Platonism, the most notable of these were Greek atomistic Materialism, Greek Skepti-

Revival of other ancient Greek schools

cism, and Roman Stoicism. The discovery of the manuscript of Lucretius (and the atomistic doctrines of Democritus) finally came to influence Galileo, Bruno, and, later, Pierre Gassendi, a modern Epicurean, through the insights into nature reflected in this work. The recovery of the manuscript of Sextus Empiricus, with its carefully argued Skepticism presented in a printed text in 1562, produced "a skeptical crisis" in French philosophy, which dominated the period from Montaigne to René Descartes. And the Stoicism of Seneca and Epictetus became almost the official ethics of the Renaissance—to appear prominently in the *Essais* (1580–88) of Montaigne, in the letters that Descartes wrote to the princess Elizabeth of Bohemia and to Queen Christina of Sweden, and in the later sections of the *Ethics* (first published 1675) of the Rationalist Benedict de Spinoza.

Philosophy of nature. Philosophy in the modern world is a self-conscious discipline. It has managed to define itself narrowly, so as to differentiate itself on the one hand from religion and on the other from exact science. But this narrowing of focus came about very late in its history—certainly not before the 18th century. The earliest philosophers of Greece were theorists of the physical world; Pythagoras and Plato were at once philosophers and mathematicians; and in Aristotle no clear distinction between philosophy and natural science can be maintained. The Renaissance continued this breadth of conception characteristic of the Greeks. Galileo and Descartes were mathematicians, physicists, and philosophers at once; and physics retained the name of "natural philosophy" at least until the death of Sir Isaac Newton (1727).

Had the Renaissance been painstakingly self-aware in the matter of definition (which it was not), it might have defined philosophy, on the basis of its actual practice, as "the rational, methodical, and systematic consideration of man, civil society, and the natural world." The areas of its interests would in no case have been in doubt. But exactly what constitutes "rational, methodical, and systematic consideration" would have been extremely controversial. For knowledge advances through the discovery and advocacy of new philosophic methods; and, because the diverse methods advocated depend for their validity upon the acceptance of different philosophic criteria of truth, meaning, and importance, the crucial philosophic quarrels of the 16th and 17th centuries were at bottom quarrels in the advocacy of methods. It is this issue, rather than any disagreement over subject matter or areas of attention, that separated the greatest Renaissance philosophers—such as Francis Bacon, René Descartes, and Thomas Hobbes.

The great new fact that confronted the Renaissance was the immediacy, the immensity, and the uniformity of the natural world. But what was of primary importance was the new perspective in which this fact was interpreted. To the Middle Ages the universe was hierarchical, organic, and God-ordained. To the Renaissance it was pluralistic, machinelike, and mathematically ordered. In the Middle Ages scholars thought in terms of purposes, of ends, of divine intentions; in the Renaissance they thought in terms of forces, mechanical agencies, and physical causes.

Leonardo's three theses

All of this had become clear by the end of the 15th century. Within the early pages of the *Notebooks* of Leonardo da Vinci, the great Florentine artist, scientist, Humanist, and mechanical genius, occur the following three propositions:

> (1) Since experience has been the mistress of whoever has written well, I take her as my mistress, and to her on all points make my appeal.
> (2) Instrumental or mechanical science is the noblest and above all others the most useful, seeing that by means of it all animated bodies which have movement perform all their actions.
> (3) There is no certainty where one can neither apply any of the mathematical sciences, nor any of those which are based upon the mathematical sciences.

Here are enunciated respectively: (1) the principle of Empiricism, (2) the advocacy of mechanistic science, and (3) the faith in mathematical explanation; and it is upon these three formulations, as upon a rock, that the science and philosophy of the Renaissance built its foundations. From each of Leonardo's theses descended one of the great streams of Renaissance philosophy: from the empirical principle the work of Francis Bacon; from mechanism the work of Thomas Hobbes; and from mathematical explanation that of René Descartes.

Any adequate philosophic treatment of scientific method surely contains both an empirical principle and a faith in mathematical explanation; and, in Leonardo's thinking, as in scientific procedure generally, there need be no conflict between them. Yet they do represent two poles of emphasis, each capable of exaggeratedly excluding the other. Moreover, the peculiar accidents of Renaissance scientific achievement did present some evidence for their mistaken separation: for the revival of medical studies on the one hand and the novel blooming of mathematical physics on the other emphasized opposite virtues in scientific methodology. This polarity was represented by the opposing figures of Vesalius and Galileo.

In the mid-16th century Andreas Vesalius, a Belgian physician, was astounding all of Europe with the unbelievable precision of his anatomical dissections and drawings. Having invented new tools for this precise purpose, he successively laid bare the vascular, the neural, and the musculature systems of the human body; and this procedure seemed to demonstrate the virtues of empirical method, of physiological experiment, and of the precision and disciplined skill in sensory observation, which made his demonstrations classics of inductive procedure.

Galileo's experiential dynamics

Only slightly later, the Italian physicist Galileo, following in the tradition already established by Copernicus and Johannes Kepler, founders of modern astronomy (but without their more mystical and metaphysical eccentricities), attempted to do for terrestrial and sidereal movements what Vesalius had managed for the structure of the human body—by creating his experiential dynamics, however, with the help of hypotheses supplied by the quantitative calculations of mathematics. In Galileo all of the most original scientific directions of the Renaissance came to a head: the revival of Alexandrian mathematics, the experimental use of new instruments like the lens and the telescope, the search for certainty in physics based upon the undoubted applicability of mathematical theory, and the underlying faith that the search for absolute certainty in science was reasonable because matter in motion accorded with a model of mathematical simplicity. In Galileo also were some of the recurrent themes of 16th- and 17th-century philosophy: an atomism that relates changes in the relations of physical bodies to the corpuscular motion of their parts, the reduction of all qualitative differences to quantitative reasons, and the resultant important distinction between "primary" and "secondary" qualities. The former—including shape, extension, and specific gravity—were considered to be in fact a constituent part of nature and therefore "real." The latter—such as colour, odour, taste, and relative position—were taken to be simply the effect of bodily movements upon perceiving minds and therefore ephemeral, "subjective," and essentially irrelevant to the nature of physical reality.

Rise of Empiricism and Rationalism. The scientific contrast between the rigorous observational techniques of Vesalius and Galileo's reliance upon mathematical theory received further expression in the contrast between the respective philosophies of Francis Bacon and René Descartes. And, indeed, in its more abstract formulation as the contrast between Rationalism and Empiricism, it was to dominate the philosophical controversies of the 17th and 18th centuries and to present a dilemma hardly to be resolved before the advent of Immanuel Kant.

The Empiricism of Francis Bacon. Flourishing around the turn of the 17th century, Sir Francis Bacon, Baron Verulam, Viscount St. Albans (and briefly lord chancellor of England), was the oustanding apostle of Renaissance Empiricism. Less an original metaphysician or cosmologist than the advocate of a vast new program for the advancement of learning and the reformation of scientific method, Bacon's vision of philosophy was that of a new

technique of reasoning that should re-establish natural science upon a firm foundation. In the *Advancement of Learning* (1605) he charted the map of knowledge: history, which depends upon the human faculty of memory; poetry, upon that of imagination; and philosophy, upon man's reason. To reason, however, Bacon assigned a completely experiential function. Fifteen years later, in his *Novum Organum*, he made this clear: because, he said, "we have as yet no natural philosophy which is pure, . . . the true business of philosophy must be . . . to apply the understanding . . . to a fresh examination of particulars." A technique for "the fresh examination of particulars" thus constituted his chief claim to philosophic distinction.

Examination of particulars in tables of discovery

Bacon's hope for a new birth of science hinged not only upon vastly more numerous and varied experiments but chiefly upon "an entirely different method, order, and process for advancing experience." This method consisted in the construction of what he called "tables of discovery." He distinguished three kinds: tables of presence, of absence, and of degree (*i.e.*, in the case of any two properties, such as heat and friction, instances in which they appeared together, instances in which one appeared without the other, and instances in which their amounts varied proportionately); and the ultimate purpose of these tables was to order facts in such a way that the true causes of phenomena (the subject of physics) and the true "forms" of things (the subject of metaphysics—the study of the nature of Being) could be inductively established.

Bacon's was no raw Empiricism; his profound sense of fact and his belief in the primacy of observation were meant to be used for eliciting laws and generalizations. Also, his search for forms was quite un-Platonic: a form for him was not an essence but a permanent geometric or mechanical structure. His enduring place in the history of philosophy lies, however, in his single-minded advocacy of experience as the only source of valid knowledge and in his profound enthusiasm for the perfection of natural science. It is in this sense that "the Baconian spirit" continued as a source of inspiration: his elaborate classification of the sciences inspiring the French Encyclopaedists of the 18th century and his Empiricism inspiring the English philosophers of science of the 19th century.

Hobbes's system of physics and moral and civil philosophy

The mechanism of Thomas Hobbes. During his long lifetime, the English political philosopher Thomas Hobbes was acquainted with both Bacon and Galileo. With the first he shared a strong concern for philosophical method; with the second, an overwhelming interest in matter in motion. His philosophical efforts, however, were more inclusive and more complete than those of either of these contemporaries. He was a comprehensive thinker within the scope of an exceedingly narrow set of presuppositions, and he produced one of the most systematic philosophies of the early-modern period—an almost completely consistent description of nature, man, and civil society according to the tenets of mechanistic Materialism.

Hobbes's account of what philosophy is and ought to be clearly distinguished between content and method. As method, philosophy is simply reasoning or calculating by the use of words as to the causes of phenomena. When a man reasons forward from causes to effects, he reasons synthetically, and, when he reasons from effects backward to causes, he does so analytically. (His strong deductive and geometric bias favoured the former.) Hobbes's dogmatic metaphysical presupposition was that the basic reality is matter in motion. The real world is a corporeal universe in constant movement, and the phenomena, the causes and effects of which it is the business of philosophy to lay bare, are either the mutual action of bodies or the quaint effects of bodies upon minds. From this assumption follows Hobbes's classification of the fields that form the content of philosophy: (1) physics, (2) moral philosophy, and (3) civil philosophy. Physics is the science of the motions and actions of natural bodies conceived in terms of cause and effect. Moral philosophy (or, more accurately, psychology) is the detailed knowledge of "the passions and perturbations of the mind"—*i.e.*, how minds are "moved" by desire, aversion, appetite, fear, anger, and envy. And civil philosophy concerns the concerted actions of men in a commonwealth—how in detail the wayward wills of men are constrained by power (force) in the prevention of civil disorder and the maintenance of peace.

Hobbes's philosophy was a bold Renaissance restatement of Greek atomistic Materialism with applications to the realities of Renaissance politics that would have seemed strange to its ancient originators. But there are also elements in it that make it characteristically English. For Hobbes's conventionalist account of language led him to a nominalistic position—that is, to a position denying the reality of universals (general or common concepts)—in much the same fashion as did Bacon's exaggeration of particulars. Moreover, Bacon's general emphasis upon experience also had its analogue in Hobbes's sensationalist theory of knowledge: the notion that all knowledge has its origin in sense impressions and that all sensations are caused by the action of external bodies upon the organs of sense. Empiricism has been a basic and recurrent expression of British mentality, and its nominalistic and sensationalist roots were already clearly evident in both Bacon and Hobbes.

Rationalism of Descartes. But it was not their philosophy that was to dominate the last half of the 17th century but rather that of René Descartes, a French gentleman who signed himself "Lord of Perron" and who lived the 20 most productive years of his life in the tolerant and hospitable Dutch republic. Descartes is a crucial figure in the history of philosophy, who combined (however unconsciously or even unwillingly) the influences of the past into a new synthesis that seemed so striking in its originality and yet so congenial to the scientific temper of the age as to make him in the minds of all later historians the progenitor of the modern spirit in philosophy.

Synthesis of past and present thought

From the past there seeped into the Cartesian synthesis doctrines about God from Anselm and Aquinas, a theory of the will from St. Augustine, a deep sympathy with the Stoicism of the Romans, and a skeptical method taken indirectly from the ancient Skeptics Pyrrho and Sextus Empiricus. But Descartes was also a great mathematician, who invented analytic geometry, who made many physical and anatomical experiments, who knew and profoundly respected the work of Galileo, and who withdrew from publication his own cosmological treatise *Le Monde* ("The World") after Galileo's condemnation by the Inquisition in 1633.

Each of the maxims of Leonardo, which compose the Renaissance world view, found its place in Descartes: the Empiricism of his physiological researches described in his *Discours de la méthode* (1637), the mechanistic interpretations of the physical world and human action detailed in the *Principia Philosophiae* (1644) and *Traité des passions de l'âme* (1649), and the mathematical bias that dominates his theory of method in the *Regulae ad Directionem Ingenii* (published posthumously in 1701) and his metaphysics in the *Meditationes de Prima Philosophia* (2nd ed. 1642). But of these three, it is the mathematical strain that clearly predominates.

Bacon and Descartes, the founders of modern Empiricism and Rationalism, respectively, shared two pervasive Renaissance tenets: an enormous enthusiasm for physical science; and the pragmatic belief that knowledge means power—that the ultimate purpose of theoretical science is to serve the practical needs of men.

In his *Principia* Descartes defined philosophy as "the study of wisdom" or "the perfect knowledge of all one can know." Its chief utility is "for the conduct of life" (morals), "the conservation of health" (medicine), and "the invention of all the arts" (mechanics). He expressed the relation of philosophy as theoretical inquiry to practical consequences in the famous metaphor of the tree of philosophy whose root is metaphysics, whose trunk is physics, and whose branches are, respectively, morals, medicine, and mechanics. The metaphor is revealing for it indicates that, for Descartes (as for Bacon and for Galileo), the major concern was for the trunk (physics) and that he busied himself with the "roots" only in order to provide a firm foundation for the trunk. Thus the *Dis-*

cours de la méthode, which provides a synoptic view of the Cartesian philosophy, shows it to be *not* (as with Aristotle or Whitehead) a metaphysics founded upon physics but rather—that more characteristic product of the 17th century—a physics founded upon metaphysics.

Elements in Cartesian synthesis

Descartes's mathematical bias was expressed in his determination to ground natural science not in sensation and probability (as did Bacon) but in a principle of absolute certainty. Thus his metaphysics in essence consisted of three principles:

1. To employ the procedure of complete and systematic doubt to eliminate every belief that does not pass the test of indubitability (skepticism);
2. To accept no idea as certain that is not clear, distinct, and free of contradiction (mathematicism);
3. To found all knowledge upon the bedrock certainty of self-consciousness, so that "I think, therefore I am" becomes the only innate idea unshakable by doubt (subjectivism).

From the indubitability of the self, Descartes deduced the existence of a perfect God; and, from the fact that a perfect being is incapable of falsification or deception, made the inference that those ideas about the corporeal world that he has implanted within man must be true. The achievement of certainty about the natural world was thus guaranteed by the perfection of God and by the clear and distinct ideas that are his gift.

The Cartesian metaphysics was the fountainhead of Rationalism in modern philosophy, for it suggests that the mathematical criteria of clarity, distinctness, and absence of contradiction among ideas are the ultimate test of both meaningfulness and truth. This stance is profoundly anti-empirical. Bacon had said, however, that "reasoners resemble spiders who make cobwebs out of their own substance"; and he might well have been speaking of Descartes, for the Cartesian self is just such a substance, from which the innate idea of God originates and with which all deductive reasoning begins. For Descartes the understanding is vastly superior to the senses, and, in the final question of what constitutes truth in science, only man's reason can ultimately decide.

Cartesianism after Descartes

Cartesianism was to dominate the intellectual life of the Continent until the end of the 17th century. It was a fashionable philosophy, appealing alike to learned gentlemen and highborn ladies; and it was one of the few philosophical possibilities to offer competition to the decadent Scholasticism still being taught in the universities. Precisely for this reason it constituted a serious threat to established religious authority. In 1663 the Catholic Church placed Descartes's works upon the Index of Forbidden Books, and the University of Oxford forbade the teaching of his doctrines. Only in the liberal Dutch universities, such as Groningen and Utrecht, did Cartesianism make serious headway.

Certain features of the Cartesian philosophy made it an important starting point for subsequent philosophical speculation. Being essentially the meeting ground of the medieval and the modern world views, it accepted the doctrines of Renaissance science, while attempting to ground them metaphysically in the medieval notions of God and the human mind. Thus a certain dualism between God the creator and the mechanistic world of his creation and between mind as a spiritual principle and matter as mere spatial extension was inherent in the Cartesian position; and a whole generation of French Cartesians (among them Arnold Geulincx, Nicolas Malebranche, and Pierre Bayle) wrestled with the resulting problems of the interaction and reconciliation between the counterposed entities.

Rationalism of Spinoza and Leibniz. But two philosophers of genius carried on the tradition of continental Rationalism: the Dutch Jew Benedict de Spinoza in the mid-17th century and his younger contemporary Gottfried Wilhelm Leibniz, a Leipzig scholar and polymath. Bacon's philosophy had been a search for method in science, and Descartes's basic aim had been the achievement of scientific certainty; but Hobbes and Spinoza provided the most comprehensively worked out speculative systems of the early-modern period. In certain respects they had much in common: a mechanistic picture of the world, with its events guided by a strict determinism, and even a political philosophy in each case looking for political stability based upon centralized power. Yet Spinoza introduced a conception of philosophizing that was new to the Renaissance, something more personal and morally oriented—philosophy as a quest for the wisdom of life, for the achievement of human perfection.

Spinoza's aim, terminology, and method

In conducting this search, Spinoza borrowed much of the basic apparatus of Descartes: the aim at a rational understanding of principles, the terminology of "substance" and of "clear and distinct ideas," and a mathematical method that seeks to convert philosophic knowledge into a complete deductive system using the geometric model of Euclid's *Elements*. Spinoza viewed the universe pantheistically as a single infinite substance, which he called "God," with the dual attributes (or aspects) of thought and extension, and which he differentiated into plural "modes" (or particular things); and he attributed to this world as a whole the properties of a timeless logical system—of a complex of completely determined causes and effects. In so doing, Spinoza was simply seeking for man the series of "adequate" ideas that furnish the intellect and constitute human freedom. For ultimately, for Spinoza, the wisdom that philosophy seeks is achieved when one views the universe in its wholeness, through the "intellectual love of God," which merges the finite individual with the eternal unity and provides the mind with the pure joy that is the final achievement of its search.

The basic elements of the Spinozistic world view are given in his one great work, the *Ethics*, whereas Leibniz' philosophy has to be pieced together from numerous brief expositions or fragments, which seem to be mere intermissions or philosophic interludes in an otherwise busy life. But the philosophic form is deceptive. Leibniz was a mathematician and jurist (inventor of the infinitesimal calculus and codifier of the laws of Mainz), diplomat, historian to royalty, and court librarian in a princely house; yet he was also one of the most original philosophers of the early-modern period. His chief contributions were in the fields of logic, in which he was a truly brilliant innovator, and metaphysics, in which he provided a third alternative to the Rationalist constructions of Spinoza and Descartes. Leibniz saw logic as a mathematical calculus. He was the first to distinguish "truths of reason" from "truths of fact" and to contrast the "necessary" propositions of logic and mathematics (which express identities), which hold for all possible worlds, with the "contingent" (or empirical) propositions of science, which hold only for certain existential conditions; and he saw clearly that, as "the principle of contradiction" controls the first, so "the principle of sufficient reason" governs the second.

A pluralism of monads

In metaphysics Leibniz was a pluralist (as opposed to the dualism of Descartes's thought and extension and the monism of Spinoza's single substance, which is God). There were for him an infinite number of spiritual substances (which he called "monads"), each different, each a percipient of the universe around it, and each mirroring that universe from its own point of view. The chief significance of Leibniz, however, lies not in his differences from Descartes and Spinoza but in the extreme Rationalism that all three shared. In the *Principes de la nature et de la grâce fondés en raison* (1714), he stated the maxim that can stand for the entire school:

> True reasoning depends upon necessary or eternal truths, such as those of logic, numbers, geometry, which establish an indubitable connection of ideas and unfailing consequences.

Literary forms and sociological conditions. The literary forms in which philosophical exposition was couched in the early-modern period ranged from the scientific aphorisms of Bacon and the autobiographical meditations of Descartes to the systematic prose of Hobbes and the episodic propositional format of Leibniz. Two basic tendencies, however, can be discerned:

1. The early-Renaissance commitment to the dialogue

form (already noted), inspired by the rediscovery of the Platonic dialogues;

2. The later prevalence of the systematically ordered treatise, undoubtedly influenced by the enormous prestige of deductive mathematics.

The concept of serial order stressed by geometry, in which the reasoner passes from more universal axioms to more specific derivative propositions, influenced in turn the style of Hobbes, Descartes, and Spinoza. The order of presentation in Hobbes's *Leviathan* and in Descartes's *Principia Philosophiae* reflects this serial concern, while Spinoza's *Ethics* utilizes the Euclidean method so formalistically as almost to constitute an impenetrable barrier to the basic lucidity of his thought.

Medieval philosophy was characteristically associated with the centring of Christian pedagogy in the medieval university with its texts, readings, learned authorities, Disputed Questions, Quodlibetal Questions (brief academic discussions), and its Summas. It is a singular fact, however, that from the birth of Bacon in 1561 to the death of Hume in 1776—*i.e.*, for 200 years—not one first-rate philosophic mind in Europe was permanently associated with a university.

Non-professional status of philosophers

The fact is that, as the age of the saint passed into that of the gentleman, philosophers too reflected this profound change in their titles, their social status, and their economic situation. Sir Francis Bacon was a lawyer, judge, and attendant upon the royal court. Thomas Hobbes was the perpetual tutor and companion of young noblemen. René Descartes, son of a noble family, travelled and studied at leisure, retiring to Holland to live out his life on inherited income. Gottfried Wilhelm Leibniz, courtier, diplomat, and scholar, became a privy councillor and baron of the Holy Roman Empire. Thus philosophers were often of the lesser nobility or were closely associated with the nobility, to whom—like poets—they dedicated their works; and they lived not by philosophy but for it: independently, by pensions, gifts, inherited income, or in the households of the nobility.

Thus, philosophy in the 16th and 17th centuries was clearly the preoccupation of a widely scattered elite; and this meant that, despite the printed essay, much philosophical communication took place within a small, but at the same time loose and informal, circle. Treatises were circulated in manuscript; comments and objections were solicited; and a vast polemical correspondence was built up. Prior to its publication, Descartes prudently sent his *Meditationes* to the theologians of the Sorbonne for comment; and, after its publication, his friend Mersenne sent it to Hobbes, Antoine Arnauld, and Pierre Gassendi among others, who returned formal "objections," to which Descartes in turn replied. In addition, the 17th century possessed a rich repository of philosophical correspondence, such as the letters that passed between Descartes and the scientist Christiaan Huygens, between Spinoza and Henry Oldenburg (one of the first secretaries of the Royal Society), and between Leibniz and Arnauld. But philosophers were also familiar with the great monarchs and administrators of the age: Descartes giving philosophical instruction to Queen Christina of Sweden, Leibniz being intimate with Queen Sophia Charlotte of Prussia, and Spinoza enjoying the personal friendship of the Dutch politician Johan de Witt.

The chief fact, however, remains the sharp separation of creative philosophers from the formal centres of learning: Hobbes expressed extreme contempt for the decadent Aristotelianism of Oxford; Descartes, despite his prudence, scorned the medievalists of the Sorbonne; and Spinoza refused the offer of a professorship of philosophy at Heidelberg with polite aversion. It was to be another 100 years before philosophy returned to the universities, which the chief Renaissance thinkers had abandoned.

THE ENLIGHTENMENT

Although they both lived and worked in the late 17th century, Sir Isaac Newton and John Locke were the true fathers of the Enlightenment. Newton was the last of the scientific geniuses of the age, and his great *Philosophiae Naturalis Principia Mathematica* (1687) was the culmination of the entire movement that had begun with Copernicus and Galileo—the first great physical synthesis based upon the application of mathematics to nature in every detail. The basic idea of the authority and autonomy of reason, which dominated all philosophizing in the 18th century, was, at bottom, the consequence of Newton's work.

Seminal influence of Newton's *Principia*

Copernicus, Kepler, Bacon, Galileo, and Descartes—scientists and methodologists of science—performed like men urgently attempting to persuade nature to reveal her secrets. Newton's comprehensive mechanistic system made it seem as if at last she had done so. It is impossible to exaggerate the enormous enthusiasm that this fact kindled in all of the major thinkers of the 18th century from Locke to Kant. For the new enthusiasm for reason that they all instinctively shared was based not upon the mere advocacy of propagandists like Descartes and Leibniz but upon the conviction that, for the intellectual conquest of the natural world, reason had really worked.

Classical British Empiricism and its basic tasks. Two major philosophic problems remained: to account for the genetic origins of the reason that had proved so successful and to shift its application from external nature to man. John Locke's *Essay Concerning Human Understanding* (1690) was devoted to the first; and David Hume's *Treatise of Human Nature* (1739–40), "being an attempt to apply the method of experimental reasoning to moral subjects," was devoted to the second.

Shift to epistemological and Empiricist philosophy

These two basic tasks reflect a shift in the direction in which philosophizing had moved in the late Renaissance. For the Renaissance preoccupation with the natural world had represented a certain "realistic" bias. Hobbes and Spinoza had each produced a metaphysics. They had been interested in the real constitution of the physical world. Moreover, the Renaissance bias in favour of mathematics had generated the profound interest in rational principles, necessary propositions, and innate ideas that was so prevalent in the philosophies of Leibniz and Descartes. On the other hand, the Enlightenment, in turning from the realities of nature to account for the structure of the mind that knows it so successfully and in attempting an experiential account of the furnishings of that mind, settled on the sensory components of knowledge rather than on the merely mathematical. Thus the school of so-called British Empiricism (John Locke, George Berkeley, and David Hume) dominated the perspective of Enlightenment philosophy until the time of Kant. And this school philosophized in terms of ideas rather than things and of experience rather than innate necessary principles. Whereas the philosophy of the late Renaissance had been metaphysical and Rationalistic, that of the Enlightenment was epistemological and Empiricist.

Origin and nature of reason in Locke and Berkeley. Locke's *Essay Concerning Human Understanding* thus marked a decisively new direction for modern philosophizing, because it proposed what amounts to a new criterion of truth. The design of his essay was "to inquire into the origin, certainty, and extent of human knowledge," which involved three tasks:

1. To discover the origin of men's ideas;
2. To exhibit their certainty and evidential value;
3. To examine the claims of all knowledge that is less than certain.

What was crucial for Locke, however, was that the second task is dependent upon the first. Following the general Renaissance custom, Locke defined an "idea" as a mental content, as "whatever is the object of the understanding when a man thinks"; but, whereas for Descartes and the entire Rationalistic school the certainty of ideas had been a function of their self-evidence—*i.e.*, of their clarity and distinctness—for Locke their validity hinged expressly upon the mode and manner of their origin. A genetic criterion of truth and validity replaced an intrinsic one.

Locke's exhaustive survey of mental contents is useful, if elaborate. Though he distinguished between ideas of sensation and ideas of reflection, the whole thrust of his efforts and those of his Empiricist followers was to re-

duce the latter to the former, to minimize the originative power of the mind in favour of its passive receptivity to the sensory impressions received from without. Locke's classification of ideas into simple and complex was an attempt to distinguish mental contents such as blueness or solidity, which come from a single sense like sight or touch, and those such as figure, space, extension, rest, and motion, which are the product of several senses combined, on the one hand, from those complicated and compounded (complex) ideas of universals such as triangle or gratitude, of substances, and of relations such as identity, diversity, and cause and effect, on the other.

Locke's effort to build a total world

Locke's *Essay* was a dogged attempt to produce the total world of man's conceptual experience out of a set of elementary sensory building blocks, moving always from sensation toward thought and from the simple to the complex. The basic outcome of his epistemology was therefore:

1. That the ultimate source of men's ideas is sensation;
2. That all mental operations are a combining and compounding of simple sensory materials into complex conceptual tools.

It was a theory of knowledge based upon a kind of sensory atomism, which sees the mind as an agency of discovery rather than of creation and views its ideas as "like" the objects that are the sources of the sensations it receives. But it sees also that an important distinction must still be made between those "primary qualities" such as solidity, figure, extension, motion, and rest, which are the actual characteristics of objects themselves, and those "secondary qualities" such as colour, taste, and smell, which are simply the internal consequences of how the mind is affected by them. This important Lockean distinction (already found in Galileo) was only a mirroring in the theory of knowledge of the physics of Sir Isaac Newton, which was its contemporary.

It was precisely this dualism presupposed by the science of the time between primary qualities, which belong to matter, and secondary qualities, which belong to mind, that Locke's successor George Berkeley sought to overcome. Though Berkeley was a bishop in the Anglican Church who wanted to combat "atheistic Materialism" and to sustain the power of God and spirituality, his importance for the theory of knowledge lies rather in the way in which he demonstrated that, in the end, primary qualities are reducible to secondary qualities. His Empiricism led to a denial of abstract ideas, because he believed that general notions are simply fictions of the mind and vehemently denied that one may validly distinguish between objects and the sensory impressions that repose in the mind. Science, he argued, can easily dispense with the concept of matter: nature is simply that which men perceive by their senses, and this means that sense data can be considered as "objects for the mind" rather than as "qualities adhering in a substance." A thing is simply a recurrent group of sense qualities. With this important reduction of substance to quality, Berkeley thus became the true father of the epistemological position known as phenomenalism, which has remained an important influence in British philosophizing to the present day.

Basic science of man in Hume. The third, and in many ways the most famous, of the British Empiricists was the 18th-century Skeptic David Hume. Hume's philosophic intention was to reap, humanistically, the harvest sowed by Newtonian physics, to apply the method of natural science to human nature, and to create a basic science of man. The paradoxical result of this admirable purpose, however, was to create a skeptical crisis more devastating than that of the early French Renaissance and to reduce human certainty once more to the state that it was in before Descartes had reached the dogmatic halting point in his procedure of methodical doubt.

Sense impressions and association of ideas

Hume followed Locke and Berkeley in approaching the problem of knowledge from a psychological perspective. He too found the origin of knowledge in sense impressions. But whereas Locke had found a certain trustworthy order in the compounding power of the mind, and Berkeley had found mentality itself expressive of a certain spiritual power, Hume's relentless analysis discovered as much contingency in mind as in the external world. All uniformity in perceptual experience, he held, comes from "an associating quality of the mind." The "association of ideas" is a fact, but the relations of resemblance, contiguity, and cause and effect that it produces have no intrinsic validity, because they are the product of an inexplicable "mental habit." Thus the causal principle, upon which all knowledge rests, indicates no necessary connections between things but is simply the accident of their constant conjunction in men's minds. Moreover, the mind itself, far from being an independent power, is simply "a bundle of perceptions" without unity or cohesive quality. Hume's denial of a necessary order of nature, on the one hand, and of a substantial or unified self, on the other, precipitated a philosophic scandal from which Enlightenment philosophy was not to be definitely rescued until the work of Kant.

Nonepistemological movements in the Enlightenment. Though the school of British Empiricism represented the mainstream of Enlightenment philosophy until the time of Kant, it was by no means the only type of philosophy that the 18th century produced. For the Enlightenment, which was based upon a few great fundamental ideas—such as the dedication to reason, the belief in intellectual progress, the confidence in nature as a source of inspiration and value, and the search for tolerance and freedom in political and social institutions—produced many cross-currents of intellectual and philosophic expression.

Materialism and scientific discovery. The profound influence of Locke spread to France, where it not only produced the skeptical Empiricism of Voltaire but also united with the mechanistic side of the teachings of Descartes to create an entire school devoted to a sensationalistic Materialism. Julien de La Mettrie in *L'Homme-machine* (1747; Eng. trans., *Man a Machine*, 1750), Étienne de Condillac in his *Traité des sensations* (1754), and Paul, baron d'Holbach, in his *Système de la nature* (1770) represented a limited world view in which matter in motion is the only reality, determining the functioning of the human brain and its sensations according to ironclad and necessary laws. This position even found its way into many of the articles of the great French encyclopaedia edited by Denis Diderot and Jean d'Alembert, which was almost a complete compendium of the scientific and humanistic accomplishments of 18th-century intellectual life.

Optimism and the beginning of many disciplines

Though the Middle Ages and the Renaissance had not referred to themselves by these names, the 18th century called itself "the Enlightenment" with self-conscious enthusiasm and pride. It was an age of optimism with a sense of new beginnings. Great strides were made in chemistry and biological science. Jean, chevalier de Lamarck, Georges Cuvier, and Georges, comte de Buffon, were perfecting a system of animal classification. And, in the eight years between 1766 and 1774, Henry Cavendish discovered hydrogen; Daniel Rutherford, nitrogen; and Joseph Priestley, oxygen. It was the period when foundations were being laid in psychology and the social sciences and in ethics and aesthetics. Anne-Robert-Jacques Turgot, Marquis de Condorcet, and Montesquieu in France, Giambattista Vico in Italy, and Adam Smith in England marked the beginning of history, economics, sociology, and jurisprudence as sciences. Hume, Jeremy Bentham, and the British moral-sense philosophers were turning ethics into a specialized field of philosophical inquiry; and Anthony, 3rd earl of Shaftesbury, Edmund Burke, Johann Gottsched, and Alexander Baumgarten were laying the foundations for a systematic aesthetic.

Social and political philosophy. But, outside of the theory of knowledge, the most significant contribution of the Enlightenment came in the field of social and political philosophy, as Locke's *Two Treatises of Civil Government* (1690) and Jean-Jacques Rousseau's *Du contrat social* (1762) proposed a justification of political association grounded in the newer political requirements of the age. The Renaissance political philosophies of Machiavelli, Bodin, and Hobbes had centred around the absolute power of kings and rulers. But the Enlightenment theories of Locke and Rousseau turned rather to the free-

Freedom, equality, justice, rights

dom and equality of citizens. It was a natural historical transformation. The 16th and 17th centuries had comprised the age of absolutism; the political problem had been largely that of internal order, and political theory had been presented in the language of national sovereignty. But the 18th century was the age of the democratic revolutions (including the French Revolution); the political problem was that of freedom and the revolt against injustice, and political theory was expressed in the idiom of natural and inalienable rights.

Locke's political theory was an express denial of the divine right of kings and the absolute power of the sovereign as enthroned in the doctrines of Hobbes. Instead, he insisted that all men have a natural right to freedom and equality. The state of nature in which men originally live is not, as in Hobbes, intolerable; but it has certain inconveniences, and so men band together to form society, as Aristotle taught, "not simply to live, but to live well." Political power can never be exercised apart from its ultimate purpose, which is the common good; for men enter the political contract in order to preserve life, liberty, and property.

Locke thus stated one of the fundamental principles of the liberal tradition: that there can be no subjection to power without consent, although once political society has been founded, there is an obligation to submit to the decisions of the majority. It is the legislature that makes these decisions, although the ultimate power of choosing the legislature rests with the people; and even the powers of the legislature are not absolute, because the laws of nature remain as a permanent standard and as principles of protection against any arbitrary authority.

Rousseau's more radical political doctrines were built upon Lockean foundations. For him, too, the convention of the social contract formed the basis of all legitimate authority among men, although his conception of citizenship was much more organic, much less individualistic than Locke's. The surrender of natural liberty for civil liberty means that all individual rights (among them property rights) become subordinate to the general will. For Rousseau the state is a moral person whose life is the union of its members, whose laws are acts of the general will, and whose end is the liberty and equality of the citizens. It follows that, when any government usurps the power of the people, the social compact is broken; and not only are the citizens no longer compelled to obey, but they also have an obligation to rebel. Rousseau's defiant collectivism was clearly a revolt against Locke's systematic individualism; for him the fundamental category is not "natural person" but "citizen." Nevertheless, however much they differed, in these two social theorists of the Enlightenment is to be found the germ of all modern liberalism: its faith in representative democracy, in civil liberties, and in the basic dignity of man.

Professionalization of philosophy. In his *Éléments de philosophie* (1759), Jean d'Alembert, an 18th-century French mathematician and Encyclopaedist, wrote:

> Our century is the century of philosophy par excellence. If one considers without bias the present state of our knowledge, one cannot deny that philosophy among us has shown progress.

D'Alembert was calling attention to that reflective self-examination for which the 18th century is famous, and he was undoubtedly referring to the activities of mathematicians like himself, jurists, economists, and amateur moralists rather than to narrow philosophical specialists. But the 18th century was clearly the "century of philosophy par excellence" in a more technical sense also. For it was the period in which philosophizing first began to pass from the hands of gentlemen and amateurs into those of true professionals. The chief sign of this shift was the return of reputable philosophy to the universities.

University of Halle and Christian Wolff

This transformation first occurred in Germany and is chiefly associated with the University of Halle (founded 1694). In the time of the generation that lies between Leibniz (died 1716) and Kant (born 1724), the philosophical climate changed profoundly. The chief representative of this change was Christian Wolff, who taught philosophy at Halle and was the intermediary between the ideas of Leibniz and those of Kant and the profoundly different conceptions of philosophizing for which they stood.

Kant later called Wolff "the real originator of the spirit of thoroughness in Germany"; and Wolff was indeed a pioneer in those techniques that transform philosophy into a professional discipline: the self-conscious adoption of a systematic approach and the creation of a specialized philosophical vocabulary. Wolff carefully distinguished the various fields of philosophy; wrote textbooks in each of them, which were used in the German universities for many years; and created many of the specialized philosophical terms that have survived to the present day.

The German Enlightenment was the first modern period to produce "specialists in philosophy." In England, philosophizing in the universities did not become serious until well after the time of Hume, but already philosophical fields had been sufficiently distinguished to be represented by distinct professorships. The titles professor of mental or moral or metaphysical philosophy, as they arose at Oxford and Cambridge, were the product of the late 18th and early 19th centuries. Two other factors of the German Enlightenment, however, are also relevant: (1) the founding of the first professional journals and (2) the rising concern of philosophy with its own history.

Journals and histories of philosophy

The learned journal, like the scientific society, was an innovation of the 17th century. But what had begun as a general intellectual endeavour became in the German 18th century a specifically philosophical enterprise; and a rash of journals broke out, such as the *Acta Philosophorum* (Halle, 1715–26); the *Philosophische Büchersaal* (Leipzig, 1741–44); and the short-lived *Neues philosophisches Magazin* (Leipzig, 1789–91), devoted exclusively to the philosophy of Kant.

More interesting still is the flowering of voluminous German histories of philosophy after 1740. By the 18th century there was already a vast accumulation of historical materials, and the self-conscious feeling that philosophy constitutes a specific field the past of which is worth examination and careful ordering combined with the German spirit of thoroughness to produce a series of massive histories such as Johann Brucker's *Historia Critica Philosophiae*, 6 vol.; Johann Buhle's *Lehrbuch der Geschichte der Philosophie*, 8 vol.; Dietrich Tiedemann's *Geist der spekulativen Philosophie von Thales bis Berkeley*, 6 vol.; and Gottlieb Tennemann's *Geschichte der Philosophie*, 11 vol.

Critical examination of reason in Kant. All of these developments led directly to Immanuel Kant, professor at Königsberg, the greatest philosopher of the modern period, who marks the true culmination of the philosophy of the Enlightenment. Historically speaking, Kant's great substantive contribution was to relate both the sensory and the a priori elements in knowledge and thus to mend the breach between the extreme Rationalism of Leibniz and the extreme Empiricism of Hume. But in addition to the brilliant content of his philosophical doctrines, Kant is responsible for three crucial philosophical innovations: (1) a new definition of what philosophy is, (2) a new conception of philosophical method, and (3) a new structural model for the writing of philosophy.

Kant's definition of philosophy culminated the Enlightenment, for it took reason to be the very heart of the philosophic enterprise. Philosophy's sole task, in his view, is to determine what reason can and cannot do. Philosophy, he said, "is the science of the relation of all knowledge to the essential ends of human reason"; and its true aim is both constructive ("to outline the system of all knowledge arising from pure reason") and critical ("to expose the illusions of a reason that forgets its limits"). Philosophy is thus a calling of great dignity, for its aim is wisdom, and its practitioners are themselves "lawgivers of reason." But in order for philosophy to be "the science of the highest maxims of reason," the philosopher must be able to determine the source, the extent, and the validity of human knowledge and the ultimate limits of reason. And these tasks require a special philosophic method.

Sometimes Kant called this the "transcendental method," but more often the "critical method"; for his pur-

The transcendental, or critical, method of examining reason

pose was to reject the dogmatic assumptions of the Rationalist school, and his wish was to return to the semiskeptical position with which Descartes had begun before his dogmatic pretensions to certainty took hold. Kant's method was to conduct a critical examination of the powers of an a priori judging reason and to inquire what reason can achieve when all experience is removed. The method was based upon a doctrine that he himself called "a Copernican revolution" in philosophy (by analogy with the change from geocentric to heliocentric cosmology): the assumption not that man's knowledge must conform to objects but that objects must conform to man's apparatus of knowing. The question then became: What is the exact nature of that knowing apparatus?

Unlike Descartes, Kant could not question *that* knowledge exists. Mathematics and Newtonian physics were too real for any Enlightenment mentality to entertain this doubt. Kant's methodological question was rather: *How* is mathematical and physical knowledge possible? What must be the structure of man's knowing process to have made these sciences secure? Exactly the attempt to answer these questions was the task of Kant's great *Critik der reinen Vernunft* (1781; Eng. trans., *Critique of Pure Reason*, 1929).

But Kant's great aim was to examine reason not merely in one of its domains, but in each of its employments, according to the threefold structure of the human mind, which he had taken over from Wolff. Thus the critical examination of reason in thinking (science) is undertaken in the *Critik der reinen Vernunft*, that of reason in willing (ethics) in the *Kritik der praktischen Vernunft* (1788; Eng. trans., *Critique of Practical Reason*, 1949), and that of reason in feeling (aesthetics) in *Kritik der Urteilskraft* (1790; Eng. trans., *Critique of Judgment*, 1951).

Literary forms. The literary format of Enlightenment philosophizing was essentially simple and straightforward. Except for an occasional reversion to the dialogue form as in Berkeley's *Three Dialogues Between Hylas and Philonous* (1713) and Hume's *Dialogues Concerning Natural Religion* (1779), it consisted of inquiries, discourses, treatises, dissertations, and essays, generally well written in clear, relatively nontechnical, prose style. But Kant not only introduced a formidable technical philosophical terminology into his works but was, in fact, the originator of a new philosophical form—the "critique" or "critical examination"—which had its own special architectonic structure. Each of Kant's three critiques consists of the same division into three parts: (1) an "Analytic," or analysis of reason's right functioning; (2) a "Dialectic," or logic of error, showing the pitfalls into which a careless reason falls; and (3) a "Methodology," which is an arrangement of rules for practice. It is a form that was unique with Kant, but it raised certain problems of "oppositional" thinking, to which 19th-century philosophers such as Hegel, Schopenhauer, and Kierkegaard were subsequently to turn.

THE 19TH CENTURY

The end of Enlightenment

Kant died in 1804. Formally, it was the end of the Enlightenment. But the 19th century ushered in new philosophical problems and new conceptions of what philosophy ought to do. It was a century of great philosophic diversity and variety. In the Renaissance the chief intellectual fact had been the rise of mathematics and natural science, and the tasks that this fact imposed upon philosophy determined its direction for two centuries. In the Enlightenment attention had turned to the character of the mind that had so successfully mastered the natural world, and Rationalists and Empiricists had contended for mastery until the Kantian synthesis. As for the 19th century, however, if one single feature of its thought could be singled out for emphasis, it might be called the discovery of the irrational. But many philosophical schools were present, and they contended, one with another, in a series of distinct and powerful oppositions: Pragmatism against Idealism; Positivism against irrationalism; Marxism against liberalism.

Politically, the 19th century began with the consulate of Napoleon and ended with the Golden Jubilee of Queen Victoria; but it is the intellectual and social changes that fell in between that have philosophic consequences. These changes were chiefly: the Romantic movement of the 1820s and 1830s, which was a poetic revolt against reason in favour of feeling; the maturation of the Industrial Revolution, which caused untold misery within society and called forth a multitude of philosophies of social reform; the Revolution of 1848 in Paris, Germany, and Vienna, which symbolized class divisions and first enthroned in the European consciousness the concepts of "the bourgeoisie" and "the proletariat"; and, finally, the great surge in biological science with Darwin and the publicizing of the idea of biological evolution. Romanticism influenced both German Idealism and philosophers of irrationalism. Experiences of economic discord and social unrest produced the ameliorative social philosophy of English Utilitarianism and the revolutionary doctrines of Karl Marx. And the developmental ideas of Darwin provided the prerequisites for American Pragmatism.

A synoptic view of philosophy in the 19th century reveals an interesting chronology. The early century was dominated by the German school of absolute Idealism (Johann Fichte, Friedrich Schelling, and G.W.F. Hegel). The midcentury was marked by a rebirth of interest in science and its methods (Auguste Comte in France and John Stuart Mill in England) and by liberal (Mill) and radical (Marx) social theory. The late century saw a second flowering of Idealism, this time in England (T.H. Green, F.H. Bradley, and Bernard Bosanquet) and, with Charles Sanders Peirce and William James, the rise of American Pragmatism. The new philosophies of the irrational in the highly individual thinkers Søren Kierkegaard, Arthur Schopenhauer, and Friedrich Nietzsche ran through the century in its entirety.

German Idealism of Fichte, Schelling, and Hegel. The Enlightenment, inspired by the example of natural science, had accepted certain bounds to the possibility of knowledge; *i.e.*, it had recognized certain limits to reason's ability to penetrate ultimate reality by methods that surpass the boundaries of scientific method. In this particular modesty, the philosophies of Hume and Kant were much alike. But the early 19th century marked a resurgence of the metaphysical spirit at its most ambitious and extravagant extreme. German Idealism reinstated the speculative pretensions of Leibniz and Spinoza at their height. This turn was partly a consequence of the Romantic influence but, more importantly, of a new alliance of philosophy not with science but with religion. It was not accidental that all of the great German Idealists were university professors, whose fathers were Protestant pastors or who had themselves studied theology: Fichte at Jena and Leipzig (1780–84); Schelling and Hegel at the Tübingen seminary (1788–95). And it is probably this circumstance that gave to German Idealism its intensely serious, its quasi-religious, and its dedicated character.

Resurgence of religiously inspired metaphysics

The consequence of this religious alignment was that philosophic interest shifted from Kant's *Critique of Pure Reason* (in which he had attempted to account for natural science and denied the possibility of certainty in metaphysics) to his *Critique of Practical Reason* (in which he had explored the nature of the moral self) and his *Critique of Judgment* (in which he had treated of the purposiveness of the universe as a whole). For absolute Idealism was based upon three premises:

1. That the chief datum of philosophy is the human self and its self-consciousness;
2. That the world as a whole is spiritual through and through; is, in fact, something like a cosmic Self;
3. That, in both the self and the world, it is not primarily the intellectual element that counts but, rather, the volitional and the moral.

Thus, to understand the self, self-consciousness, and the spiritual universe became for Idealistic metaphysics *the* task of philosophy.

From the point of view of doctrine, Fichte, Schelling, and Hegel had much in common. Fichte, professor of philosophy at the newly founded University of Berlin (1809–14) and a great symbol of German patriotism

through the Napoleonic Wars, combined in a workable unity the subjectivism of Descartes, the cosmic monism of Spinoza, and the moral intensity of Kant. He saw human self-consciousness as the primary metaphysical fact through the analysis of which the philosopher finds his way to the cosmic totality that is "the Absolute." And, just as the moral will is the chief characteristic of the self, so also is it the activating principle of the world. Thus Fichte provided a new definition of philosophizing that made it central in dignity in the intellectual world. The sole task of philosophy is "the clarification of consciousness." And the highest degree of self-consciousness is achieved by the philosopher, because he alone recognizes "Mind," or "Spirit," as the central principle of reality.

Hegel: "the rational is the real"

This line of thought was carried further by G.W.F. Hegel, Fichte's successor at Berlin and perhaps the single most comprehensive and influential thinker of the 19th century. Kant's problem had been the critical examination of reason's role in human experience. For Hegel, too, the function of philosophy is to discover the place of reason in nature, in experience, and in reality; to understand the laws according to which reason operates in the world. But, whereas Kant had found reason to be the form that mind imposes upon the world, Hegel found it to be constitutive of the world itself—not something that mind imposes but that it discovers. As Fichte had projected consciousness from mind into reality, so Hegel projected reason; and the resultant Hegelian dictates—that "the rational is the real" and that "the truth is the whole"—although they express an organic and a totalitarian theory of truth and reality, tend to blur the usual distinctions that previous philosophers had made between logic and metaphysics, between subject and object, and between thought and existence. For the basic tenet of Idealism, that reality is spiritual, generates just such a vague inclusiveness.

To the Fichtean foundations, however, Hegel added one crucial corollary: that the Absolute, or Whole, which is a concrete universal entity, is not static but undergoes a crucial development in time. Hegel called this evolution "the dialectical process." By stressing it, Hegel accomplished two things: (1) he indicated that reason itself is not eternal but "historical," and (2) he thereby gave new meaning and relevance to the changing conditions of human society in history—which added to the philosophic task a cultural dimension that it had not possessed before.

The philosopher's vocation, in Hegel's view, was to approach the Absolute through consciousness, to recognize it as Spirit expressing and developing itself ("realizing itself" was his own phrase) in all of the manifold facets of human life. For struggle is the essence of spiritual existence, and self-enlargement is its goal. For these reasons the various branches of intellect and culture become stages in the unfolding of the World-Spirit:

1. The psychological characteristics of man (habit, appetite, judgment) representing "Subjective Spirit";
2. His laws, social arrangements, and political institutions (the family, civil society, the state) expressing "Objective Spirit";
3. His art, religion, and philosophy embodying "Absolute Spirit."

What began, therefore, in Hegel as a metaphysics of the Absolute ended by becoming a total philosophy of human culture.

Positivism and social theory in Comte, Mill, and Marx. The absolute Idealists wrote as if the Renaissance methodologists of the sciences had never existed. But if, in Germany, the Empirical and scientific tradition in philosophy lay dormant, in France and in England in the middle of the 19th century it was very much alive. In France, Auguste Comte wrote his great philosophical history of science, *Cours de philosophie positive* (1830–42) in six volumes. Influenced by Bacon and the entire school of British Empiricism, by the doctrine of progress put forward by Turgot and Condorcet during the 18th century, and by the very original social reformer Henri de Saint-Simon, Comte called his philosophy "Positivism," by which he meant a philosophy of science so narrow that it denied any validity whatsoever to "knowledge" not derived through the accepted methods of science. But the *Cours de philosophie positive* made its point not by dialectic but by an appeal to the history of thought, and here Comte presented his two basic ideas:

Comte's *Philosophie positive*

1. The notion that the sciences have emerged as sciences in strict order, beginning with mathematics and astronomy, being followed by physics, chemistry, and biology in that order, and culminating in the new science of sociology, to which Comte was the first to ascribe the name;
2. The so-called "law of the three stages," which views thought in every field as passing progressively from superstition to science by first being (*a*) religious, then (*b*) abstract, or metaphysical, and finally (*c*) positive, or scientific. Comte's permanent contribution was to initiate an antireligious and an antimetaphysical bias in the philosophy of science that has passed into the 20th century.

In mid-19th-century England the chief representative of the Empirical tradition from Bacon to Hume was John Stuart Mill. Mill's theory of knowledge, best presented in his *Examination of Sir William Hamilton's Philosophy* (1865), was not particularly original but rather a judicious combination of the doctrines of Berkeley and Hume; but it symbolized his mistrust of vague metaphysics, his denial of the a priori element in knowledge, and his determined opposition to any form of intuitionism. It is in his enormously influential *System of Logic* (1843), however, that Mill's chief theoretical ideas are to be found.

This work, as part of its subtitle, *the Principles of Evidence and the Methods of Scientific Investigation,* shows, was concerned less with formal logic than with scientific methodology. Mill made here the fundamental distinction between deduction and induction, defined induction as the process by which men discover and prove general propositions, and presented his "four methods of experimental inquiry" as the heart of the inductive method. These methods were, in fact, only an enlarged and refined version of Francis Bacon's "tables of discovery." But the most significant section of *A System of Logic* was its conclusion: book 6, "On the Logic of the Moral Sciences."

Mill had taken men's experience of the uniformity of nature as the warrant of induction. Here he reaffirmed the belief of Hume that it is possible to apply the principle of causation and the methods of physical science to moral and social phenomena. These may be so complex as to yield only "conditional predictions," but in this sense there *are* "social laws." Thus Comte and Mill agreed upon the possibility of a true social science.

Mill's *Logic* was extremely influential, and it continued to be taught at Oxford and Cambridge well into the 20th century; but in the end his importance lay less in logic and theory of knowledge than in ethics and political theory. For Mill was the great apostle of political liberalism in the 19th century, a true follower of John Locke. And, just as Locke and Rousseau had represented the liberal and the radical wings of social theory in the early modern period, so Mill and Karl Marx represented the liberal and the radical approaches to social reform 100 years later.

Mill's political liberalism and ethics

Raised by social reformers (his father James Mill and Jeremy Bentham) to be a social reformer himself, Mill's social theory was an attempt, by gradual means arrived at democratically, to combat the evils of the Industrial Revolution. His ethics, expressed in his *Utilitarianism* (1861), followed the formulations of Bentham in finding the end of society to consist in the production of the greatest quantity of happiness for its members, but he gave to Bentham's cruder (but more consistent) doctrines a humanistic and individualistic slant. Thus, the moral self-development of the individual becomes the ultimate value in Mill's ethics.

This trend was also expressed in his essay *On Liberty* (1859) and *Considerations on Representative Government* (1861). In the former he stated the case for the freedom of the individual against "the tyranny of the majority," presented strong arguments in favour of complete freedom of thought and discussion, and argued that no state or society has the right to prevent the free development of human individuality. In the latter he pro-

vided a classic defense for the principle of representative democracy, asked for the adequate representation of minorities, urged renewed public participation in political action for necessary social reforms, and pointed up the dangers of class-oriented or special-interest legislation.

Marx and social revolution

A radical counterbalance to Mill's liberal ideas was provided by Karl Marx, a German philosopher, social revolutionary, and political economist. Taking over from Hegel the idea of estrangement (which Hegel had used in a metaphysical sense), Marx used this notion prior to 1848 to indicate the alienation of the worker from the enjoyment of the products of his work, the crass treatment of human labour as a mere commodity (and man as a thing), and, in fact, the general dehumanization of man in a selfish, profit-seeking capitalist society.

In the famous *Communist Manifesto* (1848), Marx, yielding to the revolutionary temper of the times, called (as Rousseau had done before the French Revolution) for the violent overthrow of the established order. All of history, Marx said, is the struggle between exploiting minorities and the underlying population, between bourgeois and proletarians; and he advocated the formation of a Communist Party to stimulate proletarian class consciousness toward the seizure of power and the institution of a just and democratically managed Socialist society.

Marx's revolutionary fervour may have tended to dampen his philosophic reputation in the West, and his philosophic achievement remains a controversial point; but certain of his ideas (some Hegelian in inspiration, some more original) have endured. Among these are the ideas:

1. That society is a moving balance (dialectic) of antithetical forces that produce social change;
2. That there is no conflict between a rigid economic determinism and a program of revolutionary action;
3. That ideas (including philosophic theories) are not purely rational and thus independent of external circumstance but depend upon the nature of the social order in which they arise.

Independent and irrationalist movements. The end of the 19th century saw a flowering of many independent philosophical movements. After Hegel was almost forgotten in Germany, his influence migrated to England, where T.H. Green, F.H. Bradley, and Bernard Bosanquet initiated a new Hegelian renaissance. Bradley's *Appearance and Reality* (1893) constituted the high-water mark of the rediscovery of Hegel's dialectical method. In America, a strong reaction against Idealism led to the beginnings of the Pragmatic movement initiated by Charles Sanders Peirce and William James. Peirce was an exact logician who recognized that the function of all inquiry is to eradicate doubt and that the meaning of a concept consists in the practical consequences that the concept might be said to have. James transferred Peirce's pragmatic theory of meaning into a pragmatic theory of truth and, in his *Will to Believe* (1897), asserted that men have a right to believe even in the face of inconclusive evidence and that, since knowledge is an instrument for the sake of life, the true test of a belief is the practical consequences that it entails. Meanwhile, in Austria, Franz Brentano, who taught at Vienna from 1874 to 1895, and Alexius Meinong, who taught at Graz, were developing an empirical psychology and a theory of objects, which were to have considerable influence upon the new philosophy of Phenomenology.

Irrationalism in Kierkegaard and Schopenhauer

It was not, however, any of these late-19th-century developments but, rather, the emphasis upon the irrational, which started almost at the century's beginning, that gave the philosophy of the period its peculiar flavour. Hegel, despite his commitment to systematic metaphysics, had nonetheless carried on the Enlightenment tradition of faith in human rationality. But soon his influence was challenged from two different directions. Just before midcentury, the Danish Christian thinker Søren Kierkegaard challenged the logical pretensions of the Hegelian system; and one of his contemporaries, Arthur Schopenhauer, himself a German Idealist and constructor of a bold and imaginative system, contradicted Hegel by asserting that the irrational is the truly real.

Kierkegaard's criticism of Hegel was an appeal to the concrete as against the abstract. He satirized Hegelian Rationalism as a perfect example of "the academic in philosophy"—of detached, objective, abstract theorizing and system building, blind to the realities of human existence, to its subjective, living, emotional character. What a man requires in life, said Kierkegaard, is not infinite inquiry but the boldness of resolute decision and commitment. Man's essence is not contained in thinking but in the existential conditions of his emotional life, in his anxiety and despair. The titles of three of Kierkegaard's books—*Frygt og Baeven* (1843; Eng. trans., *Fear and Trembling*, 1939), *Begrebet Angest* (1844; Eng. trans., *The Concept of Dread*, 1944), *Sygdommen til Døden* (1849; Eng. trans., *The Sickness unto Death*, 1941)—indicate his preoccupation with states of consciousness quite unlike the usual philosophic concentration on cognition and the validity of knowledge.

Schopenhauer, even though, for a short time, he competed unsuccessfully with Hegel at the University of Berlin, soon withdrew and thereafter waged a lifelong battle against academic philosophy. His own system, though orderly and carefully worked out, was written in a vivid and engaging literary style. Schopenhauer agreed with Kant that the world of appearances, of phenomena, is governed by the conditions of space, time, and causality. But he held that science, which investigates this world, cannot itself penetrate the real world behind appearances, which is dominated by a strong, blind, striving, universal cosmic Will that expresses itself in the vagaries of human instinct, in sexual striving, and in the wild uncertainties of all animal behaviour. Everywhere in nature one sees strife, conflict, and inarticulate impulse; and these, rather than rational processes or intellectual clarity, are man's true contacts with ultimate reality.

Schopenhauer's great work *Die Welt als Wille und Vorstellung* (Eng. trans., *The World As Will and Idea*, 1883) was published in 1819; and Kierkegaard's uneven masterpieces appeared between 1843 and 1849. But Friedrich Nietzsche, the third of the irrationalist triumvirate, wrote between 1872 and 1889. A prolific but unsystematic writer, presenting his patchwork of ideas in swift atoms of thought, Nietzsche saw the task of the philosopher as that of a man who destroys old values, creates new ideals, and through them a new civilization. He agreed with Schopenhauer that mind is an instrument of instinct to be used in the service of life and power, and he held that illusion is as necessary to man as truth. Nietzsche spent much time in the analysis of such states as resentment, guilt, bad conscience, and self-contempt.

Nietzsche's Dionysian irrationalism

Kierkegaard, Schopenhauer, and Nietzsche provided for the 19th century a new, nonrational conception of human nature, and they viewed the mind not with the rational clarity of Locke and Hume but as something dark, obscure, hidden, deep. But, above all, they initiated a new style of philosophizing. Schopenhauer wrote like an 18th-century essayist; Kierkegaard was a master of the methods of irony and paradox; and Nietzsche used aphorism and epigram with conscious literary intention. For these three, the philosopher should be less a crabbed academician than a man of letters.

THE 20TH CENTURY

Despite the tradition of philosophic "professionalism" established during the Enlightenment by Wolff and Kant, philosophy in the 19th century was still largely created by men outside of the universities. Comte, Mill, Marx, Kierkegaard, Schopenhauer, and Nietzsche were not professors. Only the German Idealist school was rooted in academic life. Today, however, it is difficult to find one first-rate philosophic mind outside of the academy (George Santayana after his early years and Jean-Paul Sartre are the notable exceptions). The agencies of professionalism—the philosophic congress, the national philosophical societies, the narrowly specialized journals—have become increasingly important. Philosophers more and more employ a technical vocabulary and deal with specialized problems; and they write not for a broad intellectual public but for one another.

Professionalism also has: (1) sharpened the divisions

between philosophic schools and (2) made the definition of what philosophy is and ought to be, a matter of the sharpest controversy. Philosophy has become extremely self-conscious about its own method and nature. Moreover, intellectual competitiveness has been further sharpened by geographical and political considerations. The tradition of clear analysis, inaugurated by Locke and Hume, still dominates the Anglo-Saxon world; the Hegelian tradition of speculative metaphysics and the Nietzschean tradition of bold psychological generalization still dominate the Continent. And the very special framework of thinking established by Karl Marx still dominates philosophy in the Socialist world.

Individual philosophies of Bergson, Dewey, and Whitehead. Except, then, for philosophical Marxism and a few individual philosophers such as Henri Bergson, John Dewey, and A.N. Whitehead, who evade easy classification, the main currents of contemporary philosophy are: (1) Analytic philosophy with its two chief branches, (*a*) Logical Empiricism and (*b*) Linguistic Analysis; and (2) Continental philosophy with its two branches, (*a*) Existentialism and (*b*) Phenomenology.

Henri Bergson, John Dewey, and Alfred North Whitehead have been called the three most important speculative philosophers of the first half of the 20th century. The French Bergson and the American Dewey shared an instrumental theory of mind; *i.e.*, influenced by Darwinism, they both held that mind has emerged in the course of biological evolution as an instrument of man's adaptation to his environment and, therefore, that man's mental functions are chiefly utilitarian agencies of action. Bergson and the English Whitehead were both adherents of a process philosophy; *i.e.*, each held that actuality is, like a river, fluent, that ultimate reality is mobile and dynamic. Yet, in fact, each of the three defined philosophy differently and saw it as performing a uniquely different task. Bergson's vision of philosophy was intuitional, Whitehead's was speculative, and Dewey's was pragmatic.

Bergson's two ways of knowing

Bergson, in the *Introduction à la métaphysique* (1903) and in his masterpiece, *L'Évolution créatrice* (1907), distinguished between two profoundly different ways of knowing: the method of analysis, which is that of science; and the method of intuition, which is just that intellectual sympathy by which men enter into objects and persons and identify with them. All basic metaphysical truths, Bergson held, are grasped by philosophical intuition. This is how men know their deepest selves, the duration, which is the essence of all living things, and the vital spirit, which is the mysterious creative agency in the world.

For Whitehead, philosophy was primarily metaphysics (or "speculative philosophy"), the effort "to frame a coherent, logical, necessary system of general ideas in terms of which every element of our experience can be interpreted." Whitehead's philosophy was, thus, an attempt to survey the world with a large generality of understanding, an end toward which his great trilogy, *Science and the Modern World* (1925), *Process and Reality* (1929), and *Adventures of Ideas* (1933), was directed.

While Bergson and Whitehead were principally metaphysicians and philosophers of culture, Dewey was an all-around philosopher who dominated the fields of ethics, metaphysics, and methodology in the United States for many years and who stressed the unity, interrelationship, and organicity of all forms of philosophic knowledge. He is chiefly notable, however, because his notion of philosophy stressed so powerfully the conceptions of practicality and moral purpose. One of the guiding aims of Dewey's philosophizing was the effort to find the same warranted assertibility for ethical and political as for scientific judgments. Philosophy, he said, should be oriented not to professional pride but to human need.

The social thought of Mill and Marx presented opposed solutions to the social problems of the 19th century; that of Dewey and Lenin presented opposed solutions to the problems of today. Not revolution, but the continuous application of the method of intelligence to social affairs was Dewey's answer. He believed in social planning, in conscious intelligent intervention to produce desirable social change; and he proposed a new "experimentalism" in social affairs as the guide to enlightened public action to promote the aims of a democratic community. His pragmatic social theory is the major social philosophy that liberal democracy has produced in the modern world.

Marxist thought. The framework of 19th-century Marxist thought, augmented by the philosophical suggestions of Lenin, still serves as the starting point of all philosophizing in the Socialist countries of eastern Europe. To Marx's conception of the dialectic (that the social world progresses through the rise and fall of oppositions), Lenin added a strong tinge of metaphysical Materialism (material things and their systems of relations exist independently of mind, and the laws of nature are objective) and a naïvely Realist theory of knowledge (ideas and sensations are true copies of independently existing things). But much of Lenin's thinking was also devoted to rather practical issues like the tactics of violence and the role of the Communist Party in bringing about and consolidating the proletarian revolution.

Unity of theory and practice in Marxism

Later Marxism has continued this concern with practical issues, because it represents the basic Marxist conception of what philosophy is and ought to be. For Marxism (like Pragmatism) tries to assimilate theoretical issues to practical needs. It asserts the basic unity of theory and practice by finding that the function of the former is to serve the latter. Marx and Lenin both held that theory is always in fact expressive of class interests; consequently, they wished philosophy to be transformed into a tool for furthering the class struggle: not abstractly to discover the truth but concretely to forge the intellectual weapons of the proletariat is philosophy's dominant task. Thus philosophy becomes inseparable from ideology.

Since the discovery and publication in 1927 and 1932 of some of Marx's early writings, many have sensed a breath of fresh air in Marxist philosophizing outside the Soviet Union. For the concern of the young Marx with alienation, with a theory of the nature of man, and with moral values has given Marxism a new humanistic dimension and has inspired some notable revisionist thinking in Poland, Hungary, and particularly in Yugoslavia.

Analytic philosophy. *Logical Positivism.* Logical Positivism, or Logical Empiricism, the earliest branch of modern Analytic philosophy was jointly inspired by Hume and by the new logic of Russell and Whitehead, authors of the *Principia Mathematica* (3 vol.; 1910–13). The school, formally instituted at the University of Vienna in a seminar of Moritz Schlick in 1923, continued there as the Vienna Circle until 1938. Its great period began in 1926 when Rudolf Carnap, who became a leading semanticist and philosopher of science, arrived at the University. Its manifesto of 1929, *Wissenschaftliche Weltauffassung: Der Wiener Kreis* ("Scientific Conception of the World: The Vienna Circle"), and its journal *Erkenntnis* ("Knowledge"), founded a year later, marked its self-consciousness as a philosophical movement.

Logical Empiricism's basic contribution has been a profound alteration in the conception of the role that philosophy itself must play. Philosophy, it claimed, must henceforth be scientific. It should seek less a content than a function: it should produce not complicated pictures of the world but clear thinking. This was best stated by Ludwig Wittgenstein in his *Tractatus Logico-Philosophicus* (1922), first published as *Logische-philosophische Abhandlung* (1921):

> The object of philosophy is the logical clarification of thoughts. Philosophy is not a theory but an activity. A philosophical work consists essentially of elucidations. The result of philosophy is not a number of 'philosophical propositions,' but to make propositions clear.

Despite this emphasis upon philosophy as a pure activity, Logical Empiricism did propound, at the same time, a series of revolutionary theses:

1. All *meaningful* discourse consists either of (*a*) the formal sentences of logic and mathematics or (*b*) the factual propositions of the special sciences;
2. Any assertion that claims to be factual has meaning only if it is possible to say how it might be verified;
3. Metaphysical assertions, coming under neither of the two classes of (1), are *meaningless*;

4. All statements about moral, aesthetic, or religious values are scientifically unverifiable and meaningless.

Logical Empiricism's radical denial of the meaningfulness of metaphysics and of assertions of value at first produced something of a philosophical scandal. But meanwhile, Bertrand Russell and his student Wittgenstein (who had emigrated from Austria) were propounding similar doctrines in England; and, after the Nazis overran Austria and Carnap, Hans Reichenbach (a Berlin Positivist), and many others from the Vienna Circle moved to America, this philosophy proved remarkably influential in the Anglo-American world. The Logical Empiricists were mainly interested in three basic themes: logic, language, and perception. Whereas Russell began with logic, turned to problems of perception, and ended with semantics, Carnap, having begun with perception in *Der logische Aufbau der Welt* (1928; *The Logical Structure of the World*, 1967), turned to problems of semantics in *Logische Syntax der Sprache* (1934; Eng. trans., *The Logical Syntax of Language*, 1937) and ended with logic in *Meaning and Necessity* (1947). But of these three themes, it was that of language that proved most enduring. This emphasis Logical Empiricism shared with its successor, Linguistic Analysis.

Linguistic Analysis. Sometimes called "ordinary language" Analysis or even "Oxford philosophy" from its present stronghold, this movement was, in fact, largely the product of two philosophers of the first half of the century who were associated with the University of Cambridge, G.E. Moore and Wittgenstein. Moore made the examination of the assertions of other men philosophically popular—*i.e.*, the practice of directing one's critical energies to the mistakes of his contemporaries. The major subjects of his interest were ethics and the theory of knowledge; his thought was always Realistic and commonsensical, and he introduced into philosophy an unbelievably precise and closely applied analytical method. It was Moore's passion for clarity and his infinite pains "to get everything *exactly* right" that served as moral inspiration to a succeeding generation of younger men who had attended his Cambridge lectures over a span of 30 years.

Seminal influence of Moore and Wittgenstein

Wittgenstein had begun as a kind of adjunct to the Vienna Circle. There was a time when his philosophical position was much like that of Russell and Carnap; but as he later became more skeptical of the foundations of mathematics and logic, his interest turned from logic and artificial language systems toward a critical examination of ordinary natural language. This change was principally registered in his *Philosophical Investigations*, posthumously published in 1953, which has become the true bible of Linguistic Analysis. For it is here that Wittgenstein showed how a man's entire world is constituted by his linguistic experience and suggested that "all philosophy is critique of language." Wittgenstein thought that to ask "Why do we use this particular word or expression?" was the crucial philosophical question, since a focussing of philosophy not upon the world but upon the mechanisms of linguistic use would solve most of the perplexities that have plagued philosophizing. Thus Linguistic philosophy was to perform a therapeutic function.

Wittgenstein's example found many followers in the United States and, particularly, in England. There philosophers Gilbert Ryle, J.L. Austin, and P.F. Strawson have examined "category mistakes," "systematically misleading expressions," "mental-conduct concepts" (such as "other minds"), and all of the major "perception words" such as "looking," "seeing," "sensing," "feeling," and the like. And the effect of this concentration not upon things or ideas but upon words has, for members of this school, turned the customary works on ethics, aesthetics, or the philosophy of religion into treatises entitled "The Language of Ethics" or "The Language of Aesthetics" or "The Language of Religion." And these efforts are all, in Wittgenstein's terms, "a battle against the bewitchment of our intelligence by means of language."

Continental philosophy. The Analytic philosophy that, by the mid-1970s, had dominated Anglo-American philosophical thinking for almost half a century has had comparatively little influence on the continent of Europe. For there the metaphysical and speculative tradition remained strong; and in the 1950s and 1960s the interest in both, but particularly in Phenomenology, posed an increasing threat to philosophical Analysis in Britain and the United States. Continental philosophy may be divided into Phenomenology and Existentialism, but the division is not sharp. Thinkers such as Martin Heidegger, Jean-Paul Sartre, and Maurice Merleau-Ponty combine the two strands. And the phenomenological method has been one of the formative forces in Existentialism.

Phenomenology of Husserl and others. Edmund Husserl, a German mathematician turned philosopher, was the true father of Phenomenology. An extremely complicated and technical thinker, whose views changed considerably over the years, his chief contributions were the phenomenological method, developed early, and the concept of the life-world, appearing only in his later writings. Husserl proposed the phenomenological method as a technique for conducting phenomenological analyses —*i.e.*, for making possible "a descriptive account of the essential structures of the directly given." The emphasis here is upon the immediacy of experience, the attempt to isolate it and set it off from all assumptions of existence or causal influence and to lay bare its actual intrinsic structure; for it is this structure that constitutes its essence. Phenomenology restricts the philosopher's attention to the pure data of consciousness uncontaminated by metaphysical theories or scientific assumptions. Husserl's concept of the life-world expressed this same idea of immediacy. It is the individual's personal world as directly experienced, with the ego at the centre and with all of its vital and emotional colourings.

Husserl's phenomenological method and life-world

With the appearance of the *Jahrbuch für Philosophie und Phänomenologische Forschung* (1913–30), under Husserl's chief editorship, his personal philosophizing flowered into an international movement. Its two most notable adherents were Max Scheler, a scholar of life-philosophy, whose *Formalismus in der Ethik und die materiale Wertethik* was published in the *Jahrbuch* (1913–16), and the ontological Existentialist Martin Heidegger, whose famous *Sein und Zeit* (Eng. trans., *Being and Time*, 1962) appeared in the *Jahrbuch* in 1927. Despite the differences in their individual philosophies, their work showed clearly the influence of the phenomenological method: in Scheler's careful explorations of the role that emotion plays in men's moral values and in his descriptive analyses of the human attitudes of resentment, sympathy, and love; and in Heidegger's startlingly original investigations of human existence with its unique dimensions of being-in-the-world, dread, care, and being toward death. In all of these painstaking and enormously detailed inquiries, however, both Scheler and Heidegger held to the phenomenological principle that philosophy is not empirical but is the strictly self-evident insight into the structure of experience. Later, Maurice Merleau-Ponty, a French student of philosophical psychology, in his *Phénoménologie de la perception* (1945), took over Husserl's conception of the life-world and used the notions of the lived body and its facticity to build up a hierarchical order of man's lived experience.

Scheler, Heidegger, and Merleau-Ponty

Existentialism of Jaspers and Sartre. Continental Existentialism, true to its roots in Kierkegaard and Nietzsche, is oriented about two major themes: (1) the analysis of human Being and (2) the centrality of human choice; thus its chief theoretical energies are devoted to ontology and decision. Existentialism as a philosophy of human existence finds its best spokesman in the German Karl Jaspers, who came to philosophy from medicine and psychology. As a philosophy of human decision, its spokesman is the French philosopher and playwright Jean-Paul Sartre. These two preoccupations lead to two different conceptions of philosophy's function.

For Jaspers, as for Dewey, the aim of philosophy is practical; for Dewey, however, it guides human doing, whereas for Jaspers its purpose is the achievement of human Being. Philosophizing is an inner activity by which the individual finds and becomes himself: it is a *revelation of Being*. It is an attempt to answer the question of what man is and what he can become; and this

Human existence and choice in Jaspers and Sartre

activity, wholly unlike that of science, is one of mere thought through the "inwardness" of which a man becomes aware of the deepest levels of Being.

But if, for Jaspers, philosophy is devoted to "the illumination of existence," that illumination is achieved when a person recognizes that human existence is revealed most profoundly in his experience of those "extreme" situations that define the human condition—conflict, guilt, suffering, and death. It is in man's confrontation with these extremes that he achieves his existential humanity.

Sartre, too, has some concern for man's Being and his dread before the threat of Nothingness. But Sartre finds the essence of this Being in man's liberty—in his duty of self-determination and his freedom of choice—and therefore spends much time in describing the human tendency toward "bad faith"—reflected in man's perverse attempts to deny his own responsibility and to flee from the truth of his inescapable freedom. Sartre, like Schopenhauer and Nietzche, considers himself a man of letters and has written novels and plays that assert the Existential dogma of human freedom and explore the inexhaustible mechanisms of bad faith. It is not that Sartre overlooks the legitimate series of obstacles that freedom has to face in the givens of one's place, past, environment, fellowmen, and death; but his Existentialist bias demands that man surmount these limitations by acts of conscious decision. For only in acts of freedom does human existence achieve authenticity.

Concluding comments. In generally surveying the field of contemporary philosophy, it seems clear that, in the 50 years around the mid-20th century, philosophizing has been uncommonly concerned with problems of language, symbolism, and communication. Renaissance philosophy was primarily preoccupied with nature, with an external physical world and the objects contained within it. The Enlightenment turned to the mind that knows the world; its philosophy spoke of the genesis of ideas, the relation of concepts, the quality of appearances. But current philosophizing largely states its problems in terms of symbolic manipulation and the characteristics of words.

Nor is this merely the preoccupation of Analytical philosophy: Dewey was interested in social communication; Whitehead wrote a little volume on *Symbolism: Its Meaning and Effect* (1927); Heidegger has turned to poetry and the etymology of words for the revelation of Being; Ernst Cassirer, a German Neo-Kantian, has produced a "philosophy of symbolic forms"; and Jaspers has tried to decipher the meanings reflected in human speech and gesture.

Apart from this underlying tendency, there seems little prospect for philosophical unification. The scientific and the metaphysical tempers still pursue their opposite courses, and the subjectivity of Existentialism and the objectivity of Logical Positivism still express their opposition in mutual contempt. Thus, in the contemporary philosophical universe, multiplicity and division still reign. (A.W.L.)

BIBLIOGRAPHY

General histories: G.W.F. HEGEL, *Vorlesungen über die Geschichte der Philosophie*, 3 vol. (1840–44; Eng. trans., *Lectures on the History of Philosophy*, 1892–96); FRIEDRICH UEBERWEG, *Grundriss der Geschichte der Philosophie*, 12th ed., 5 vol. (1923–28; Eng. trans. from 4th ed., *A History of Philosophy from Thales to the Present Time*, 2 vol., 1872–74); WILHELM WINDELBAND, *Lehrbuch der Geschichte der Philosophie*, 14th ed. with supplement "Die Philosophie im zwanzigsten Jahrhundert" by HEINZ HEIMSOETH (1948; Eng. trans., *A History of Philosophy*, 1905, with a brief notice on English thought replacing Heimsoeth's supplement), a problems approach, particularly good on Kant and German Idealism; ALFRED WEBER, *Historie de la philosophie européenne*, 7th ed., 3 vol. (1905; Eng. trans., *History of Philosophy*, 1925, with supplement by RALPH B. PERRY); FRANK THILLY, *A History of Philosophy*, 3rd ed. rev. by LEDGER WOOD (1957); EMILE BREHIER, *Histoire de la philosophie*, new ed., 2 vol. (1941–47), particularly good on French thinkers, less so on English; JULIAN MARIAS, *Historia de la filosofía*, 22nd ed. (1966; Eng. trans., 1967); BERTRAND RUSSELL, *A History of Western Philosophy* . . . (1946); W.T. JONES, *A History of Western Philosophy*, 2nd ed. (1969); F.C. COPLESTON, *History of Philosophy*, 8 vol. (1946–66); JOHN HERMAN RANDALL, *The Career of Philosophy*, 2 vol. (1962–65), an informed, comprehensive, and judicious treatment. The best source for bibliographic information in philosophy is the annual *Répertoire bibliographique de la philosophie*, which the Institut Supérieur de Philosophie at Louvain began publishing in 1949.

Ancient Greek and Roman philosophy: (*Studies*): The most detailed work on the whole course of Greek and Roman philosophy is still EDUARD ZELLER, *Die Philosophie der Griechen*, 7th ed., 3 pt. (1920–23; Eng. trans. from earlier ed., *History of Greek Philosophy*, 2 vol., 1881–97). Still more detailed and up to date is the Italian translation of Zeller's first two volumes by RUDOLF MONDOLFO, with a wealth of additional notes and discussions. Equally thorough and up to date will be the great work by W.K.C. GUTHRIE, *A History of Greek Philosophy* (1962–), when completed; 3 vol. have appeared to date. For short introductions to Greek philosophy in English, see M.E.J. TAYLOR, *Greek Philosophy* (1924); REX WARNER, *The Greek Philosophers* (1958); and the excellent one by W.K.C. GUTHRIE, *The Greek Philosophers from Thales to Aristotle* (1960). For somewhat more detailed ones, see (in addition to the above work by Copleston) CORNELIA J. DE VOGEL (ed.), *Greek Philosophy*, 3 vol. (1950–59). (*Collections of fragments and works*): An influential source from perhaps the 3rd century AD is DIOGENES LAERTIUS, *Peri biōn dogmatōn kai apophthegmatikōn tōn en philosophia eudokimēsantōn* (Eng. trans., *Diogenes Laërtius: Lives of Eminent Philosophers*, 2 vol., 1925). The best comprehensive collection of the fragments of the pre-Socratic philosophers is still HERMANN DIELS and WALTHER KRANZ, *Die Fragmente der Vorsokratiker*, 11th ed., 3 vol. (1961–64), made more readily accessible for English-speaking readers by KATHLEEN FREEMAN, *Ancilla to the Pre-Socratic Philosophers: A Complete Translation of the Fragments* . . . (1948). A good selection of texts from the whole history is CORNELIA J. DE VOGEL, *Greek Philosophy: A Collection of Texts, Selected and Supplied with Some Notes and Explanations*, 3 vol. (1950–59). A more comprehensive edition of the main extant works and fragments of Epicurus is CYRIL BAILEY, *Epicurus: The Extant Remains* (1926), with short critical apparatus, translation, and notes. The works of Plato and Aristotle have been edited with English translations in the "Loeb Classical Library." (*Bibliographies*): The most detailed bibliography to 1925 is in KARL PRAECHTER, *Die Philosophie des Altertums*, 12th ed. (1926, reprinted 1953); the most complete bibliography from 1926 to 1963 is WILHELM TOTOK, *Handbuch der Geschichte der Philosophie*, vol. 1, *Altertum* (1964), 353 pages of bibliography, including ancient Indian philosophy.

Medieval philosophy: MAURICE DE WULF, *Histoire de la philosophie médiévale*, 6th ed., 3 vol. (1934–47; Eng. trans., 3rd ed., 2 vol., 1935–38, with reissue of vol. 1, 1952), contains much valuable information but has been superseded by more recent histories. ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955), is the best account of medieval philosophy. AIME FOREST, FERNAND VAN STEENBERGHEN, and MAURICE DE GANDILLAC, *Le Mouvement doctrinal du IX^e^ au XIV^e^ siècle* (1951), traces doctrinal developments from the 9th to the 14th century. FERNAND VAN STEENBERGHEN, *La Philosophie au XIII^e^ siècle* (1966), gives a different interpretation of 13th-century philosophy from that of Gilson. His *Aristotle in the West* (1955), is a valuable account of the introduction of Aristotle's works into western Europe. DAVID KNOWLES, *The Evolution of Medieval Thought* (1962), is the work of an eminent historian of medieval religion. ARMAND A. MAURER, *Medieval Philosophy* (1962), sketches medieval philosophy from Augustine to the Renaissance. The following are short accounts of philosophy in the Middle Ages: GORDON LEFF, *Medieval Thought: St. Augustine to Ockham* (1958); PAUL VIGNAUX, *Philosophie au moyen âge*, 3rd ed. (1958; Eng. trans., *Philosophy in the Middle Ages*, 1959); JULIUS R. WEINBERG, *A Short History of Medieval Philosophy* (1964). Useful information about medieval Arab and Jewish philosophy is contained in T.J. DE BOER, *The History of Philosophy in Islam* (Eng. trans. 1903); GOFFREDO QUADRI, *La Philosophie arabe dans l'Europe médiévale* (French trans. 1947); ISAAC HUSIK, *A History of Mediaeval Jewish Philosophy* (1916, reprinted 1969); and GEORGES VAJDA, *Introduction à la pensée juive du moyen âge* (1947). Collections of translated texts include RICHARD MCKEON, *Selections from Medieval Philosophers*, 2 vol. (1929–30); HERMAN SHAPIRO, *Medieval Philosophy: Selected Readings from Augustine to Buridan* (1964); ARTHUR HYMAN and JAMES J. WALSH (eds.), *Philosophy in the Middle Ages: The Christian, Islamic, and Jewish Traditions* (1967); and JOHN F. WIPPEL and ALLAN B. WOLTER (eds.), *Medieval Philosophy, from St. Augustine to Nicholas of Cusa* (1969).

Modern and contemporary philosophy: (Comprehensive classical studies): JOHANN EDUARD ERDMANN, *Versuch einer wissenschaftlichen Darstellung der Geschichte der neuern Philosophie*, 3 vol. (1834–53; Eng. trans., *A History of Philosophy*, 3 vol. 1891–93); KUNO FISCHER, *Geschichte der neuern Philosophie*, 5th ed., 2 vol. (1909–12; partial Eng. trans., *History of Modern Philosophy: Descartes and His School*, 1887); HARALD HOFFDING, *Den nyere Filosofis Historie: En Fremstilling af Filosofiens Historie fra Renaissancens Slutning til vore Dage*, 2 vol. (1894–95; Eng. trans., *A History of Modern Philosophy*, 2 vol., 1900, reprinted 1955). *(Contemporary works)*: RAYMOND KLIBANSKY (ed.), *Philosophy in the Mid-Century: A Survey* (1958), and *Contemporary Philosophy* (1968), broad coverage in essays by leading scholars; ALBERT WILLIAM LEVI, *Philosophy and the Modern World* (1959), a broad treatment; WOLFGANG STEGMULLER, *Hauptströmungen der Gegenwartsphilosophie*, 4th ed. (1969; Eng. trans., *Main Currents in Contemporary German, British, and American Philosophy*, 1969), narrower, more technical treatment; WALTER KAUFMANN (ed.), *Existentialism from Dostoevsky to Sartre* (1956), selections with introductions; MARVIN FARBER (ed.), *L'Activité philosophique contemporaine en France et aux États-Unis*, 2 vol. (1950; Eng. trans., *Philosophic Thought in France and the United States: Essays Representing Major Trends in Contemporary French and American Philosophy*, 2nd ed., 1968). *(Chiefly Anglo-American)*: JOHN PASSMORE, *A Hundred Years of Philosophy*, 2nd ed. (1966); G.J. WARNOCK, *English Philosophy Since 1900*, 2nd ed. (1969); HERBERT W. SCHNEIDER, *A History of American Philosophy*, 2nd ed. (1963), standard reference work with excellent bibliographies. *(Chiefly continental)*: INNOCENTIUS M. BOCHENSKI, *Europäische Philosophie der Gegenwart*, 2nd ed. (1951; Eng. trans., *Contemporary European Philosophy*, 1956); MICHELE FEDERICO SCIACCA, *La filosofia oggi*, 3rd ed., 2 vol. (1958; Eng. trans., *Philosophical Trends in the Contemporary World*, 1964); HERMANN NOACK, *Die Philosophie Westeuropas im zwanzigsten Jahrhundert* (1962); HERBERT SPIEGELBERG, *The Phenomenological Movement*, 2nd ed., 2 vol. (1965–66); ANNA-TERESA TYMIENIECKA, *Phenomenology and Science in Contemporary European Thought* (1962). For specific individuals and schools of philosophy, see the bibliographies of relevant articles.

(Kv.Fr./A.A.Ma./A.W.L.)

Phonetics

Phonetics is the study of pronunciation. The subject has two main branches—the description of the way in which the speech organs are used to produce the sounds of speech, and the specification of the resulting sound waves. A great deal of attention is also given to studying how languages and the dialects within a language differ in the sounds they use. Phonetics is thus part of linguistics, which is the study of human language as a whole, but its subject matter also overlaps with that of other disciplines such as physiology and acoustics.

Phonetics is a necessary part of linguistics, not only for adequately describing different languages but also for explaining the processes of word and sentence formation, and for understanding historical sound changes. Some knowledge of the subject is also useful for speech pathologists and for teachers who are concerned with the teaching of pronunciation. Because the movements of the vocal organs are the most rapid, intricate, and precisely controlled of all human muscular activities, knowledge of the processes that are involved in their control is also of considerable interest to physiologists and neurophysiologists.

STANDARDS OF PRONUNCIATION

Phoneticians are often expected to be concerned with standards of pronunciation. From a phonetician's point of view, however, everybody has an accent, and there is no such thing as good or bad speech. What is called good speech is simply that style of speech preferred by a particular class or group of people. One form of speech may become accepted as a standard within a region, or even within a country as a whole; however, what is accepted by a certain group may not, in many instances, be acceptable to another.

It should also be remembered that there are no genetic or racial characteristics that determine the way in which a child speaks. As a person grows up, he usually continues to talk in the way he wanted to talk during his formative years, when he was trying to win the approval of those who were important to him; *e.g.*, his parents. An accent thus marks a person as belonging, usually by birth, to a particular group of people. With the exception of those who suffer from disorders of speech, such as cleft palate, there is nothing to prevent anyone from changing his accent, provided he devotes enough effort to the problem. In doing so, a person can make use of phoneticians' descriptions of speech styles used by other groups.

Determinants of pronunciation

Within the English-speaking world there is no single way of speaking that is accepted as the most prestigious, but each country or region does have its own standard forms. In the United States there are several regions—*e.g.*, New England, New York City, the Appalachian Mountains, and Texas—that have characteristic styles of speech. At the same time, there is a growing tendency for news commentators on television and radio, actors, particularly those in commercials, and others who want to reach a wide audience throughout the country to use a style of speech that is comparatively free of local regional characteristics, and that is spoken by a large proportion of the population who live in the Midwest and the Far West. This style is sometimes called General American.

In Britain the situation is different. In Scotland and Wales there are regional accents; but in England, in addition to several local accents, there is an accent known as Received Standard English, or, as a technical term in phonetics, RP (Received Pronunciation, in the sense of received that means acceptable at court). This is spoken all over England and indicates not the regional origin of the speaker but rather that he belongs to a certain social class. Other English-speaking countries, such as Ireland, Canada, Australia, New Zealand, and, in another category, India, Ghana, and Nigeria, also have their own forms of English with their own regional standards.

ARTICULATORY PHONETICS

The traditional method of describing speech sounds is in terms of the movements of the vocal organs that produce them. The main structures that are important in the production of speech are the lungs and the respiratory system, together with the vocal organs shown in Figure 1. The airstream from the lungs passes between the vocal cords, which are two small muscular folds located in the larynx at the top of the windpipe. The space between the

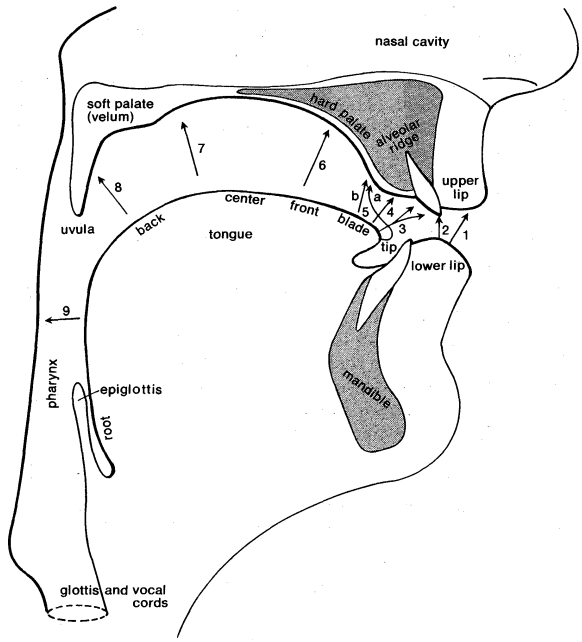


Figure 1: *Location of vocal organs and possible places of articulation.*
(1) Bilabial. (2) Labiodental. (3) Dental and interdental. (4) Alveolar. (5) Postalveolar; (a) retroflex, (b) palato-alveolar. (6) Palatal. (7) Velar. (8) Uvular. (9) Pharyngeal.

vocal cords is known as the glottis. If the vocal cords are apart, as they are normally when breathing out, the air from the lungs will have a relatively free passage into the pharynx (see Figure 1) and the mouth. But if the vocal cords are adjusted so that there is a narrow passage between them, the airstream will cause them to be sucked together. As soon as they are together there will be no flow of air, and the pressure below them will be built up until they are blown apart again. The flow of air between them will then cause them to be sucked together again, and the vibratory cycle will continue. Sounds produced when the vocal cords are vibrating are said to be voiced, as opposed to those in which the vocal cords are apart, which are said to be voiceless.

Voiced and voiceless sounds

The air passages above the vocal cords are known collectively as the vocal tract. For phonetic purposes they may be divided into the oral tract within the mouth and the pharynx, and the nasal tract within the nose. Many speech sounds are characterized by movements of the lower articulators—*i.e.*, the tongue or the lower lip—toward the upper articulators within the oral tract. The upper surface includes several important structures from the point of view of speech production, such as the upper lip and the upper teeth; Figure 1 illustrates most of the terms that are commonly used. The alveolar ridge is a small protuberance just behind the upper front teeth that can easily be felt with the tongue. The major part of the roof of the mouth is formed by the hard palate in the front, and the soft palate or velum at the back. The soft palate is a muscular flap that can be raised so as to shut off the nasal tract and prevent air from going out through the nose. When it is raised so that the soft palate is pressed against the back wall of the pharynx there is said to be a velic closure. At the lower end of the soft palate is a small hanging appendage known as the uvula.

As may be seen from Figure 1, there are also specific names for different parts of the tongue. The tip and blade are the most mobile parts. Behind the blade is the so-called front of the tongue; it is actually the forward part of the body of the tongue and lies underneath the hard palate when the tongue is at rest. The remainder of the body of the tongue may be divided into the centre, which is partly beneath the hard palate and partly beneath the soft palate; the back, which is beneath the soft palate; and the root, which is opposite the back wall of the pharynx.

The major division in speech sounds is that between vowels and consonants. Phoneticians have found it difficult to give a precise definition of the articulatory distinction between these two classes of sounds. Most authorities would agree that a vowel is a sound that is produced without any major constrictions in the vocal tract, so that there is a relatively free passage for the air. It is also syllabic. This description is unsatisfactory in that no adequate definition of the notion syllabic has yet been formulated.

Consonants. In the formation of consonants, the airstream through the vocal tract is obstructed in some way. Consonants can be classified according to the place and manner of this obstruction. Some of the possible places of articulation are indicated by the arrows going from one of the lower articulators to one of the upper articulators in Figure 1. The principal terms that are required in the description of English articulations, and the structures of the vocal tract that they involve are: bilabial, the two lips; dental, tongue tip or blade and the upper front teeth; alveolar, tongue tip or blade and the teeth ridge; retroflex, tongue tip and the back part of the teeth ridge; palato-alveolar, tongue blade and the back part of the teeth ridge; palatal, front of tongue and hard palate; and velar, back of tongue and soft palate. The additional places of articulation shown in Figure 1 are required in the description of other languages. Note that the terms for the various places of articulation denote both the portion of the lower articulators (*i.e.*, lower lip and tongue) and the portion of the upper articulatory structures that are involved. Thus velar denotes a sound in which the back of the tongue and the soft palate are involved, and retroflex implies a sound involving the tip of the tongue and the back part of the alveolar ridge. If it is necessary to distinguish between sounds made with the tip of the tongue and those made with the blade, the terms apical (tip) and laminal (blade) may be used.

Classification of consonants

There are six basic manners of articulation that can be used at these places of articulation: stop, fricative, approximant, trill, tap, and lateral.

Stops. Stops involve closure of the articulators, so that the airstream cannot go out of the mouth. This manner of articulation can be considered in terms of nasal and oral stops. If the soft palate is down so that air can still go out through the nose, there is said to be a nasal stop. Sounds of this kind occur at the beginning of the words *my* and *nigh*. If, in addition to the articulatory closure in the mouth, the soft palate is raised so that the nasal tract is blocked off, then the airstream will be completely obstructed, the pressure in the mouth will be built up, and an oral stop will be formed. When the articulators come apart the airstream will be released with a plosive quality. This kind of sound occurs in the consonants in the words *pie, tie, kye, buy, die,* and *guy*. Many authorities refer to these two articulations as nasals, meaning nasal stops (closure of the articulators in the oral tract implied), and stops, meaning oral stops (raising of the soft palate to form a velic closure implied).

Fricatives. A fricative sound involves the close approximation of two articulators, so that the airstream is partially obstructed and a turbulent airflow is produced. The mechanisms used in the production of these sounds may be compared to the physical forces involved when the wind "whistles" round a corner. Examples are the initial sounds in the words *fie, thigh, sigh* and *shy*. Some authorities divide fricatives into slit and grooved fricatives, or rill and flat fricatives, depending on the shape of the constriction in the mouth required to produce them. Other authorities divide fricatives into sibilants, as in *sigh* and *shy*, and nonsibilants, as in *fie* and *thigh*. This division is based on acoustic criteria (see below).

Varieties of fricatives

Approximants. Approximants are produced when one articulator approaches another but does not make the vocal tract so narrow that a turbulent airstream results. The terms frictionless continuant, semivowel, and glide are sometimes used for some of the sounds made with this manner of articulation. The consonants in the words *we* and *you* are examples of approximants.

Trills. A trill results when an articulator is held loosely fairly close to another articulator, so that it is set into vibration by the airstream. The tongue tip and blade, the uvula, and the lips are the only articulators than can be used in this way. Tongue tip trills occur in some forms of Scottish English in words such as *rye* and *ire*. Uvular trills are comparatively rare but are used in some dialects of French, but not Parisian French. Trills of the lips are even rarer but do occur in a few African languages.

Taps. A tap is produced if one articulator is thrown against another, as when the loosely held tongue tip makes a single tap against the upper teeth or the alveolar ridge. The consonant in the middle of a word such as *letter* or *Betty* is often made in this way in American English. The term flap is also used to describe these sounds, but some authorities make a distinction between taps as defined here and flaps, in which the tip of the tongue is raised up and back and then strikes the alveolar ridge as it returns to a position behind the lower front teeth. Some languages—*e.g.*, Hausa, the principal language of Northern Nigeria—distinguish between words containing a flap and words containing a tap. The distinction between a trill and a tap is used in Spanish to distinguish between words such as *perro*, meaning "dog," and *pero*, meaning "but."

Laterals. When the airstream is obstructed in the midline of the oral tract, and there is incomplete closure between one or both sides of the tongue and the roof of the mouth, the resulting sound is classified as a lateral. The sounds at the beginning and end of the word *lull* are laterals in most forms of American English.

The production of many sounds involves more than one of these six basic manners of articulation. The sounds at the beginning and end of the word *church* are stops combined with fricatives. The articulators—tongue tip or

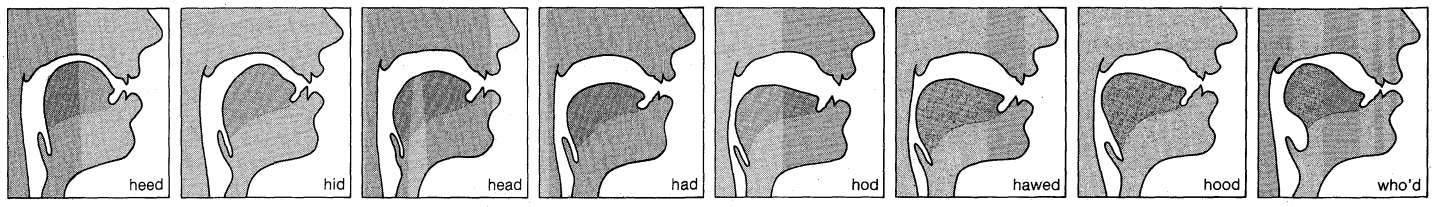


Figure 2: Tongue position for several vowel sounds.

blade, and alveolar ridge—come together for the stop, and then, instead of coming fully apart, they separate only slightly so that a fricative is made at the same place of articulation. This kind of combination is called an affricate. Lateral articulations may also occur in combination with other manners of articulation. The laterals in a word such as *lull* might more properly be called lateral approximants, in that the airstream passes out freely between the sides of the tongue and the roof of the mouth without a turbulent airstream being produced. But in some sounds in other languages the sides of the tongue are closer to the roof of the mouth and a lateral fricative occurs; an example is the sound spelled *ll* in Welsh words such as *llan* "church" and the name *Lluellyn*.

Affricates

Secondary articulations. When an approximant articulation occurs at the same time as another articulation is being made at a different place in the vocal tract, the approximant is said to form a secondary articulation. There are special terms for some of these possibilities. Added lip rounding is called labialization; it occurs in the formation of several English sounds—*e.g.*, during the pronunciation of the palato-alveolar fricative at the beginning of the word *shoe*. Raising of the front of the tongue while simultaneously making another articulation elsewhere in the vocal tract is called palatalization. It is the distinguishing characteristic of the soft consonants in Russian and also occurs, to a lesser extent, in English; *e.g.*, in the first consonant in the word *leaf*. Raising of the back of the tongue to form a secondary articulation is called velarization; it occurs in the last consonant in the word *feel*, which therefore does not contain the same sounds as those in the reverse order in the word *leaf*. Retracting of the root of the tongue while making another articulation is called pharyngealization; it occurs in Arabic in what are called emphatic consonants.

The states of the glottis, places of articulation, and manners of articulation discussed above are sufficient to distinguish between the major contrasts among the consonants of English and many other languages. But additional possibilities have to be taken into account in a more detailed description of English, or in descriptions of several other languages. Among these possibilities are variations in the timing of the states of the glottis. In addition to the contrast between the voiced and voiceless states of the glottis that occur during an articulation, there may be variations in the state of the glottis during the release of the articulation. Thus both the *p* in *pin* and that in *spin* are voiceless bilabial stops, but they differ in that the glottis remains in a voiceless position for a short time after the release of the bilabial stop in *pin*, whereas in *spin* the voicing starts as soon as the lips come apart. When there is a period of voicelessness during the release of an articulation, the sound is said to be aspirated. The main difference between the consonants in *pea* and *bee*, when these words are said in isolation, is not that the one is voiceless and the other voiced, but that the first is aspirated and the second is unaspirated. Some languages distinguish between both voiced–voiceless and aspirated–unaspirated sounds. Thus Thai has contrasts between voiceless aspirated stops, voiceless unaspirated stops, and voiced unaspirated stops.

States of the glottis

Several languages use more than just the voiced and voiceless states of the glottis. In Hindi and many of the other languages of India, some sounds are produced while the vocal cords are vibrating for part of their length but are apart, so that a considerable amount of air escapes between them at one end. This phenomenon is known as breathy voice, or murmur. Other languages have sounds in which the vocal cords are held tightly together so that only part of their length can vibrate. This kind of sound, which is usually very low pitched, is sometimes called creaky voice, or vocal fry. It is used to make contrasts between consonants in several American Indian languages. An additional glottal state that is widely used —*e.g.*, in the Austronesian (Malayo–Polynesian) languages of the Philippines—is a glottal stop, a tight closure of the two vocal cords. This articulation also occurs in many forms of English as the usual pronunciation of *t* in words such as *bitten* and *fatten*.

Types of airstream. In English, all sounds are produced with an airstream caused by the expiration of the air from the lungs. This is known as a pulmonic airstream. Other mechanisms for producing an airstream also occur. If there is a glottal stop and the closed glottis is moved rapidly upward or downward it can act like a piston pushing or pulling the air in the pharynx. This is the glottalic airstream mechanism. When there is an upward movement of the closed glottis the resulting sound is called an ejective. Amharic, the national language of Ethiopia, uses this mechanism to produce both ejective stops and fricatives, which contrast with the more usual stops and fricatives made with a pulmonic airstream mechanism. A downward movement of the glottis is used in the production of implosive sounds, which occur in many American Indian, African, and other languages. The use of movements of the tongue to suck air into the mouth is known as the velaric airstream mechanism; it occurs in the production of clicks, which are regular speech sounds in many languages of southern Africa.

To summarize, a consonant may be described by reference to seven factors: (1) state of the glottis, (2) secondary articulation (if any), (3) place of articulation, (4) type of airstream, (5) central or lateral articulation, (6) velic closure—oral or nasal, and (7) manner of articulation. Thus the consonant at the beginning of the word *swim* is a (1) voiceless, (2) labialized, (3) alveolar, (4) pulmonic, (5) central, (6) oral, (7) fricative. Unless a specific statement is made to the contrary, consonants are usually presumed to have a pulmonic airstream and no secondary articulation, and it is also assumed that they are not laterals or nasals. Consequently, points 2, 4, 5, and 6 are often disregarded and a three-term description —*e.g.*, voiceless alveolar fricative is sufficient.

Factors in consonant description

Vowels. Vowels traditionally have been specified in terms of the position of the highest point of the tongue and the position of the lips. Figure 2 shows these positions for eight different vowels. The highest point of the tongue is in the front of the mouth for the vowels in *heed*, *hid*, *head*, and *had*. Accordingly, these vowels are classified as front vowels, whereas the vowels in *hod*, *hawed*, *hood*, and *who'd* are classified as back vowels. The tongue is highest in the vowels in *heed* and *who'd*, which are therefore called high, or close, vowels, and lowest in the vowels in *had* and *hod*, which are called low, or open, vowels. The height of the tongue for the vowels in the other words is between these two extremes, and they are therefore called midvowels. Lip positions may be described as being rounded, as in *who'd*, or unrounded or spread, as in *heed*.

The specification of vowels in terms of the position of the highest point of the tongue is not entirely satisfactory for a number of reasons. In the first place, it disregards the fact that the shape of the tongue as a whole is very different in front vowels and in back vowels. Second, although the height of the tongue in front vowels varies by approximately equal amounts for what are called equidistant steps in vowel quality, this is just not factually true in descriptions of back vowels. Third, the width of

the pharynx varies considerably, and to some extent independently of the height of the tongue, in different vowels.

Tenseness and laxness

Some authorities use terms such as tense and lax to describe the degree of tension in the tongue muscles, particularly those muscles responsible for the bunching up of the tongue lengthways. Other authorities use the term tense to specify a greater degree of muscular activity, resulting in a greater deformation of the tongue from its neutral position. Tense vowels are longer than the corresponding lax vowels. The vowels in *heed* and *hayed* are tense, whereas those in *hid* and *head* are lax.

In many languages there is a strong tendency for front vowels to have spread lip positions, and back vowels to have lip rounding. As will be seen in the next section, this results in vowels that are acoustically maximally distinct. But many languages—*e.g.*, French and German—have front rounded vowels. Thus French has a contrast between a high front unrounded vowel in *vie,* "life," and a high front rounded vowel with a very similar tongue position in *vu,* "seen," as well as a high back rounded vowel in *vous,* "you." Unrounded back vowels also occur —*e.g.*, in Vietnamese.

Nasalized vowels, in which the soft palate is lowered so that part of the airstream goes out through the nose, occur in many languages. French distinguishes between several nasalized vowels and vowels made with similar tongue positions but with the soft palate raised. Low vowels in many forms of English are often nasalized, especially when they occur between nasal consonants, as in *man.*

Cardinal vowels

Because of the difficulty of observing the precise tongue positions that occur in vowels, a set of eight vowels known as the cardinal vowels has been devised to act as reference points. This set of vowels is defined partly in articulatory and partly in auditory terms. Cardinal vowel number one is defined as the highest and farthest front tongue position that can be made without producing a fricative sound; cardinal vowel number five is defined as the lowest and farthest back vowel. Cardinal vowels two, three, and four are a series of front vowels that form auditorily equidistant steps between cardinal vowels one and five; and cardinal vowels six, seven, and eight are a series of back vowels with the same sized auditory steps as in the front vowel series. Phoneticians who have been trained in the cardinal vowel system are able to make precise descriptions of the vowels of any language in terms of these reference points.

Suprasegmentals. Vowels and consonants can be considered to be the segments of which speech is composed. Together they form syllables, which in turn make up utterances. Superimposed on the syllables there are other features that are known as suprasegmentals. These include variations in stress (accent) and pitch (tone and intonation). Variations in length are also usually considered to be suprasegmental features, although they can affect single segments as well as whole syllables. All of the suprasegmental features are characterized by the fact that they must be described in relation to other items in the same utterance. It is the relative values of the pitch, length, or degree of stress of an item that are significant. The absolute values are never linguistically important, although they may be of importance paralinguistically, in that they convey information about the age and sex of the speaker, his emotional state, and his attitude.

Many languages—*e.g.*, Finnish and Estonian—use length distinctions, so that they have long and short vowels; a slightly smaller number of languages, among them Luganda (the language spoken by the largest tribe in Uganda) and Japanese, also have long and short consonants. In most languages segments followed by voiced consonants are longer than those followed by voiceless consonants. Thus the vowel in *cad* before the voiced *d* is much longer than that in *cat* before the voiceless *t*. Variations in stress are caused by an increase in the activity of the respiratory muscles, so that a greater amount of air is pushed out of the lungs, and in the activity of the laryngeal muscles, resulting in significant changes in pitch. In English, stress has a grammatical function, distinguishing between nouns and verbs, such as *an insult* versus *to insult.* It can also be used for contrastive emphasis, as in *I want a RED pen, not a black one.*

Pitch and intonation

Variations in laryngeal activity can occur independently of stress changes. When they do so, the resulting pitch changes can affect the meaning of the sentence as a whole, or the meaning of the individual words. The pitch pattern in a sentence is known as the intonation. In English the meaning of a sentence such as *That's a cat* can be changed from a statement to a question by the substitution of a mainly rising for a mainly falling intonation. Pitch patterns that affect the meanings of individual words are known as tones and are common in many languages. In Chinese, for example, a syllable that is transliterated as *ma* means "mother" when said on a high tone, "hemp" on a midrising tone, "horse" on the falling-rising tone, and "scold" on a high-falling tone.

ACOUSTIC PHONETICS

Speech sounds consist of small variations in air pressure that can be sensed by the ear. Like other sounds, speech sounds can be divided into two major classes—those that have periodic wave forms (*i.e.*, regular fluctuations in air pressure) and those that do not. The first class consists of all the voiced sounds, because the vibrations of the vocal cords produce regular pulses of air pressure.

Fundamental frequency

From a listener's point of view, sounds may be said to vary in pitch, loudness, and quality. The pitch of a sound with a periodic wave form—*i.e.*, a voiced sound—is determined by its fundamental frequency, or rate of repetition of the cycles of air pressure. For a speaker with a bass voice, the fundamental frequency will probably be between 75 and 150 cycles per second. Cycles per second are also called hertz (Hz); this is the standard term for the unit in frequency measurements. A soprano may have a speaking voice in which the vocal cords vibrate to produce a fundamental frequency of over 400 hertz. The relative loudness of a voiced sound is largely dependent on the amplitude of the pulses of air pressure produced by the vibrating vocal cords. Pulses of air with a larger amplitude have a larger increase in air pressure.

The quality of a sound is determined by the smaller variations in air pressure that are superimposed on the major variations that recur at the fundamental frequency. These smaller variations in air pressure correspond to the overtones that occur above the fundamental frequency. Each time the vocal cords open and close there is a pulse of air from the lungs. These pulses act like sharp taps on the air in the vocal tract, which is accordingly set into vibration in a way that is determined by its size and shape. In a vowel sound, the air in the vocal tract vibrates at three or four frequencies simultaneously. These frequencies are the resonant frequencies of that particular vocal tract shape. Irrespective of the fundamental frequency that is determined by the rate of vibration of the vocal cords, the air in the vocal tract will resonate at these three or four overtone frequencies as long as the position of the vocal organs remains the same. In this way a vowel has its own characteristic auditory quality, which is the result of the specific variations in air pressure caused by the superimposing of the vocal tract shape on the fundamental frequency produced by the vocal cords.

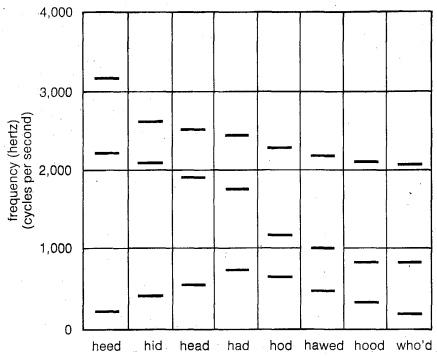


Figure 3: Schematic spectrogram showing frequencies of the first three formants of the vowels in similar English words.

Vowel formants. The resonant frequencies of the vocal tract are known as the formants. The frequencies of the first three formants of the vowels in the words *heed, hid, head, had, hod, hawed, hood,* and *who'd* are shown in Figure 3. Comparison with Figure 2 shows that there are no simple relationships between actual tongue positions and formant frequencies. There is, however, a good inverse correlation between one of the labels used to describe the tongue position and the frequency of the first, or lowest, formant. This formant is lowest in the so-called high vowels, and highest in the so-called low vowels. When phoneticians describe vowels as high or low, they probably are actually specifying the inverse of the frequency of the first formant.

Most people cannot hear the pitches of the individual formants in normal speech. In whispered speech, however, there are no regular variations in air pressure produced by the vocal cords, and the higher resonances of the vocal tract are more clearly audible. It is quite easy to hear the falling pitch of the second formant when whispering the series of words *heed, hid, head, had, hod, hawed, hood, who'd.* Conversely, the auditory effect of the second and higher formants is lessened when speaking in a creaky voice. Under such conditions, it is possible to hear the rise in pitch of the first formant during the first four of these words, and the fall in pitch during the last.

Consonant formants. Voiced consonants such as nasals and laterals also have specific vocal tract shapes that are characterized by the frequencies of the formants. They differ from vowels in that in their production the vocal tract is not a single tube. There is a side branch formed when the nasal tract is coupled in with the oral tract, or, in the case of laterals, when the oral tract itself is obstructed in the centre. The effect of these side branches is that the relative amplitudes of the formants are altered; it is as if one or more of the possible superimposed variations in air pressure had been lessened because it had been trapped in the cavity formed at the side. Nasals and laterals can therefore be specified in terms of their formant frequencies, just like vowels. But in a complete specification of these consonants the relative amplitudes of the formants also have to be given, because they are not completely predictable.

Other voiced consonants such as stops and approximants (semivowels) are more like vowels in that they can be characterized in part by the resonant frequencies—the formants—of their vocal tract shapes. They differ from vowels in that during a voiced stop closure there is very little acoustic energy, and during the release phase of a stop and the entire articulation of a semivowel the vocal tract shapes are changing comparatively rapidly. These transitional movements can be specified acoustically in terms of the movements of the formant frequencies.

Pitch of voiceless sounds

Voiceless sounds do not have a periodic wave form with a well-defined fundamental frequency. Nevertheless, some sensations of pitch accompany the variations in air pressure caused by the turbulent airflow that occurs during a voiceless fricative, or in the release phase of a voiceless stop. This is because the pressure variations are far from random. During the first consonant in *sea* these have a tendency to be at a higher centre frequency, and hence a higher pitch, than in the pronunciation of the first consonant in *she.* There is also a difference in the average amplitude of the wave form in different voiceless sounds. All voiceless sounds have much less energy—*i.e.*, a smaller amplitude—than voiced sounds pronounced with the same degree of effort. Other things being equal, the fricatives in *sin* and *shin* have more amplitude—*i.e.*, are louder—than those in *thin* and *fin.*

In summary, speech sounds are fairly well defined by nine acoustic factors. The first three factors include the frequencies of the first three formants; these are responsible for the major part of the information in speech. Characterizing the vocal tract shape, these formant frequencies specify vowels, nasals, laterals, and the transitional movements in voiced consonants. The frequencies of the fourth and higher formants do not vary significantly. The fourth factor is the fundamental frequency—roughly speaking, the pitch—of the larynx pulse in voiced sounds, and the fifth, the amplitude—roughly speaking, the loudness—of the larynx pulse. These last two factors account for suprasegmental information; *e.g.*, variations in stress and intonation. They also distinguish between voiced and voiceless sounds, in that the latter have no larynx pulse amplitude. The centre frequency of the high-frequency hissing noises in voiceless sounds constitutes the sixth acoustic factor, and the seventh is the amplitude of these high-frequency noises. These two factors characterize the major differences among voiceless sounds. In more accurate descriptions it would be necessary to specify more than just the centre frequency of the noise in fricative sounds. The eighth and ninth factors include the amplitudes of the second and third formants relative to the first formant; the amplitudes of the formants as a whole are determined by the larynx pulse amplitude. These latter factors are the least important in that they convey only supplementary information about nasals and laterals.

Sound spectrograph

Instruments for acoustic phonetics. The principal instrument used in acoustic phonetic studies is the sound spectrograph. This device gives a visible record of any kind of sound. A spectrographic analysis of the phrase *speech pictures* is shown in Figure 4. The time of occurrence of each item is given on the horizontal scale. The vertical scale shows the frequency components at each moment in time, the amplitude of the components being shown by the darkness of the mark. (Figure 3 diagrams

By courtesy of Peter Ladefoged, University of California, Los Angeles

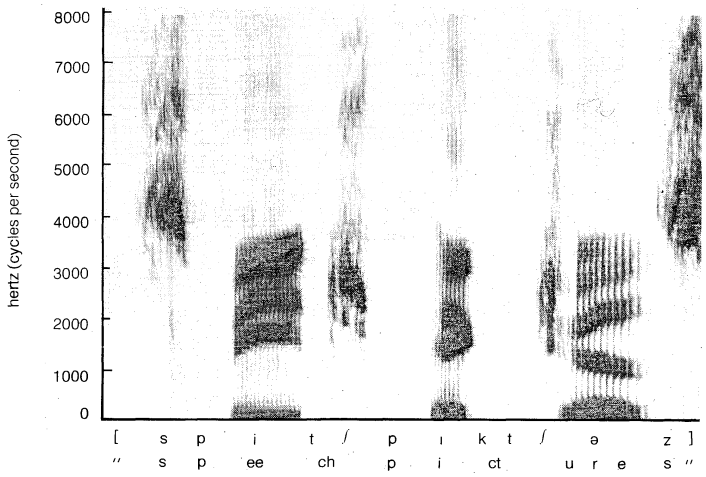


Figure 4: A spectrogram of the phrase "speech pictures." The horizontal dimension represents time, and the vertical scale represents frequency. The amplitude of the component frequencies at each moment in time is shown by the darkness of the mark.

the formant frequencies in a set of English vowels in the same way and might be regarded as a schematic spectrogram.) In the phrase *speech pictures* the first consonant has a comparatively random distribution of energy, but, as Figure 4 shows, it is mainly in the higher frequencies. The second consonant is a voiceless stop, which produces a short gap in the pattern. The next segment, the first vowel, has four formants that appear as dark bars with centre frequencies of 300, 2,000, 2,700, and 3,400 hertz. Each of the other segments has its own distinctive pattern.

Much information has also been gained from the use of speech synthesizers, which are instruments that take specifications of speech in terms of the acoustic factors summarized above and generate the corresponding sounds. Some speech synthesizers use electronic signal generators and amplifiers; others use digital computers to calculate the values of the required sound waves. Good synthetic speech is hard to distinguish from high-quality recordings of natural speech. The principal value of a speech synthesizer is its precisely controllable "voice" that an experimenter can vary in a systematic way to determine the perceptual effects of different acoustic specifications.

LINGUISTIC PHONETICS

Phonetics is part of linguistics in that one of the main

Relation of phonetics to linguistics

aims of phonetics is to determine the categories that can be used in explanatory description of languages. One way of looking at the grammar of a language is to consider it to be a set of statements that explains the relation between the meanings of all possible sentences in a language and the sounds of which they are composed. In this view, a grammar may be divided into three parts: the syntactic component, which is a set of rules describing the ways in which words may form sentences; the lexicon, which is a list of all the words and the categories to which they belong; and the phonological component, which is a set of rules that relates phonetic descriptions of sentences to the syntactic and lexical descriptions.

Phonological rules. In the lexicon of a language, each word is represented in its underlying, or basic, form, which discounts all of the alternations in pronunciation that are predictable by phonological rules. For example, there are phonological rules that will account for the variations in the placement of stress and the alternations of vowel quality that occur in sets of words such as *harmOny*, *harmOnic*, *harmOnious* and *melOdy*, *melOdic*, *melOdious*. The rules that predict the pronunciation of the capitalized *O*'s are general, rather than specific for each word, and the grammar should state such rules so that the regularities are revealed. Accordingly, each of these words must be entered in the lexicon in a way that represents simply its underlying form, and that allows the alternations that occur to be generated by phonological rules. The underlying form is known as the phonemic—sometimes morphophonemic, or phonological—representation of the word. The phonemes of a language are the segments that contrast in the underlying forms. American English may be said to have at least 13 vowel phonemes, which contrast in the underlying forms of words such as *bate*, *bat*, *beat*, *bet*, *bite*, *bit*, *bout*, *but*, *boat*, *dot*, *bought*, *balm*, and *boy*. Some authorities consider that there are additional vowel phonemes exemplified in the words *bush* and *beaut(y)*, but others believe that these can be derived from the same underlying vowel as that in the word *bud*. Phonemes are traditionally written between slanting lines, as /p/, /m/, or /l/.

The variants of phonemes that occur in phonetic representations of sentences are known as allophones. They may be considered to be generated as a result of applying the phonological rules to the phonemes in underlying forms. For example, there is a phonological rule of English that says that a voiceless stop such as /p/ is aspirated when it occurs at the beginning of a word (*e.g.*, in *pin*), but when it occurs after a voiceless alveolar fricative (*i.e.*, after/s/), it is unaspirated (*e.g.*, in *spin*). Thus the underlying phoneme /p/ has an aspirated and an unaspirated allophone, in addition to other allophones that are generated as a result of other rules that apply in other circumstances. Allophones are conventionally written inside brackets—*e.g.*, [p] or aspirated [p^h].

Classes of phonemes

In stating phonological rules it is necessary to refer to classes of phonemes. Consider part of the rule for the formation of the plural in English: there is an extra vowel in the suffix if the word ends in the same sound as occurs at the end of *horse, maze, fish, rouge, church*, or *judge*. The plural forms of words of this kind are one syllable longer than the singular forms. The phonological rules of English could simply list the phonemes that behave in the same way in the rules for plural formation; the rules for the possessive forms of nouns and for the 3rd person singular of the present tense of verbs are similar in this respect. The rules are more explanatory, however, if they show that these phonemes behave in a similar way because they form a natural class, or set, whose members are defined by a common property. In the case of these plural forms, the phonemes are all, and only, those that have a high-frequency fricative component; they may be called the sibilant, or strident, phonemes.

Other phonological rules that refer to natural classes of phonemes have already been mentioned. The rule about voiceless stops being aspirated in some circumstances and unaspirated in others refers to the subset of phonemes that are both voiceless and stops. Similarly, the variations in vowel length in *cat* and *cad* can be stated with reference to the set of phonemes that are vowels, and also to the set that comprises both voiceless sounds and stops.

Features. Each of the phonemes that appears in the lexicon of a language may be classified in terms of a set of phonetic properties, or features. Phoneticians and linguists have been trying to develop a set of features that is sufficient to classify the phonemes in each of the languages of the world. A set of features of this kind would constitute the phonetic capabilities of man. To be descriptively adequate from a linguistic point of view, the set of features must be able to provide a different representation for each of the words that is phonologically distinct in a language; and if the feature set is to have any explanatory power it must also be able to classify phonemes into appropriate natural classes as required in the phonological rules of each language.

In the earlier work on feature sets, emphasis was placed on the fact that features were the smallest discrete components of language. Not much attention was paid to their role in classifying phonemes into the natural classes required in phonological rules. Instead, they were considered to be the units to which a listener attends when listening to speech. Features were justified by reference to their role in distinguishing phonemes in minimal sets of words such as *bill, pill, fill, mill, dill, sill, kill.*

Jakobson, Fant, and Halle features. As a result of studying the phonemic contrasts within a number of languages, Roman Jakobson, Gunnar Fant, and Morris Halle concluded in 1951 that segmental phonemes could be characterized in terms of 12 distinctive features. All of the features were binary, in the sense that a phoneme either had, or did not have, the phonetic attributes of the feature. Thus phonemes could be classified as being consonantal or not, voiced or not, nasal or not, and so on. In 1968, Noam Chomsky and Morris Halle stated that nearer 30 features are needed for a proper description of the phonetic, and linguistic, capabilities of man. In agreement with Jakobson, they claimed that each feature functions as a binary opposition that can be given the value of plus or minus in classifying the phonemes in underlying forms. But they suggested that the features may require more precise systematic phonetic specifications.

Binary quality of features

Chomsky–Halle features. Some of the binary features proposed by Chomsky and Halle are listed in Table 1. The first group comprises major class features, because these features are required for dividing sounds into classes such as vowels, consonants, and semivowels. There are several problems in giving satisfactory definitions of the phonetic properties of these features, but there is no doubt that binary oppositions of this kind are needed for describing phonological patterns.

The next group, the manner of articulation features, includes continuant/noncontinuant (in which noncontinuant is exactly equivalent to the notion of stop as defined above), and delayed release (the comparatively slow parting of the articulators that occurs in an affricate). The source features refer to the action of the vocal cords (voice) or to fricative noise mechanisms (strident).

The cavity features include nasal and lateral, which are used in the same sense as they were in the section on articulatory phonetics, and the features that determine the place of articulation of consonants and the quality of vowels. The most important features specifying the place of articulation of consonants are anterior, made in the

Table 1: Part of the Set of Features Proposed by Chomsky and Halle (1968)

Major class features	Cavity features
consonantal	nasal
syllabic	lateral
sonorant	anterior
Manner of articulation features	coronal
continuant	high
delayed release	low
tense	back
Source features	round
voice	
strident	

Table 2: The Distinctive Feature Composition of Some English Segments

	i	u	e	ɑ	ɪ	w	r	l	p	t	d	θ	ð	n	s	z	ʃ	k	h
Syllabic	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Consonantal	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	−
High	+	+	−	−	+	+	−	−	−	−	−	−	−	−	−	−	+	+	−
Back	−	+	−	+	−	+	−	−	−	−	−	−	−	−	−	−	−	+	−
Low	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+
Anterior	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	−	−	−
Coronal	−	−	−	−	−	−	+	+	−	+	+	+	+	+	+	+	+	−	−
Round	−	+	−	−	−	+													
Tense	+	+	+	+	−	−													
Voice							+	+	−	−	+	−	+	+	−	+	−	−	−
Continuant							+	+	−	−	−	+	+	−	+	+	+	−	+
Nasal							−	−	−	−	−	−	−	+	−	−	−	−	−
Strident							−	−	−	−	−	−	−	−	+	+	+	−	−
Lateral							−	+	−	−	−	−	−	−	−	−	−	−	−

front of the mouth, and coronal, made with the tip or blade of the tongue raised toward the teeth or teeth ridge. These two features can be used to specify four places of articulation: bilabial (+anterior, −coronal); dental, or alveolar (+anterior, +coronal); postalveolar, or palatoalveolar (−anterior, +coronal); velar (−anterior, −coronal). There is still some disagreement concerning whether consonantal places of articulation are specified appropriately by binary oppositions of this kind.

Vowels and binary features

There is even more disagreement over the advisability of describing vowels in terms of binary features. Chomsky and Halle use the features high/nonhigh and low/nonlow to specify the height of the tongue, midtongue positions being considered to be simply those that are −high, −low; the feature back/nonback is employed to specify the front/back distinctions among vowels. But these three features can be combined to specify only six basic tongue positions: high front, high back, midfront, midback, low front, and low back. It is true that each of these possibilities can have tense/nontense (lax) variants and rounded/nonrounded (spread) variants. But the Chomsky–Halle feature system does not permit the specification within underlying forms of central vowels, nor of more than three degrees of tongue height. Moreover, their binary oppositions of vowel height do not make it clear that the difference between low vowels and midvowels is the same as that between midvowels and high vowels.

Table 2 shows the feature composition of a number of segments that occur in English. The phonetic symbols at the top of each column are used with the values discussed in the following section.

PHONETIC TRANSCRIPTION

There are many different kinds of phonetic transcription. In some circumstances a phonetic symbol can be simply an abbreviation for a phonetic description. The symbol [s] may then be regarded as exactly equivalent to the phrase "voiceless, alveolar, fricative." When a linguist tries to describe an unknown language he begins by writing it down using symbols in this way. Later, when he has learned about the function of sounds and the underlying forms in the language, he might make a more systematic transcription, known as a broad transcription, in which each phoneme is represented by a simple symbol.

Occasionally it is convenient to use a transcription in which some of the allophones are represented by specific symbols, or some of the phonemes are designated by the symbols for a more restricted set of categories. If, for example, the transcription were to be used in teaching pronunciation, the difference between the aspirated and unaspirated allophones of /p/ might be represented by transcribing *pan* as [pʰan] and *span* as [span]; or the vowel phoneme in each of these words might be designated by the more specific symbol [æ], which represents a low front vowel of a certain type. In a narrow transcription the symbols are more specific, either because allophones are differentiated, or because the phonetic quality of the sounds is shown more precisely.

Narrow transcription

The most widely used set of symbols is that of the International Phonetic Association (IPA). The latest version of the IPA chart is shown in Table 3. In general, the consonants have the same values as the corresponding letters in many European languages; the vowel symbols have similar values to the corresponding letters in a language such as Italian.

EXPERIMENTAL PHONETICS

Experimental phonetics employs the methods of investigation commonly used in other disciplines—*e.g.*, physics, physiology, and psychology—for measuring the physical and physiological dimensions of speech sounds and their perceptual characteristics. The sound spectrograph and speech synthesizers were mentioned in the section on acoustic phonetics. Other techniques include the use of X-rays; air-pressure and air-flow recording; palatography, a method of registering the contacts between the tongue and the roof of the mouth; and cinematography. All of these techniques have been used for studying the actions of the vocal organs.

Much of the work in experimental phonetics has been directed toward obtaining more accurate descriptions of the sounds that characterize different languages. There have also been several studies aimed at determining the relative importance of different features in signalling contrasts between sounds. But experimental phoneticians are probably most concerned with trying to discover the central cerebral processes involved in speech.

Motor theory of speech perception

One issue of this kind that has been extensively discussed is the so-called motor theory of speech perception. There is a great deal of evidence that the way in which people speak greatly influences their perception of what is said to them. For example, speakers of Spanish cannot pronounce the different vowels in words such as *ship* and *sheep* in English. These people also have difficulty in hearing the difference between these two vowels. But when they have learned, by trial and error methods, to say them correctly, then they can easily hear the difference. Similarly, using synthetic speech stimuli it is possible to make a series of consonant sounds that go by acoustically equidistant steps from [b] through [d] to [g]. When listeners hear these synthetic sounds they do not consider the steps between them to be auditorily equidistant. The steps that correspond to the large articulatory movements between the consonants are heard as being much larger than the equal size acoustic steps that do not correspond to articulatory movements occurring in the listener's speech. Facts such as these have led some phoneticians to believe that the perception of speech is structured more in motor—articulatory—terms than in acoustic terms. Other phoneticians have claimed that the evidence does not really distinguish between these two possibilities but demonstrates simply that the perception of speech is structured in terms of linguistic categories.

Perception and size of speech units

Another major problem is the size of the units that are involved in the perception of speech. Some authorities have claimed that a listener distinguishes between words by making a series of binary decisions concerning the features in each segment that he hears. Others hold that the listener takes in information in much larger temporal pieces and perhaps processes speech in terms of units of at least the size of a syllable. All authorities agree on the importance of context. Speech conveys information in a redundant way. Experiments have shown that a listener need attend to only a part of the information presented to him in order to understand all that is being said.

Table 3: The International Phonetic Alphabet*
(revised to 1951)

	bilabial	labio-dental	dental and alveolar	retroflex	palato-alveolar	alveolo-palatal	palatal	velar	uvular	pharyngal	glottal
Consonants											
Plosive	p b		t d	ʈ ɖ			c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ	n	ɳ			ɲ	ŋ	ɴ		
Lateral fricative			ɬ ɮ								
Lateral nonfricative			l	ɭ			ʎ				
Rolled			r						ʀ		
Flapped			ɾ	ɽ					ʀ		
Fricative	ɸ β	f v	θ ð s z ɹ	ʂ ʐ	ʃ ʒ	ɕ ʑ	ç j	x ɣ	χ ʁ	ħ ʕ	h ɦ
Frictionless continuants and semivowels	w ɥ	ʋ	ɹ				j (ɥ)	(w)	ʁ		

	front	central	back
Vowels			
Close	(y ʉ u)		
	i y	ɨ ʉ	ɯ u
Half-close	(ø o)		
	e ø		ɤ o
Half-open	(œ ɔ)		
	ɛ œ	ɘ	ʌ ɔ
	æ	ɐ	
Open	(ɒ)		
	a		ɑ ɒ

*Secondary articulations are shown by symbols in parentheses.
Other sounds—palatalized consonants: ʇ, ɗ, etc.; palatalized ʃ, ʒ : ʄ, ʒ̠. Velarized or pharyngalized consonants: ɬ, ɗ, ʐ, etc. Ejective consonants (with simultaneous glottal stop): p', t', etc. Implosive voiced consonants: ɓ, ɗ, etc. ɾ fricative trill. ʊ, ɻ (labialized θ, ð, or s, z). ɬ, ʒ (labialized ʃ, ʒ). ǀ, ǃ, ǂ (clicks, Zulu *c*, *q*, *x*). ɺ (a sound between r and l). ɳ Japanese syllabic nasal. ʃ (combination of x and ʃ). ʍ (voiceless w). ɪ, ʏ, ɷ (lowered varieties of i, y, u). ɜ (a variety of ə). ɵ (a vowel between ø and o).
Affricates are normally presented by groups of two consonants (ts, tʃ, d , etc.), but, when necessary, ligatures are used (t͡s, t͡ʃ, d͡ʒ, etc.), or the marks ‿ or ‿ (f‿s or t‿s, etc.). ‿ ‿ also denote synchronic articulation (m‿ŋ = simultaneous m and ŋ). c, ɟ may occasionally be used in place of tʃ, dʒ, and ʦ, ʣ for ts, dz. Aspirated plosives: ph, th, etc. r-coloured vowels: eɹ, aɹ, ɔɹ, etc., or $e^{ɹ}$, $a^{ɹ}$, $o^{ɹ}$, etc., or ɐ̯, ä̯, ɒ̯, etc.; r-coloured ə: əɹ or $ə^{ɹ}$ or ɹ or ə̯ or ɚ.
Length, stress, pitch—: (full length). ˑ (half length). ˈ (stress, placed at beginning of the stressed syllable). ˌ (secondary stress). ˉ (high level pitch); ˍ (low level); ˊ (high rising); ˏ (low rising); ˋ (high falling); ˎ (low falling); ˆ (rise-fall); ˇ (fall-rise).
Modifiers— ˜ nasality. ˭ breath (l̥ = breathed l). ˬ voice (s̬ = z). ʻ slight aspiration following p, t, etc. ˷ labialization (n̪ = labialized n). ̪ dental articulation (t̪ = dental t). ʲ palatalization (ż = ʐ). ̣ specially close vowel (ę = a very close e). ̩ specially open vowel (ę = a rather open e). ̟ tongue raised (e̟ or ę̟ = ę). ̠ tongue lowered (e̠ or ę̠ = ę). ˖ tongue advanced (u˖ or ʉ̟ = an advanced u, t̠ = t̪). - or ̈ tongue retracted (i- or =i̠, t̠ = alveolar t). ˃ lips more rounded. ˂ lips more spread. Central vowels: ï (= i), ü (= ʉ), ë (= ə̟), ö (= ɵ), ë, ɜ̈.
̩ (e.g. n̩) syllabic consonant. ̯ consonantal vowel. ʃ˕ variety of ʃ resembling s, etc.

A related problem is that of the temporal structure of speech production. There may be very little structure, and a speaker may simply time the movements of his vocal organs by allowing each gesture to run its course before starting on the next one. Alternatively, he may impose a hierarchical structure on the gestures by requiring, for instance, each major stress in a sentence to occur at some predetermined moment, and the articulatory movements to be speeded up or slowed down depending on the number of movements that have to occur before the major stress. There is some evidence in favour of this latter possibility as a result of experiments in which a speaker is asked to say a given phrase first slowly and then fast. When he is speaking at a rate that is twice as fast as some other rate, then the interval between the major stresses is about halved. But the duration of each segment is not halved. The consonants are only slightly reduced in length, whereas the vowels are considerably shortened. Some authorities have used the results of experiments of this kind to argue that the stress group is the major unit in the temporal organization of speech.

HISTORY OF PHONETICS

The earliest phoneticians were the Indian grammarians who wanted to preserve what they thought to be the correct pronunciation of Sanskrit. Their motives were religious, in that it was very important to them that their holy works should be pronounced in the traditional way. The works of Pāṇini, a Sanskrit scholar who lived about 2,300 years ago, are still highly regarded by grammarians.

None of the classical civilizations of the Mediterranean basin produced any phonetic descriptions of comparable importance, but it must be remembered that the Greeks were primarily responsible for the greatest phonetic invention of all time—the development of a writing system in which syllables are represented in terms of their component parts. All other writing systems suffer either from ambiguity in that many different syllables have to be written in the same way, or from requiring a very large number of different symbols. The realization that each vowel and each consonant could be represented by a separate symbol made it possible to write down any word that was said with a small inventory of symbols.

In Renaissance times, Leonardo da Vinci made numerous observations on the physiology of speech. Apart from a few individual pieces of work that did not have much influence on the development of phonetic knowledge, however, there were no systematic descriptions of speech until the 17th century. Then grammarians such as John Wallis (1616–1703) included lengthy sections on speech in their work, and several publications by teachers of pronunciation and spelling reformers show that the method of production of most speech sounds was becoming generally understood.

It was not until the rise of linguistic science in the 19th century that the forerunners of modern phonetics can be clearly discerned. Following the earlier work of Sir Isaac Pitman and Alexander J. Ellis, Alexander Melville Bell published *Visible Speech* in 1867. This book proved to be a milestone in the history of phonetics, in that it provided a greatly improved set of categories for describing the sounds of any language. By using a very detailed notational system, Bell and his co-workers, including his son, Alexander Graham Bell, the inventor of the telephone, were able to write down small variations in speech sounds with an accuracy previously impossible.

The emergence of phonetics as an independent field can be said to date from this time. The German scholar Eduard Sievers published his book on the physiology of speech sounds in 1876, and the *Handbook of Phonetics*, by the great English phonetician Henry Sweet, appeared in the following year. Scholars in many other European countries were becoming interested in describing and comparing the sounds of languages. Paul Passy in France and Otto Jespersen in Denmark joined with Bell, Sievers, and Sweet, and many others to form the International Phonetic Association in 1886.

International phonetic association

The leading phonetician during most of the first half of the 20th century was Daniel Jones, who started teaching in 1907 and retired as Professor of Phonetics at University College in London in 1949. He was president of the International Phonetic Association from 1950 until his death in 1967 at the age of 86. His numerous publications, particularly those on the pronunciation of the British accent known as RP (see above *Standards of pronunciation*), served as a model for phoneticians all over the

world. Their influence can be seen in the phonetic aspects of the work of American linguists such as Leonard Bloomfield, Bernard Bloch, Kenneth Pike, and many other leading figures in the period before the recent work of Chomsky and Halle.

The phonetic roots of Chomsky and Halle's work derive from the distinctive feature theories originated by the linguists in Prague headed by Nikolay S. Trubetskoy and Roman Jakobson. The culmination of Trubetskoy's work was his *Grundzüge der Phonologie* (1939; Eng. trans., *Principles of Phonology*, 1969), published in the year after he died. Jakobson, who emigrated to the United States in 1941, collaborated with Gunnar Fant and Morris Halle on a complete statement of the theory of distinctive features; this appeared in 1951.

BIBLIOGRAPHY. D. ABERCROMBIE, *Elements of General Phonetics* (1967), a readable survey of selected topics; K.L. PIKE, *Phonetics* (1943), and P. LADEFOGED, *Preliminaries to Linguistic Phonetics* (1971), critical analyses of phonetic theory; R. JAKOBSON, C.G.M. FANT, and M. HALLE, *Preliminaries to Speech Analysis* (1952), the first major account of distinctive feature theory; N. CHOMSKY and M. HALLE, *The Sound Pattern of English* (1968), a comprehensive description of English phonology; P. LADEFOGED, *Elements of Acoustic Phonetics* (1962), an introduction to the physics of speech; I. LEHISTE (comp.), *Readings in Acoustic Phonetics* (1967), a collection of major articles; B. MALMBERG (ed.), *Manual of Phonetics* (1968), authoritative chapters on many aspects of phonetics.

(P.N.L.)

Phoronida

Phoronids, which comprise the phylum Phoronida, are little-known marine invertebrate animals characterized by an elongated, nonsegmented body that is topped by a tuft of tentacles. Each adult lives within a membranous tube to which sand particles, shells, and other materials may adhere. A few species penetrate calcareous shells and rocks; most phoronids, however, burrow in sand or mud. The animal extends its tentacles beyond the tube to feed and breathe; when disturbed, however, a phoronid quickly retreats into its tube. The sedentary (*i.e.*, permanently attached to some surface) adults lead quiet lives, feeding, growing, and reproducing. An adult lives several years, each year producing young that may number thousands. The free-swimming young, called actinotrochs, live several weeks as members of the oceanic plankton before changing to young adults.

Although most phoronids are found intertidally or in shallow ocean regions in temperate zones, their habitats also include tropical regions and ocean depths. In general, phoronids are not common; certain species, however, may carpet the bottoms of shallow bays. Phoronids are of little practical value to man, but their unusual life cycle and obscure ancestry have made them objects of biological interest. Since their discovery in 1846, phoronids have been associated with every animal phylum, and despite continued investigations, the phylogeny of phoronids is not yet known with certainty.

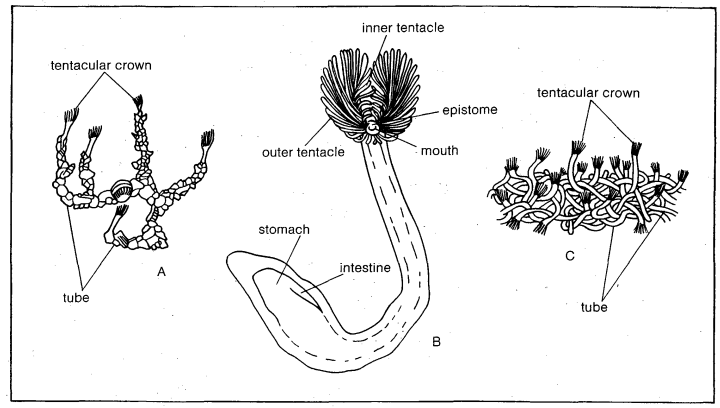
Natural history. A phoronid may be male, female, or contain the reproductive organs of both sexes. The gonads are simple, and paired excretory funnels double as ducts (called gonoducts) for releasing eggs or sperm. The eggs, fertilized within the phoronid, usually pass directly into seawater and develop there as members of the plankton; in a few cases, early embryos are protected within the tentacles of the parent, and only the advanced stages are planktonic.

The actinotroch larva has a straight digestive tract and excretory tubules called protonephridia. Adults have three body regions, each of which contains a body cavity called a coelom. The body wall of the larval trunk pushes inward during development; during metamorphosis of a larva to an adult, this invagination turns outward, forming the body wall of the adult trunk and simultaneously drawing out the digestive tract to establish the U shape of the adult gut. Metamorphosis, which lasts only a few minutes, changes the actinotroch larva from an organism adapted for a planktonic environment into a sessile adult, specialized for a habitat at the bottom of the ocean.

Asexual reproduction is known for one species. All phoronids undergo autotomy, however, during which the lophophoral region, containing tentacles, mouth, anus, excretory organs, and nervous tissue, is pinched off. Autotomy occurs when conditions are unfavourable; after suitable conditions again are present, the phoronid regenerates the lost parts.

Phoronids have formed few symbiotic associations; one species is known to live with an anemone, and a ciliated protozoan lives on another species. Although phoronids have few predators, fish may nibble at the tentacles of individuals that fail to retract quickly within their tube.

From L. Hyman, *The Invertebrates: Smaller Coelomate Groups*, vol. 5 (1959); McGraw-Hill Book Company



The structure of phoronids.
(A) *Phoronis psammophila* with tubes coated with rock grains and shells. (B) Young adult, *Phoronis architecta*. (C) Aggregation of *P. hippocrepia*.

Form and function. The adult body can be divided into three regions. The trunk extends almost the length of the animal (from 0.5 to 25 centimetres). A small distal region called the lophophore includes a ring of tentacles; lophophores vary in shape from an indented circle to an elaborate double spiral. The third region is a very small liplike epistome that overhangs the mouth. Each body region contains a coelom. The coelom of the lophophore extends into each tentacle and is continuous with the small coelomic channels of the epistome. The large trunk coelom is subdivided by mesenteries that support the digestive tract, blood vessels, excretory organs, and gonads.

The adult digestive tract is U shaped; as a result, the mouth is next to the anus. The tentacles encircle the mouth (but not the anus), and their cilia create water currents that bring diatoms and other planktonic food close to the mouth. Trapped in mucus, organisms are ingested, pass through the esophagus, and are broken down in the stomach. Nutrients are distributed to all organs by the circulatory and coelomic fluids. Indigestible wastes are eliminated through the anus. The closed circulatory system consists of arteries, veins, and capillary networks for respiratory exchange and nutrient transfer. Hemoglobin in nucleated red blood cells serves as a respiratory pigment. The nervous system is simple.

Evolution and classification. The relationship of phoronids to other animals is controversial. They have not been preserved in the geological record, although fossil tubes (called *Scolithus*) may belong to phoronids. Embryological stages often are useful in relating animal groups, but the actinotroch larva of phoronids is unique. On the basis of adult body form, especially similar body regions, the phoronids are grouped near the colonial bryozoans and the shell-bearing brachiopods. The sixteen species in the phylum Phoronida are very similar in body form; only two genera, *Phoronis* and *Phoronopsis* (and no higher groups), are recognized.

BIBLIOGRAPHY. L.H. HYMAN, "The Lophophorate Coelomates—Phylum Phoronida," in *The Invertebrates*, vol. 5, *Smaller Coelomate Groups*, pp. 228–274 (1959), the most comprehensive review article in English; P.A. MEGLITSCH, "Lophophorate Coelomates," in *Invertebrate Zoology* (1971), a brief account of the phylum.

(R.L.Zi.)

Phosphate Minerals

The phosphate minerals are those containing phosphorus, a reactive element that combines readily with oxygen. In the course of this oxidation, phosphorus becomes a positively charged atom (an ion) surrounded by and bonded to four oxygen ions that are arranged as a tetrahedron. This arrangement is called the orthophosphate ion, or more simply the phosphate ion; it is the basic building block of all phosphate minerals.

Importance of the phosphate minerals

There are more than 200 reported species of phosphate minerals, but most are rare; and some are of questionable validity. By far the most widely distributed and commercially important phosphate minerals are those of the apatite group, principally the fluoride bearing species, fluorapatite. This mineral is ubiquitous as an accessory mineral in all rock types of all geological ages. Sedimentary rocks composed principally of fluorapatite are termed phosphorites or phosphate rock.

Phosphate rock is a raw material for many industrial and agricultural products. Commonly, it is reacted with acid to yield a superphosphate, a soluble component for inorganic fertilizers. Various forms of phosphorus compounds are used extensively in detergents, plasticizers, insecticides, and in water treatment. The world reserve of phosphate rock is ample for more than 1,000 years at present rates of consumption. Major deposits are in the northwestern and southeastern United States, North Africa, and Europe.

Monazite is the second most important phosphate mineral, because of its content of rare earth oxides. These comprise a group of elements of similar chemical properties that are used as additives in glass, alloys, and solid state devices. Monazite is associated with zircon and titanium minerals in placer concentrates (alluvial deposits) derived from igneous (crystalline) rocks. There are extensive monazite deposits in the southeastern United States, Australia, India, Brazil, and Africa.

Ornamental varieties and crystal localities

Turquoise is a secondary mineral that generally occurs as cryptocrystalline-massive aggregates—*i.e.*, the individual crystals are too small to be seen. Variegated turquoise green has been valued since antiquity as a gemstone. The best specimens have for centuries come from the ʿAlī Mīrzā Kūh, near Neyshābūr, Iran, but good materials have been obtained from the Sinai Peninsula. In the United States there are several turquoise localities in Arizona, New Mexico, Nevada, and Colorado.

Gemlike crystals of yellow-green fluorapatite from the open-pit iron mines of Cerro Mercado, Durango, Mexico, are so readily available that they are used in mineral sets that represent Mohs scale of hardness (ranging from talc [1] to diamond [10]; apatite is [5]). Beautiful faceted stones are cut from this material, but the lack of hardness limits their value. Pegmatites of New England, especially Mt. Apatite at Auburn, Maine, are famous for their coloured apatite crystals. Southern California pegmatites (very coarse-grained igneous-like rocks) have yielded fine specimens, but they are rare. Marble in Sussex County, New Jersey, and Grenville marble in Ottawa County, Quebec, and in Haliburton County, Ontario, have yielded good specimens, some very large.

Deposits with enormous crystals of apatite occur in Kola Peninsula, Soviet Union. Fine, clear crystals are found in veins at Saint Gotthard, Switzerland; deposits also occur in Europe, South America, and Asia.

This article treats the physical and chemical properties of the principal phosphate minerals and their occurrence in nature in terms of environmental conditions of formation. For further information on the relation of phosphate minerals to other mineral groups, see MINERALS. See also CRYSTALLOGRAPHY, for general coverage of structural arrangements and symmetry elements of crystals, and the articles PHOSPHORITES and GEMSTONES, for additional information on economic deposits of phosphate minerals and ornamental varieties, respectively.

DESCRIPTIVE MINERALOGY

The phosphate tetrahedron has a net negative charge of three, and this can be supplemented by additional negative charges from hydroxyls (OH—) or halogens (*i.e.*, fluoride, chloride, bromide, and iodide). Cations (positively charged ions) then balance these charges. A simple classification of phosphate minerals consists of subdivision into the anhydrous group; the hydrous group with hydroxyls or halogens; and the hydrous group with hydroxyls or halogens and water.

Anhydrous phosphate minerals. *Xenotime.* Xenotime is an yttrium phosphate (YPO_4) that occurs as an accessory mineral in igneous rocks and pegmatites. The mineral generally is found as small, brown, tetragonal crystals —those referable to three mutually perpendicular crystallographic axes, two of which are of equal length and the third shorter or longer—which are similar in form, structure, and properties to zircon. Although it can be confused with zircon, it has inferior hardness and perfect cleavage (*i.e.*, a tendency to split along crystallographic planes) perpendicular to one of the two equal axes of the tetragonal system. Xenotime provides the simplest example of balancing of the phosphate anion group (with a negative charge of three) by a trivalent cation with both the phosphorus and the yttrium in fourfold coordination —*i.e.*, bound to four oxygens (see CHEMICAL BONDING).

Berlinite. Berlinite is a rare aluminum phosphate ($AlPO_4$) mineral that has been found at a single locality in Sweden. The mineral resembles quartz in its form, structure, and properties.

Monazite. Monazite [$(Ce,La)PO_4$] is an accessory mineral in igneous and metamorphic rocks and pegmatites. The crystals are monoclinic (*i.e.*, referable to three unequal crystallographic axes, two intersecting obliquely and the third perpendicular to both of them), generally small, and euhedral (perfectly formed). They have a flattened tabular form with angular terminations. Although the cerium–lanthanum ratio of the mineral ordinarily is about 1:1, most monazite contains thorium that substitutes for the rare earths and causes metamictization (alterations caused by radioactive emanations) of the structure. Because of the range of composition and metamictization, the physical properties show considerable variation. Colour is tan to brown, and hardness on the Mohs scale ranges from 5 to 5.5. The cerium–lanthanum phosphate has a density of about 4.6, but thorium substitution results in increased density, to 5.4. The mineral can be distinguished from sphene by its optical characteristics, and from zircon and xenotime on the basis of external form.

Berlinite, xenotime, and monazite are referred to different crystal systems, thus illustrating the crystallographic changes that result from increased ionic radii of the cation.

Whitlockite. Whitlockite [$Ca_3(PO_4)_2$] is a rare phosphate mineral found in the Palermo pegmatite, New Hampshire, and also in some phosphorite rock and cave deposits. It is the simplest mineral in which the charge of the phosphate anion is satisfied by divalent cations (in this case by three calcium ions). The crystals are hexagonal (referable to three equal crystallographic axes that intersect at 120°, in a plane perpendicular to a fourth axis of different length), clear or white, and are optically distinguishable from apatite, an important phosphate mineral discussed later (see below *Apatite series*).

Triphylite and lithiophilite. Triphylite ($LiFePO_4$) is one end member of a solid solution series (a single crystalline phase that varies in chemical composition within finite limits—between end members) in which lithium and ferrous iron combine with the phosphate anion. Lithiophilite ($LiMnPO_4$) is the other end member, manganese replacing the iron. Physical properties vary with composition, but hardness typically is 4.5, and density is about 3.3. Colour ranges from bluish-green for the ferrous member to yellow-brown for the manganese member. The solubility of the minerals in acid and its fusibility are distinctive tests. Triphylite and lithiophilite are orthorhombic (referable to three mutually perpendicular crystallographic axes of unequal length) and have a structure similar to that of olivine, an iron–magnesium silicate. Thus, the phosphorus has tetrahedral (fourfold) coordination, and both the lithium and iron or

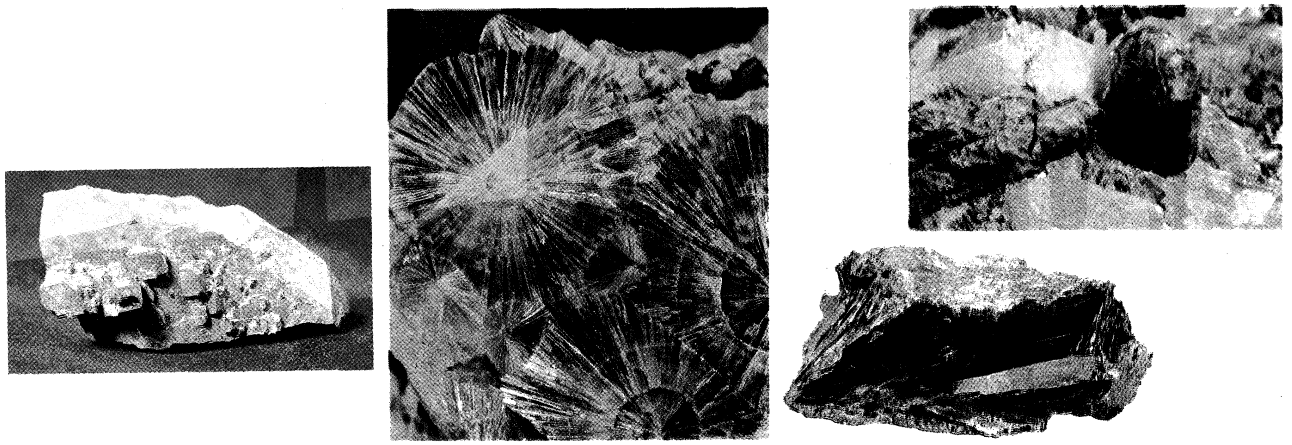


Figure 1: (Left) Monazite from Elk Mountain, New Mexico. (Centre) Wavellite from Buckville, Arkansas. (Top right) Fluorapatite on limestone from Renfrew, Ontario, Canada. (Bottom right) Vivianite from Wannon River, Victoria, Australia.

By courtesy of the (left) American Museum of Natural History, (centre) Ted and Elsie Boente collection, (top right) MacFall collection, (bottom right) American Museum of Natural History; photographs, EB Inc., (left) Emil Javorsky, (centre) John H. Gerard, (top right) Mary A. Root, (bottom right) Emil Javorsky

manganese atoms are surrounded by six oxygens, which is octahedral coordination.

Triphylite and lithiophilite occur in pegmatites and are commonly distinctive because of a dark altered surface or zone. Most material is massive, and rare crystals lack distinct faces.

Apatite minerals and their structure

Phosphate minerals containing hydroxyl or halogens. *Apatite series.* Apatite is the name for a mineral series that is subdivided with each name indicating the major halogen or hydroxyl; the general formula is $Ca_{10}(PO_4)_6(OH,F,Cl)_2$. Thus, there is fluorapatite (containing fluoride), chlorapatite (containing chloride), and hydroxylapatite (containing hydroxyl). The mineral exhibits toleration of extensive substitutions for phosphate and for calcium. Some of these chemical variants have mineral species names. Besides the chemical variants of the calcium phosphate apatites, numerous compounds can be synthesized that exhibit a structure identical or closely related to that of apatite. The name collophane is sometimes given to the massive, seemingly noncrystalline apatite that is the principal constituent of phosphate rock.

Most good crystals of apatite are fluorapatite, which occurs disseminated in igneous and metamorphic rocks and in ore deposits. Crystals are euhedral hexagonal prisms that generally range from clear to various hues of yellow and green; crystals of all colours occur in nature, however. The mineral serves as the reference material for the hardness of 5 on Mohs scale. It has a density of 3.2 and is distinguished from beryl and quartz on this basis and by reason of its hexagonal form and inferior hardness.

The apatite structure (Figure 2) is basically a hexagonal column with six calcium ions lying at each corner of the hexagon in a plane. Each of these calcium ions is surrounded by eight oxygens. At the centre of the hexagonal column lies the fluoride ion that is surrounded by three calcium ions, in a plane parallel to that containing the six calcium, but at a different level in the structure. Two phosphorus ions lie off each join or connection of two of the calciums of the hexagonal ring. Each phosphorus ion shares oxygens with calcium of both the sixfold ring and the threefold group. Thus, the structure appears as a framework with hexagonal form and a central cavity; this cavity is blocked twice in a repeat (unit) distance by centrally located fluoride that is strongly bonded to three calcium ions, however. The great stability of this structure is shown by the lack of polymorphic transitions (changes to other forms) with heating and by a melting point in excess of 1,600° C (2,900° F).

Good crystals of hydroxylapatite are extremely rare. They occur in three known localities in talc schists (foliated metamorphic rocks) or serpentine rocks. Their rarity indicates that crystallizing apatite is a scavenger for minute quantities of fluoride in the system—which is why fluorapatite predominates in nature.

Hydroxylapatite is similar to fluorapatite in properties and structure. Structural changes involve only the substi-

By courtesy of the U.S. Geological Survey

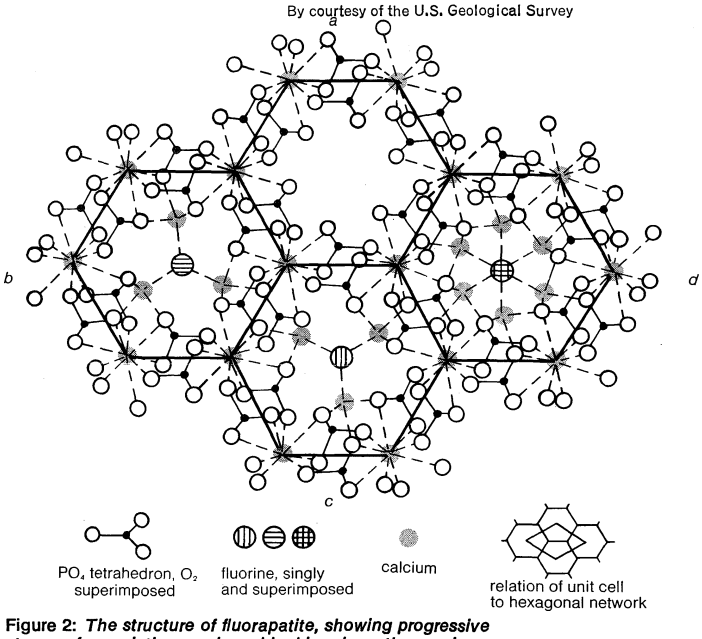


Figure 2: *The structure of fluorapatite, showing progressive stages of completion as viewed looking down the c axis.* (*a*) Incomplete structure, showing hexagonal framework and unoccupied channels; (*b, c*) addition of the two planar groups of three channel-calcium atoms and one fluoride atom, each at a different vertical height in the structure; and (*d*) complete structure.

tution of hydroxyl for fluoride. Either the fluoride or the oxygen of the hydroxyl lies in the plane containing the three calciums; the hydrogen of the hydroxyl lies on an axis normal (perpendicular) to the plane, however. Thus, with the central axis of the hexagonal structure vertically oriented, the hydrogen is either above or below the oxygen. In an ideal crystal all of the hydrogens of the hydroxyl would tend to point either down or up. In real crystals it is likely that the column containing the hydroxyls has domains such that for part of the column the hydroxyls are oriented down and for other parts the direction is up. Such directional orientation destroys the centre of symmetry and imparts to the crystal the property of piezoelectricity (the ability to develop electric surface charges when subjected to stress).

Hydroxylapatite and chlorapatite

Chlorapatite is rare. It is reported from some meteorites, and it occurs in gabbro (igneous rocks rich in iron and magnesium) and related rocks in southeastern Nor-

way. The density of chlorapatite is about 3.2, and the structure is similar to that of other apatites but differs in the position of the halogen. Because the chloride ion is 1.8 Å (one angstrom unit [Å] equals 10^{-8} centimetre) in size, it lies between two of the triangular calcium planes and not within the plane, as does the fluoride ion. Thus, chloride is surrounded by six calciums compared to three for fluoride.

Most apatite from igneous and metamorphic rocks is fluorapatite; that is, it contains approximately two fluorides per unit cell. A complete series between fluorapatite and hydroxylapatite is possible, however, and can be synthesized in the laboratory. Intermediate members of the series are rare in nature.

Chlorapatite is not structurally similar to either fluorapatite or hydroxylapatite because chloride occupies a different site. Natural crystals are known that contain appreciable fluoride and chloride. Furthermore, these crystals have physical properties that fall between those of fluorapatite and chlorapatite. This indicates either that (1) the fluorapatite structure can accommodate appreciable chloride substitution, even though complete substitution is impossible, (2) that the crystals are disordered (irregularity of arrangement involving some randomness in site filling), or (3) there are domains or layers that alternate from the fluorapatite to the chlorapatite structure.

Trace elements in apatite

Apatite is a receptive host for trace elements—*i.e.*, elements present in very minor amounts; and it accommodates extensive substitution in both the cation and anion site. Apatite is especially receptive to substitution of divalent cations larger than calcium; for example, strontium and barium. Apatite is less tolerant for divalent ions smaller than calcium, such as magnesium, which inhibits the formation of apatite. Although the apatite structure tolerates some substitution of manganese, cadmium, and other base metals, concentration is generally small because of their low natural abundance. Apatite contains, at trace element levels, arsenic and vanadium as the arsenate and vanadate ions substituting for phosphate.

Some substitutions in apatite involve a concomitant exchange of several different ions in order to maintain a charge balance. Sodium and sulfate, for example, may substitute for calcium and phosphate with no change in net charge. Likewise, sulfate and silicate may exchange for two phosphates with no change in net charge.

The carbonate problem

Many crystals of apatite show some effervescence of carbon dioxide when dissolved in acid. A number of hypotheses have been advanced to account for this carbon dioxide; basically, they fall into two groups. One postulate is that the carbon dioxide occurs in the structure, and the other is that it occurs adsorbed or absorbed on the crystal surfaces or is present as a calcium carbonate phase. Both hypotheses have supporting data, and these have been obtained primarily on different materials. Carbonate-apatite formed under metamorphic or hydrothermal conditions is well crystallized and of sufficient size to be amenable to detailed optical and X-ray examination. The crystal size makes untenable any model involving large surface areas. Carbonate-apatite formed at low temperatures, however, is poorly crystallized and of minute size, yielding a large surface area. For this material, models involving surface sorption or a poorly crystallized, admixed, second phase are plausible.

The carbonate ion is a plane of three oxygens with a central carbon and a net charge of minus two. There is no identical arrangement in the apatite structure. An early hypothesis was that the carbonate group replaces and therefore excludes halogen. This was questioned because of the inadequate size of the site for the carbonate group and because analyses showed more than enough halogens to fill the halogen sites. Polarized infrared studies of material, prepared by a high-temperature treatment of chlorapatite under a pressure of one atmosphere of carbon dioxide, lend credence to this substitution.

The three oxygens of the carbonate group are identical in form to the three oxygens found on any face of a phosphate tetrahedron. It has been proposed that a carbonate group replaces a phosphate tetrahedron, yielding a vacancy that formerly was occupied by the fourth oxygen of the tetrahedron. This vacancy may be filled with a fluoride ion, a substitution that maintains charge neutrality and can account for the excess fluoride shown in some analyses. Or the vacancy may persist, and the charge difference of a carbonate may be balanced by sodium substituting for calcium, a coupled substitution. The basic premise that carbonate substitutes for phosphate is supported by many chemical analyses of natural carbonate-apatite, which show an inverse relationship of phosphate and carbonate content (see further CARBONATE MINERALS).

Water content

Analyses of apatite that formed at low temperatures show as much as four times the amount of water that can be assigned to the hydroxyl site. As with carbonate, some water may be sorbed and thus related to the large surface areas. Other crystals of moderate size contain excess water, however, indicating that sorption is not the only process involved.

It has been proposed that water is incorporated in the apatite structure in three sites. First, water as the hydronium ion (H_3O^+) may substitute for calcium. It is further postulated that the hydronium ion, after incorporation in the structure, may lose a water molecule, leaving behind a hydrogen ion, or proton, in a calcium site. Second, water may be incorporated in the structure as a tetrahedral form of four hydroxyls ($H_4O_4^{4-}$). The $H_4O_4^{4-}$ tetrahedron can be charge compensated by a concomitant carbonate substitution; it is efficient for the incorporation of water in the structure; and it can fill phosphate sites in phosphorus deficient apatites. Third, it is postulated that water displaces a hydroxyl. This complete displacement could yield an apatite containing 3.6 percent water, but it would require a charge compensation. The HPO_4^{2-} ion may substitute in apatite, but this ion is inefficient for the incorporation of much water in the apatite structure. It is likely that water is incorporated in apatite by a combination of the methods described and that no single model can account for all of the apatites with excess water.

Pyromorphite. Pyromorphite [$Pb_{10}(PO_4)_6Cl_2$] is a member of a phosphate group containing an arsenate and vanadate equivalent, mimetite and vanadinite, respectively. All are closely related to chlorapatite, although some are monoclinic, as is a polymorph of chlorapatite. Pyromorphite is olive-green, yellow, or gray; and it has a hardness of 3.5 to 4 and a density of 7. Crystals appear as hexagonal prisms, commonly tapered. The mineral occurs in the oxidized zone of lead ore deposits.

Amblygonite. Amblygonite [$(Li,Na)Al(PO_4)(F,OH)$] is triclinic (referable to three inclined crystallographic axes of differing lengths), and it occurs as equant to short prismatic crystals. The colour is white to creamy white, the hardness 5.5 to 6, and the density is 3.11. It fuses easily and is soluble in acid with difficulty. Amblygonite occurs in granitic pegmatite and can be mistaken for albite, a sodium-rich feldspar mineral.

Lazulite. Lazulite [$(Mg,Fe)Al_2(PO_4)_2(OH)_2$] means blue stone, and this colour is its most distinctive feature. It is monoclinic, has a hardness of 5.5 to 6, and a density of 3.08. Lazulite occurs in granitic pegmatites and aluminum and silica-rich metamorphic rocks.

Phosphate minerals containing water. *Brushite.* Brushite is very similar to gypsum ($CaSO_4 \cdot 2H_2O$) in form, structure, and physical properties, but in its composition HPO_4^{2-} replaces SO_4^{2-} of the gypsum, thus yielding the formula $CaHPO_4 \cdot 2H_2O$. Precipitated gypsum has a long prismatic or needlelike form, and usually the crystals are twinned. Brushite has a blocky prismatic form, and twinning, the intergrowth of two or more grains of the same crystalline phase according to some crystallographic pattern, is much rarer than in gypsum.

At low temperature, brushite loses both water molecules and forms monetite ($CaHPO_4$). Both monetite and brushite are rare minerals, found in phosphorites formed from guano, and are the only phosphate minerals containing principally the HPO_4^{2-} ion, even though this ion is the most abundant phosphate ion in natural aqueous solutions.

Occurrence of the Phosphate Minerals

	igneous rocks*		pegmatites	hydrothermal veins	metamorphic rocks†		sedimentary rocks‡			alteration products
	basic	acid		ore deposits	high grade	low grade	detrital	indigenous	authigenic	
Calcium phosphates										
Fluorapatite	X	X	X	X	X	X	X	X	X	
Hydroxylapatite						X			X	
Chlorapatite	X									
Whitlockite			X						X	
Brushite									X	
Monetite									?	X
Anhydrous phosphates										
Berlinite (Al)				?						
Xenotime (Y)		X	X							
Monazite (Ce,La)		X	X		X		X			
Triphylite (Li)			X							
Hydrous phosphates										
Amblygonite (Li)			X							
Lazulite			X	X	X					
Hydrous phosphates with water										
Wavellite				X rare		X rare			X	X
Vivianite										X
Turquoise										X
Autunite										X
Other										
Pyromorphite										X

*Acid igneous rocks are those rich in silica and basic igneous rocks are those rich in iron and magnesium. †The term grade in metamorphism refers to the extent or degree of alterations of a rock from its original state. ‡Authigenic minerals in sedimentary rocks are those formed in place, during the transformation of sediments to rock; this is in contrast to detrital minerals, which are transported grains of pre-existing rocks.

Vivianite. Vivianite [$Fe_3(PO_4)_2 \cdot 8H_2O$] is monoclinic, with phosphate tetrahedrons linked with ferrous iron in octahedral coordination. This linking yields a sheet structure having a perfect cleavage in one direction; it is rare to find crystals or massive material lacking this cleavage. Vivianite, colourless when first mined, darkens rapidly to a very deep blue, or purple; it has a hardness of 1.5 to 2, a density of 2.7, and occurs as bladelike crystals.

Vivianite occurs in gossans (hydrated iron oxide deposits that usually form above iron sulfide veins or ore bodies) or from the weathering of primary phosphate minerals of pegmatites. It is present in fossil teeth, bones, and in phosphatic shells or hard parts of organisms, to which it imparts a blue colour.

Phosphate minerals containing water and hydroxyl. *Wavellite.* Wavellite is an aluminum phosphate containing water and hydroxyl, and it has the formula $Al_3(PO_4)_2(OH)_3 \cdot 5H_2O$. It is a secondary mineral found in near surface deposits of alumina-rich rocks. The mineral commonly occurs as concretions, starlike forms, and masses that radiate out from a common centre; it is associated with limonite, an iron oxide. Wavellite has a hardness of 3.25 to 4, a density of 2.37, and crystallizes in the orthorhombic system. Wavellite tolerates limited substitution of ferric iron for aluminum, and this results in a colour range from yellowish-tan to light green.

Turquoise. Turquoise is a copper and aluminum phosphate containing water and hydroxyl with the formula $CuAl_6(PO_4)_4(OH)_8 \cdot 4H_2O$. It is a secondary mineral prized as a semiprecious material for polishing and carvings. Most turquoise occurs as cryptocrystalline-massive nodules or vein fillings; crystals are extremely rare. The mineral is triclinic (referable to three inclined crystallographic axes of unequal lengths) and is distinctive by reason of its green hue and waxy lustre. Turquoise is commonly associated with limonite, chalcedony, and kaolinite (a clay mineral). Bone turquoise is a name incorrectly applied to fossil teeth and bone coloured with vivianite.

Copper and uranium phosphates

Autunite. Autunite is a secondary uranium phosphate mineral with the formula $Ca(UO_2)_2(PO_4)_2 \cdot 10\text{–}12H_2O$. It crystallizes in the tetragonal system, and specimens appear tabular, foliated, and even micaceous (in tablets, thin plates, and sheets). The mineral is distinctive by reason of its form, bright yellow to green-yellow colour, and strong fluorescence. Autunite readily loses water and inverts to metaautunite, a compound with two to six water molecules in its formula.

OCCURRENCE OF PHOSPHATE MINERALS

Phosphorus ranks tenth in abundance of elements in the crust of the Earth, but its absolute abundance is only about 0.1 percent by weight of all crustal material. Calcium is 32 times more abundant than phosphorus, and calcium phosphates are therefore likely to form. Sodium, magnesium, and aluminum also are abundant elements, but phosphate compounds of these elements are limited because of their high solubility and small range of thermal stability. Phosphate compounds of lithium, yttrium, cerium, lanthanum, copper, and uranium are rare because of the low natural abundance of these cations. The occurrence of phosphate, therefore, can be related principally to environmental temperatures in nature and to those rocks the formation of which involved enrichment of rare elements (see the Table).

Natural environments. *Calcium phosphates.* Fluorapatite is ubiquitous in terrestrial rocks. It is stable under conditions of elevated temperatures and pressures, and it can be synthesized at any temperature below its melting point (1,600° C, or 2,900° F). The wide distribution results from the abundance of calcium in most environments. Fluoride is present in small concentrations, but the crustal average of fluoride is about three times that necessary to combine with phosphorus to yield fluorapatite. Fluorapatite forming in seawater appears an enigma because of the low concentration of fluoride compared to the abundance of water. It is hydroxyl, not water, that competes with fluoride in the apatite structure, however, and the hydroxyl activity (a thermodynamic measure of concentration) occurring in seawater is only one-tenth to one-hundredth of the fluoride activity.

Derivation of fluoride

Hydroxylapatite occurs only in talc schists and a serpentinite. The mineral is absent in igneous rocks, even though it is thermally stable to high temperatures. It is possible that hydroxylapatite results from basic solutions active during metamorphism of the rocks. A carbonate-bearing hydroxylapatite is the mineral constituent of bone and teeth, and with fossilization such material becomes enriched in fluoride, forming a fluorcarbonate apatite. This exchange is so slow that more than a century ago it was proposed as a crude age-dating method. Although unsatisfactory for such purposes it does demonstrate that hydroxylapatite can scavenge fluoride from dilute solutions.

Chlorapatite is very rare and is limited to gabbroic rocks and meteorites. Chlorapatite is absent in marine phosphorites, even though the chloride concentration of sea-

water is thousands of times that of fluoride. Similarly, chlorapatite can be synthesized at high temperatures, but it cannot be synthesized from an aqueous solution at low temperatures. Because the halogen or hydroxyl of apatite is bound to calcium, an explanation may lie in the relative solubilities of calcium fluoride, hydroxide, and chloride in aqueous solutions. Pyromorphite, being a lead chloride apatite, indicates a high chloride environment during its formation as an alteration product of galena (lead sulfide).

Whitlockite differs chemically from apatite because it contains more phosphorus and is anhydrous. It occurs in a pegmatite with apatite, an environment in which there is little calcium. It also occurs in guano deposits in which highly phosphate-rich leachates react, under acid condition and near surface temperature, with limestone. Brushite also forms by the same process. Laboratory experiments show that the pH (a measure of the acidity of solutions ranging from 0 to 14; acid solutions have a pH of less than 7, whereas the pH of alkaline solutions is greater than 7) for brushite formation must be less than 6.5; at higher pH values, apatite forms. Monetite cannot be synthesized from simple aqueous solutions at 25° C (77° F), suggesting that its natural occurrence may be from dehydration of brushite.

Anhydrous phosphates. Xenotime, monazite, and triphylite are anhydrous phosphates, each of which contains an element with crustal abundance of only 20 to 60 parts per million. Being anhydrous, the minerals are stable at high temperatures, but they cannot form until crystallization of a magma (high-temperature silicate melt within the Earth) enriches the residual melt in these rare elements. Consequently, their occurrence is principally in acid igneous rocks and pegmatites.

Lithium requirements

Hydrous phosphates. Amblygonite and lazulite are hydrous phosphates and therefore unstable in dry magmatic systems (*i.e.*, those without water), and, in addition, amblygonite requires a system rich in lithium. Lazulite could form in a hydrous magmatic system, but there is generally sufficient calcium, which has a greater affinity for phosphate, to form apatite instead of lazulite. Thus, lazulite's formation in igneous rocks is limited by the availability of phosphorus, which is controlled by calcium, whereas amblygonite is limited in temperature and the availability of lithium. Pegmatites present an environment for both minerals. Lazulite may also form in quartz veins and calcium-poor metamorphic rocks.

Hydrous phosphates with water. Wavellite, vivianite, turquoise, and autunite are minerals that contain from four to 12 waters of crystallization (water molecules that form part of the crystal structure), and all lack stability above 200° C (390° F). Thus, they occur under near-surface conditions and commonly as an alteration product of other minerals. Wavellite has wide distribution as a secondary mineral in weathered or reworked phosphorites. Vivianite contains ferrous iron and therefore requires reducing conditions (an oxygen-free environment) for formation. Consistent with this requirement is its occurrence with peat, lignite, organic sediments, and glauconite. Autunite contains uranium and therefore is common as a weathering product of uraninite. Turquoise is formed by supergene (downward percolating) acid solutions reacting with apatite and copper sulfide ores in alumina-rich rocks.

Experimental studies. Most experimental studies of phosphate minerals have been concerned with a limited aspect of crystal chemistry or solubility. A further limitation is that most studies are on simple systems or under conditions such that the results have little application to the environments of formation.

The binary system $CaO–P_2O_5$ has been studied extensively, but the only compound to be synthesized from it having a natural mineral equivalent is whitlockite. This compound undergoes a polymorphic transition at about 1,200° C (2,200° F) to another form that may possibly be pertinent to the occurrence of whitlockite in some meteorites.

The binary systems $CaCl_2–Ca_3(PO_4)_2$ and $CaF_2–Ca_3(PO_4)_2$ were studied more than a half century ago. Fluorapatite has been shown to have a melting point in excess of 1,600° C (2,900° F), and chlorapatite is reported to melt slightly above 1,500° C (2,700° F); the polymorphism of chlorapatite was unrecognized in the early work, however.

Studies of the ternary systems $CaO–P_2O_5–H_2O$, $CaO–P_2O_5–HF$, and $CaO–P_2O_5–HCl$ or the fluoride or chloride salts of the acids are so incomplete, especially at elevated pressures, that generalizations are unjustified.

Chemical variants of apatite have received much study because of their biological and geological importance. Many crystal–chemical studies relate substitution of various ions to crystallinity and dimensional changes, length of the a_0 and c_0 lattice parameters (see Figure 3). Lattice

Chemical composition and lattice change

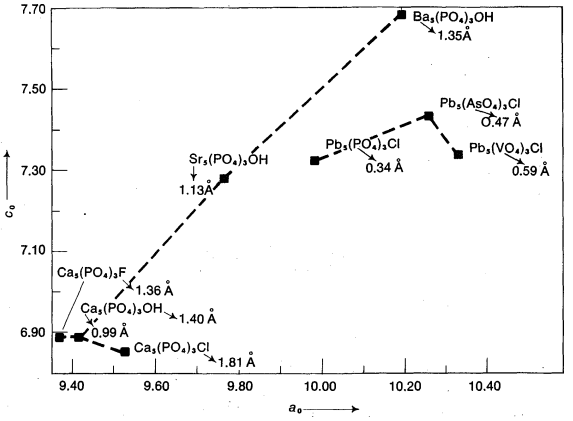


Figure 3: Relation of length of lattice parameters a_0 and c_0 in apatite and related compounds, and ionic substitution. Note that substitution of a large cation for Ca results in an equal increase of a_0 and c_0, whereas substitution in the halogen site results in an increase of a_0 and a decrease of c_0 (see text).

parameters of apatite and related compounds show that: (1) substitution of an increasingly large cation in the calcium site causes about an equal increase in a_0 and c_0; (2) substitution of a large ion in the halogen site causes an increase in the a_0 direction but a decrease in c_0; (3) and, for the pyromorphite series, increased ion size in the phosphate site yields increased a_0 and slightly increased c_0.

The many studies concerned with the solubility of apatite have produced a multitude of values because of the chemical variants of the mineral. Application of the data to natural systems is plagued by insufficient knowledge of the activity of the various species of phosphate ions in a complex solution.

The polymorphism of berlinite, $AlPO_4$, has been studied in detail because of its similarity to the silica system. Berlinite transforms at 814° C (1,500° F) to a tridymite form, with succeeding transformation at 1,025° C (1,877° F) to a cristobalite form that melts in excess of 1,600° C (2,900° F). (Tridymite and cristobalite are high temperature forms of quartz.) Of the $AlPO_4$ polymorphs, only berlinite occurs naturally.

BIBLIOGRAPHY. CHARLES PALACHE, HARRY BERMAN, and CLIFFORD FRONDEL, *Dana's System of Mineralogy*, 7th ed., vol. 2 (1951), an authoritative reference source for all phosphate minerals; W.A. DEER, R.A. HOWIE, and J. ZUSSMAN, *Rock-Forming Minerals*, vol. 5, *Non-Silicates* (1962), provides a description and discussion of apatite and monazite; J.R. VAN WAZER "Phosphorus and Its Compounds," *Chemistry* 1:419–556 (1958), a comprehensive review of the chemistry of phosphorus compounds with a chapter on condensed phosphates and salts of orthophosphoric acid; R.W.G. WYCKOFF, *Crystal Structures*, 2nd ed., vol. 3 (1965), includes a summary of structural data for phosphate minerals.

(D.R.S.)

Phosphorites

Phosphorus is an element that is present in about 200 minerals in amounts up to 5 percent or more. Most of the phosphorus in the earth's crust and in phosphorites, which are minable phosphate deposits, occurs in minerals of the

apatite family of calcium phosphates (see PHOSPHATE MINERALS). Although phosphorus is an essential constituent for all living organisms, it is one of the least abundant of the mineral nutrients. Its average content in soils and rocks of the earth's crust is 0.12 percent, whereas that of iron is 5 percent, of calcium 3.6 percent, of potassium 2.6 percent, and of magnesium 2.1 percent. Thus, when phosphorus is removed from the soil by crops, it must be replaced artificially by the addition of phosphate fertilizers. Phosphorus is also used widely in manufacturing industries. Nearly all of the phosphate fertilizers and phosphorus for industrial uses are obtained from naturally occurring rocks that contain high concentrations of phosphate-bearing minerals. Thus the availability of large quantities of phosphorites is of primary importance to agricultural and industrial nations.

Uses of phosphate

Seventy percent or more of all phosphate is used in most countries for agricultural purposes; most of this is made into fertilizers. The rest is used in making elemental phosphorus and phosphoric acid, which are, in turn, made into organic and inorganic chemicals used in preparing detergents, water softeners, leavening agents (for baking), plasticizers, insecticides, and many other products. Elemental phosphorus is also used in military smoke screens and incendiary bombs, phosphorus alloys, and pesticides. In addition, phosphoric acid is used for fertilizers, animal-feed supplement, beverages, photographic chemicals, dental and silicate cements, ceramics, catalysts, and an oil-refining agent. A complete list of phosphate compounds would embrace almost every manufacturing process and nearly all chemical techniques.

Principal types of phosphate deposits. Mineral deposits of phosphate are of six principal types, listed in approximate order of size and world importance: (1) marine phosphorites; (2) igneous apatite deposits; (3) residual phosphorites; (4) phosphatized rock; (5) river pebble; and (6) guano.

Marine phosphorites are blanket-like deposits that form on the sea floor as the result of organic or inorganic precipitation of carbonate-fluorapatite from seawater, generally in low latitudes in areas where cold, phosphate-rich water upwells onto the continental shelves. They are commonly associated with beds of black carbonaceous shale, chert or diatomite, sandstone, and minor amounts of limestone. The phosphatic layers consist mainly of pellets and nodules ranging from less than 0.1 to 20 millimetres or more in diameter. Fish teeth, scales, and bones, phosphatic brachiopod shells, and casts of small pelecypods and other fossils are common in many deposits. Individual layers are as much as 15 feet or more in thickness, may extend over hundreds of square miles, and contain millions or even billions of tons. Unweathered deposits, including those found on the present sea bottom, commonly do not contain more than about 30 percent phosphorus pentoxide (P_2O_5), but deposits that have been enriched by the leaching and oxidation of impurities may contain as much as 38 percent. Less phosphatic sandstones and shales are upgraded by separating the phosphatic particles from the matrix by screening, washing, or flotation. The marine phosphorites and some of the associated black shales contain many minor metals, some of which may be recoverable as by-products. Examples are vanadium, uranium, zinc, rare earths, and selenium.

Igneous apatite deposits are found as inclined sheets or irregular masses of intrusive igneous rock, veins, or replacement bodies, which are derived from the cooling and crystallization of silicate melts within the earth's crust. They are associated with other rocks of igneous origin, such as nephelite-syenite and carbonatite, and, in many places, with magnetite or ilmenite ores. The apatite masses range in size from thousands to millions of tons and may contain as much as 35 percent P_2O_5 (in the fertilizer industry the phosphate content is sometimes also reported as bone phosphate of lime [BPL] or tricalcium phosphate [TCL], both of which are equal to 2.18 times the P_2O_5 content). The richest deposits are used without beneficiation. The phosphate in some of the apatite-bearing ilmenite and magnetite ores is recovered as a by-product of the milling or smelting of the titanium or iron ore.

Residual phosphate deposits develop by decomposition of phosphatic limestones, located within humid and tropical climatic regions. These deposits, consisting of phosphate pellets and nodules left behind when the calcium carbonate of the limestone is removed by solution, lie on an irregularly weathered bedrock surface. In hilly country, residual phosphate deposits are narrow and conform to the outcrop pattern of the strata; but in valleys or flat country, the deposits are blanket-like. These deposits may be several feet or more in thickness and contain 30 to 38 percent P_2O_5. The higher grade deposits are utilized directly, but lower grade deposits must be beneficiated by washing, screening, or flotation in order to yield a salable product. Phosphate reserves in residual deposits amount to many millions of tons.

Phosphatized rocks develop in regions of tropical and subtropical climate when phosphate dissolved from phosphatic rocks or guano by percolating water is again deposited, by reprecipitation, in openings or replaces the minerals of limestone, igneous rocks, or clay that lie at a lower horizon. Where the phosphatized rocks are limestone, the deposits consist of apatite; but where the phosphatized rocks are igneous rocks or clay, the deposits are iron or aluminum phosphates such as strengite, crandallite, and wavellite. Some phosphatized rocks contain as much as 38 percent P_2O_5, and reserves in individual deposits may be scores of millions of tons.

River-pebble deposits consist of phosphate pellets and nodules weathered out of marine phosphatic limestone, located in humid climatic regions of low relief. The pellets and nodules are reconcentrated by stream action. Most river-pebble deposits do not contain sufficient phosphate to be mined profitably, but those in Florida have been mined and beneficiated cheaply enough to produce a commercial product. River-pebble deposits range from a few thousand to many millions of tons in size.

Guano deposits

Guano deposits form by the accumulation of seafowl or bat droppings. The largest deposits are found on islands in low latitudes along western margins of continents and in the zone of the equatorial countercurrent. Guano deposits generally are small; the larger contain a few tens of millions of tons at the most. Modern guano contains 10 to 12 percent P_2O_5; old leached guano contains 20 to 32 percent P_2O_5. Slightly decomposed deposits contain soluble ammonium alkali oxalates, sulfates, and nitrates and a variety of magnesium and ammonium magnesium phosphates; but decomposed guano consists largely of calcium phosphates such as monetite, brushite, and whitlockite.

Phosphate deposits in the United States. Examples of each of the principal types of phosphorites are present in the U.S. Although phosphate concentrations are known in 28 states, production has come largely from six fields: (1) land-pebble deposits of west central Florida; (2) phosphatized (hard-rock) deposits of Florida; (3) north Florida–south central Georgia; (4) North Carolina; (5) brown-rock deposits of Tennessee; and (6) marine phosphorites and phosphatic shales of Idaho, Montana, Utah, and Wyoming. Potentially important deposits occur also in Alaska, western California, and on the sea bottom off the California coast.

The land-pebble phosphates of Florida are blanket-like deposits covering about 2,600 square miles of west central Florida. They consist partly of residual pellets and nodules derived from phosphatic limestone of the Hawthorn Formation of Miocene age (7,000,000 to 26,000,000 years ago) and partly of residuum reworked and redeposited in the overlying Bone Valley Formation of Pliocene age (2,500,000 to 7,000,000 years ago). The phosphate deposits range from 1 to 50 feet thick and average 12 feet. They consist of nearly equal amounts of phosphate pellets and nodules, sand, and clay. The phosphate zone is overlain by as much as 60 feet of overburden, which is removed by large draglines before the phosphate is mined by similar methods. Phosphorite particles are beneficiated by washing, screening, and flotation methods. In areas of deep weathering, the upper part of the deposits has been leached and clay, in the matrix, has been altered to aluminum phosphate. River pebble derived from erosion of the Bone Valley deposits was once a source of substantial

production but is no longer mined. The hard-rock phosphate deposits of northwestern peninsular Florida are phosphatized limestone fragments and plates in the Alachua Formation of Pliocene age. These deposits are small, but some contain as much as 35 percent of P_2O_5. They are mined and beneficiated by the same methods as are used in the land-pebble fields.

The Tennessee brown-rock deposits, located in the southwestern part of Tennessee, are residual deposits derived from the weathering of phosphatic limestones of the Hermitage, Bigby, and Leipers formations, all of Ordovician age. The deposits conform closely to the outcrop pattern of the parent limestones on the rolling, hilly topography. The thickest phosphate deposits fill solution-enlarged joints (locally called cutters) in the phosphatic limestones. These deposits also are mined by open-pit methods and are beneficiated to produce a marketable product.

The marine phosphorites and phosphatic shales of the western field are in the Phosphoria Formation of Permian age. Phosphate rock is mined in Idaho, Wyoming, Montana, and Utah by open-pit and underground methods. Minable layers range from 3 to 20 feet in thickness and contain from 20 to 32 percent P_2O_5.

The north Florida–south central Georgia and North Carolina phosphorites are marine deposits of Miocene age, first developed during the early 1960s.

Minable reserves of phosphorites in the U.S. are about several billion tons, and reserves of lower-grade deposits are many times larger. About one-quarter of U.S. mine production is exported.

Phosphate deposits in other countries. The principal sources of phosphate rock outside the U.S. until the early 1950s were the marine phosphorites of Morocco, Tunisia, Algeria, and the U.A.R.; the igneous apatites of the Kola Peninsula, U.S.S.R.; and phosphatized rocks derived from the leaching of guanos on some islands of the Pacific and Indian oceans (particularly Ocean, Nauru, Makatea, and Christmas islands). Since then, large deposits of marine phosphorites have been discovered in Jordan, Israel, Syria, the Spanish Sahara, Senegal, Togo, Angola, Australia, Mexico, and Peru, and deposits that are or may be adequate for domestic needs occur in other countries. World reserves of phosphorites are shown in the Table.

World Reserves and Potential Resources of Phosphate Rock and Apatite

country	million long tons of P_2O_5
Morocco	21,000
U.S.	14,500
U.S.S.R.	7,600
Tunisia	2,000
Algeria	1,000
Brazil	600
Peru	450
Egypt	200
Togo	120
Spanish Sahara	100
Islands of the Pacific and Indian oceans	40
Senegal	40
Other countries	700
Total	48,350

Source: U.S. Department of the Interior, *Mineral Facts and Problems, 1965.*

BIBLIOGRAPHY. Z.S. ALTSCHULER, R.S. CLARKE, JR., and E.J. YOUNG, "Geochemistry of Uranium in Apatite and Phosphorite," *Prof. Pap. U.S. Geol. Surv. 314-D*, pp. 45–87 (1958); BRITISH SULPHUR CORPORATION, *A World Survey of Phosphate Deposits*, 2nd ed. (1964), includes brief descriptions of all individual deposits known at time of compilation; V.E. MCKELVEY, "Phosphate Deposits," *Bull. U.S. Geol. Surv. 1252-D* (1967); UNITED NATIONS, ECONOMIC COMMISSION FOR ASIA AND THE FAR EAST, *Proceedings of the Seminar on Sources of Mineral Raw Materials for the Fertilizer Industry in Asia and the Far East* (1969), contains 55 articles on phosphate deposits and fertilizer requirements.

(V.E.McK.)

Photius

Though he did not intend to initiate a schism with the West or with the Roman see, Photius, the patriarch of Constantinople from 858 to 867 and again from 877 to 886, became the focus of Byzantine pretensions to cultural superiority, which led to the Photian Schism. He was responsible for the first list of Latin theological deviations from the standpoint of Greek Orthodoxy. The most notable scholar among the Byzantine ecclesiastical authorities and civil servants who preserved the classics of ancient Greece—both pagan and Christian—he was also the author of a lexicon and numerous theological works.

By courtesy of the Dumbarton Oaks Research Library and Collection, Washington, D.C.



Photius, lead seal. In the Dumbarton Oaks Research Library and Collection, Washington, D.C.

Family background and early career

Photius was related through his father to Tarasius, a civil servant who was patriarch from 784 to 806, and through his mother to the empress Theodora and her brother Bardas, who was a power behind the throne from 842 to 866. The circle to which both families belonged had strained relations with some army officers, who wished to reduce the expense of government by an educated civil service, and with monks who distrusted the worldly wisdom of bureaucrats. In controversies that had arisen over the place of pictures (icons) in Christian worship, the learned civil servants resisted the iconoclasts (the destroyers of sacred images), who had support in the army. When the icons were restored in 787, however, and again after renewed controversy in 843, the civil servants were lenient with the former iconoclasts, whereas the monks, especially the Studites of the monastery of St. John of the Stoudion would have purged them, as dangerous heretics, from all important positions.

The parents of Photius were condemned to exile by an iconoclast council, probably in 815, and took him with them. Since his birthdate is not known, it is not clear whether he was a small child or a student at the time. Early in the regency of Theodora (842–856), after his parents had died in exile, Photius became a distinguished teacher. A circle gathered round him for regular readings in classical and Christian prose literature, including medical and scientific works. On the basis of notes taken at these readings, which continued after he left the schools for the civil service, he composed his *Myriobiblon* or *Bibliotheca*, a digest of Greek prose literature, with more than 270 articles. This work was begun on a diplomatic mission in the Muslim world and most probably completed during his temporary retirement from public life after 867.

Promotion to patriarch of Constantinople

He became first secretary of state, probably before 855, and in 858 he was promoted through all the ecclesiastical orders to be made patriarch of Constantinople on Christmas Day, replacing the austere Ignatius, who had fallen out with Bardas. The deposition of Ignatius offended not only the Studites and other monks, who objected to the promotion of a civil servant, but also Pope Nicholas I, who did not understand the role of laymen educated in theology and in Byzantine civilization. Photius offended him further by refusing to restore dioceses transferred

from the Roman to the Byzantine patriarchate during the iconoclastic controversy. The importance of these dioceses had been increased by the conversion to Christianity of leading chiefs among the Slavonic nations (the Moravians, Croats, and Bulgarians); jurisdictionally they might belong to either the Roman or the Byzantine patriarchate. As conflicts developed among Roman, German, and Byzantine missions, Photius wrote a circular letter to the other Eastern patriarchs complaining of theological, liturgical, and other innovations by Latin missionaries in Bulgaria. At a council in Constantinople in 867, he condemned and excommunicated Nicholas I, who had refused to recognize him as the lawful patriarch—thus bringing about the Photian Schism—and in letters to other bishops had represented him as a persistent adversary of the West.

By this time, however, the fall of Bardas had weakened the position of Photius. When he protested against the murder of the emperor Michael III by Basil the Macedonian, he was deposed, in the autumn of 867, and Ignatius was restored. Pope Adrian II, who had just succeeded Nicholas I at Rome, now envisioned a settlement of the differences between Rome and Constantinople. The terms proposed by his legates to a council in Constantinople in 869–870, however, were unacceptable to many Byzantine ecclesiastics. Ignatius himself in 870 consecrated bishops for Bulgaria. Without help from the friends of Photius, however, he could neither reach a satisfactory settlement of tensions between East and West nor solve the internal problems of the Byzantine Orthodox Church.

Photius returned to the court before 876 as tutor to the princes of the imperial family, and at the death of Ignatius in 877 or 878 he also returned to the patriarchate. Though some of the monks resented his return, as they had objected to his first promotion, he now won support from Rome, since Pope John VIII was in need of Byzantine naval assistance against the Moors, who were harrying the Italian coastline. The Pope sent legates to a new council at the church of Hagia Sophia in Constantinople in 879–880. In the resulting settlement, Bulgaria was assigned to the Roman patriarchate, but the continued presence of Greek bishops secured its cultural links with the East. Bulgaria soon became a centre from which the Byzantine liturgy in the language of the Macedonian Slavs spread to other Slavonic-speaking lands.

Rome did not press its claims to Greek dioceses in Italy and Greece under the jurisdiction of Constantinople, and the Roman legates consented to the Byzantine demand to condemn Western additions to the Nicene Creed, without explicit mention of the contentious use of the word *filioque* (Latin: "and the Son"), whereby the Holy Spirit was said by some to proceed from the Father "and the Son." This interpolation had been introduced into the Nicene Creed in Spain and had spread among the Franks, but it was not yet in use in Rome. Photius' Latin was limited, and on the *filioque* controversy his information was inadequate, though he showed more understanding of the question in his later work on *The Mystagogia of the Holy Spirit*, completed in or after his second patriarchate.

The *filioque* controversy

The settlement between Rome and Constantinople was once thought to have been obtained by fraud and repudiated by Pope John VIII, but most likely it was accepted by the Pope. John VIII was murdered in 882, and his successor, whose position was irregular, was probably not recognized at Constantinople. Photius was in communion not only with John but also with Adrian III (884–885) and with Stephen V in 886. In that year, he resigned the patriarchate on the accession to the throne of his pupil the emperor Leo VI. The Pope suspected that the resignation of Photius had been forced upon him to make way for another pupil, Prince Stephen, but if he was over 80 years old it may well have been voluntary. He died on February 6, perhaps 891, but the year is not certain.

BIBLIOGRAPHY. FRANCIS DVORNIK, *The Photian Schism: History and Legend* (1948, reprinted 1970), supersedes the older view of the contribution of Photius to the schism between the Eastern and Western Churches that is summarized by J. RUINAUT in *Le Schisme de Photius* (1910); and by LOUIS BRÉHIER, "The Greek Church: Its Relations With the West up to 1054" and "Attempts at Reunion of the Greek and Latin Churches," vol. 4, pp. 246–273, 594–626, *The Cambridge Medieval History* (1923). Informative articles by F. DVORNIK, "Constantinople and Rome," and H. GREGOIRE, "The Amorians and Macedonians, 842–1025," appear in vol. 4, pt. 2, pp. 431–472, 105–192, *The Cambridge Medieval History*, new ed. (1966), with useful bibliographies. See CYRIL MANGO (ed. and trans.), *The Homilies of Photius* (1958). Other works of his are still most easily accessible in J.P. MIGNE (ed.), *Patrologiae Graeco-Latina*, vol. 101-104 (1860) with the *Myriobiblon* or *Bibliotheca* filling volume 103. Of the older books, J.B. BURY, *A History of the Eastern Roman Empire from the Fall of Irene to the Accession of Basil I., A.D. 802–867* (1912); and J.M. HUSSEY, *Church and Learning in the Byzantine Empire, 867–1185* (1937), especially the chapter on "Learning Under the Macedonians in the Ninth and Tenth Centuries," pp. 22–36, may still be used with profit.

(Ge.E.)

Photochemical Reactions

Photochemical reactions are chemical processes initiated by the absorption of energy in the form of visible, infrared, or ultraviolet radiation. The immediate consequence of this absorption is called the primary photochemical process. Subsequent changes, called secondary processes, are part of photochemistry. The process by which a photochemical reaction is carried on is photolysis.

Photochemistry differs from most other aspects of chemistry in one regard. If an atom or molecule absorbs energy from a beam of light, it gains far more energy that it ever could by other methods; *e.g.*, from ordinary heating. Consequently, photochemical processes are sometimes extremely efficient for the conversion of energy from light into chemical energy. The most important single photochemical process for living systems, photosynthesis, is of this type: green plants convert light energy into stored chemical energy by producing carbohydrates from carbon dioxide and water. Many other natural processes are wholly or partly photochemical. Ozone in the upper atmosphere, which shuts out most of the Sun's intense ultraviolet radiation, is produced by the action of light on oxygen molecules: sunlight breaks some of the molecules into atoms, which combine with other oxygen molecules to make ozone.

Bleaching laundry in sunlight is at least partly a photochemical process, and so is the darkening of white lithopone (zinc sulfide) paint. Photography is based on a photochemical process, the action of light on grains of silver chloride or silver bromide.

GENERAL CONSIDERATIONS

Kinds of photochemical effects. The action of light on a chemical system may take various forms:

1. Light may act as a booster to maintain a reaction that might otherwise proceed at an imperceptibly slow rate, the reacting molecules exchanging little or no net energy with their environment. Some molecular rearrangements (reactions that alter the geometry of individual molecules) fall into this category.

2. Light may act as a trigger, initiating a reaction that, once started, can proceed spontaneously. An example is the initiation of the reaction between hydrogen (formula, H_2) and chlorine (Cl_2) to give hydrochloric acid (HCl). Hydrogen and chlorine can be kept mixed together for an indefinite period at room temperature in a dark container; but if violet or ultraviolet light reaches the mixture, the two substances react explosively, giving off considerable stored chemical energy as heat. The primary photochemical process in this case is the formation of two chlorine atoms from a chlorine molecule, Cl_2.

3. In some photochemical processes, absorbed light converts the reacting molecules to a new state, higher in energy than their initial state; in this case the energy of the absorbed light is converted into chemical potential energy and is stored that way. Most photochemical processes occurring in biological systems are of this type; *e.g.*, ultraviolet irradiation converts ergosterol, found in plant and animal tissues, into a biologically active substance, vitamin D.

4. A photochemical system may store light energy from the Sun as chemical energy and then, only incidentally to the photochemical process, release that energy as electrical energy. This reaction occurs in the solar battery, used to supply the power in satellites and space vehicles.

All these processes have one common characteristic: each, at some stage, requires that a large quantity of energy be available to a single reacting atom or molecule. The requirement may be a thermodynamic one; that is, that the products simply contain more stored energy than the reactants. Alternatively, the requirements may be kinetic; in this case the products may store more or less energy than the reactants, but some step along the reaction path requires a large quantity of energy to surmount a barrier, as it were. (The hydrogen–chlorine reaction has as its barrier the requirement that a chlorine–chlorine bond be broken to convert a chlorine molecule into its two constituent atoms.) In more complex molecules, the rearrangement from one shape to another exhibits an energy barrier because, between the initial and final structures, there must be a series of intermediate shapes of higher potential energy than either the initial or final shapes.

The theoretical aspects of photochemistry are derived largely from the principles of atomic and molecular physics and of quantum mechanics, while its analytical methods are mainly those of spectroscopy. The reader therefore may profitably consult the articles ATOMIC STRUCTURE; LIGHT; MOLECULAR STRUCTURE; MECHANICS, QUANTUM; and SPECTROSCOPY, PRINCIPLES OF.

The development of photochemistry. The hydrogen–chlorine reaction cited above, perhaps the first recognized photochemical reaction, was first observed about 1801 and studied in detail some 40 years later. That only the absorbed light can affect a chemical reaction was recognized early, but this qualitative idea could not be stated as a quantitative law until the development of the modern theory by Einstein. Early photochemistry was largely confined to the identification of reaction products, often resulting from the action of sunlight.

Einstein's work in light absorption

The basic process by which light is absorbed by matter was first proposed by Einstein (1905, 1912, 1916). His work was directed particularly toward the photoelectric effect, in which the primary result of the absorption of radiation by a metal is the emission of free electrons; but the interpretation is general and provides the foundation of modern photochemistry: a beam of light is a beam of photons, or quanta (energy packets), of electromagnetic radiation (see below *Photochemical quantities*). The energy of each photon is proportional to the frequency of the light and the total number of photons is proportional to the intensity of the light. The primary photochemical process then follows directly from the absorption of one photon by one atom or molecule. (Exceptions may occur if the radiation field is very intense; then the concentration of photons may be so high that some molecules absorb two photons.)

After 1910 a German physicist, Emil Warburg, began a detailed study of the formation of ozone (O_3) by the action of ultraviolet light on the normal oxygen molecule (O_2). This work represented the first measurement of the number of product molecules formed for each quantum of energy absorbed: the quantum yield of the reaction. In the next few years most of the foundations of modern photochemistry were completed.

The role of chain reactions in photoprocesses was first elucidated for the hydrogen–chlorine reaction in the period 1913–25, and photochemistry grew steadily in the next three decades. From about 1955 the field expanded rapidly as it came to be used as a tool for the synthesis of organic molecules and was extended to the study of very short wavelengths (including X-rays and gamma rays) and of chemical phenomena produced by the optical amplifier, or laser, the acronym for light amplification by stimulated emission of radiation.

THE PHOTOCHEMICAL PROCESS

Photochemical quantities. The most fundamental ideas and the basis of all quantitative interpretations in photochemistry are the quantum-mechanical principles that all energy exists in packets, called quanta, and that the quanta of electromagnetic radiation, called photons, act like particles. The energy of each photon is directly proportional to the frequency of light—a relationship that is expressed by the equation $E = h\nu$. In this equation, E is the energy in ergs, and the frequency in cycles per second is symbolized by the Greek letter nu, ν. The constant of proportionality, h, is called, after its discoverer the German physicist Max Planck, Planck's constant; it has the value 6.625×10^{-27} erg-second. (One erg equals the work required to cause a mass of one gram to accelerate at one centimetre per second per second through one centimetre.) Alternatively, since the frequency, ν, and the wavelength, symbolized by the Greek letter lambda, λ, of light are related to c, the velocity of light (3×10^{10} centimetres per second), by the equation $\nu\lambda = c$, the energy of a quantum can be related to the wavelength of the corresponding radiation rather than to its frequency. This relationship is expressed by the equation $E = hc/\lambda$. When E is expressed in ergs and λ in angstrom units, the proportionality becomes $2 \times 10^{-8}/\lambda$; when E is in electron volts per photon, however, it becomes $12{,}400/\lambda$; and when E is in kilocalories (kcal) per mole of photons it is $2.85 \times 10^{5}/\lambda$. The angstrom unit (Å) is 10^{-8} centimetre and is commonly used to designate wavelength. The electron volt (eV), 1.602×10^{-12} erg, is the usual unit for measuring energies of individual particles in atomic physics.

By means of these quantities the relationship between ordinary chemistry and photochemistry can be expressed more precisely in terms of some simple magnitudes. Green light, for example, has a wavelength of about 5000 angstroms. (The entire visible region lies between about 4000 and 7000 angstroms.) A photon of green light, consequently, has an energy of about 2.5 electron volts, or an einstein (one mole of photons) of green-light photons has an energy of 57 kilocalories. Ultraviolet light of wavelength 2500 angstroms has just twice as much energy per photon or per mole. By comparison, a weak chemical bond requires an energy of about one electron volt to be broken, while a very strong bond requires more than 10 electron volts. In contrast, however, two molecules colliding at room temperature strike each other with an energy of only about 0.05 electron volt.

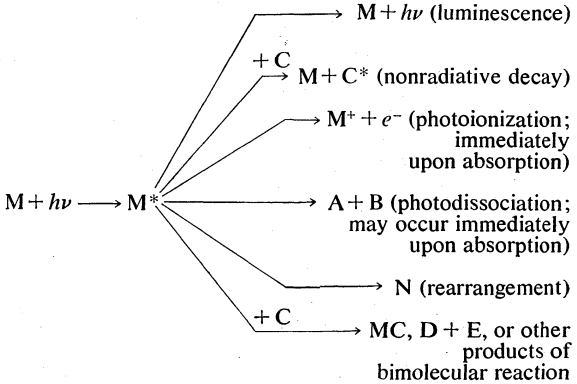
Absorption of electromagnetic energy. The history of a photochemical reaction necessarily begins with the absorption of a quantum of electromagnetic energy by an atom or molecule. Initially, the absorbing molecule or atom is in a state in which its energy is relatively low.

Types of energy in absorbers

The energy contained in the absorber is essentially of four types: (1) kinetic energy of motion through space (translational energy); (2) energy of rotational motion about its centre of mass; (3) energy of vibration of the component atoms against each other if the species is a molecule; and (4) energy of the negatively charged electrons moving in the field of force produced by the other electrons and the positively charged nuclei. The energy change initiating a photochemical process virtually always involves excitation of one electron from a state of low energy to a state of higher energy. When the absorber is a molecule, changes in rotational and vibrational energy usually also occur; the translational energy of an atom or molecule is almost completely unaffected when a quantum of radiation is absorbed, except in the case of gamma radiation.

Usually, in the initial low-energy state of an absorber, the electrons are in their lowest possible state of energy; that is, their lowest quantum state. The electrons cannot absorb all incident radiation, but only those quanta with energies corresponding to the allowed quantum jumps for that particular absorber. It has been determined that a hydrogen atom in its lowest state can absorb only those quanta having energies that fit the expression $13.7\ (n^2 - 1)/n^2$ eV, in which n may equal 2, 3, . . . ; according to this condition, it cannot absorb a six-electron-volt photon, for example. The energies that quanta must have in order to be absorbed by a molecule or atom correspond to the characteristic frequencies of the absorption spectrum of that substance.

Primary photochemical step. The occurrence of a photochemical reaction, then, begins with the absorption of a light quantum of suitable energy. The energy is used mainly or entirely to raise one electron into an excited state, with the result that the molecule or atom itself is in an activated state. The primary photochemical step then occurs; it may take the form of any of a number of different processes, as shown schematically below, in which M is the absorber; A and B indicate products formed by dissociation of M; C is a second substance, which interacts with M; N is a rearrangement product of M; D + E are secondary products that result from the interaction of M and C; e^- is an electron; and the asterisk designates electronically excited substances.



For each kind of molecule, more than one kind of excitation usually is possible. One of the origins of continuous absorption is photo-ionization, in which case the excited state of the electron is one in which it is free to escape from the atom or molecule, leaving behind a charged particle, or ion. It is the one case in which absorption and primary reaction are essentially identical. Photo-ionization can be thought of as a prototype—perhaps the most elementary photochemical process. Continuous absorption is characteristic of a transition in which a new free particle is produced: it may be an electron, as in photo-ionization, or it may be an atom or a molecular fragment. These latter two classes of particle are produced by photodissociation, a second type of photochemical process, which usually occurs when the electron being excited moves away from a locale where it can participate in a chemical bond into some other region where it is unable to use its attraction to pull two nuclei together. In normal photodissociation the excited electron remains bound to one or the other of the products of the photolysis. Photodissociation is sometimes almost as directly linked to the absorption process as photo-ionization is, but it can also be considerably slower.

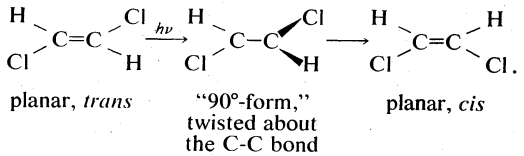
Photo-ionization and photodissociation

Absorption in a line or band spectrum leaves all the particles bound together with their geometrical configuration essentially unchanged, at least for a short time (10^{-14} second). The change produced by such an absorption involves the excitation of one electron from a closely held low-energy state to a more loosely bound excited state. In many cases there are two or more possible excited states and several electrons that could undergo a transition; because each spectral line or band corresponds to one specific electronic transition, however, it is frequently possible for an experimenter to study one kind of electron excitation at a time. This condition can be achieved by irradiating the sample with light lying only in the energy range corresponding to the transition that the experimenter is interested in. An experiment of this type can often be done with light of almost a single pure frequency (monochromatic light).

As soon as an electron is excited and the molecule is in an activated state, several events can occur. The molecule may return to its initial state, without reacting, in either of two ways: it may radiate, giving up a light quantum by luminescence; or it may lose its excitation energy, without radiating or reacting, by undergoing nonradiative decay, a process that requires effective collision partners to carry off the excitation energy. Nonradiative decay usually goes through two stages. In the first step, little or no energy is lost; the energy of electronic excitation is simply converted into vibrational energy (called intersystem crossing, or internal conversion). In the second stage, the excess energy is given up to the environment gradually through a succession of collisions.

Luminescence and nonradiative decay

If the excited molecule does react, it may do so in any of several ways. It may undergo photo-ionization or photodissociation, as indicated above. Or it may undergo an intramolecular reaction (an internal rearrangement) or an intermolecular process. An example of intramolecular rearrangement is the conversion of *trans*-1,2-dichloroethylene into its *cis* isomer (a molecule with identical numbers and types of atoms but different arrangement of them). In the equation for this reaction, the lines represent single covalent bonds and the shaded lines provide perspective; *i.e.*, their pointed ends are away from the viewer, their blunt ends toward the viewer:



The product, in this case, has almost the same energy as the starting compound; the excitation energy required to reach the twisted intermediate form must be given off somehow by the product, usually by collision.

Primary intermolecular reactions may involve nothing more than the transfer of a quantum of energy, as with mercury (Hg) and ethylene (C_2H_4):

$$Hg\,(vapour) + h\nu \longrightarrow Hg^*$$

$$Hg^* + C_2H_4 \longrightarrow Hg + C_2H_4^*,$$

leaving the excited ethylene to undergo secondary reaction. The intermediate kind of primary nondissociative reaction uses collisions in a way akin to the nonradiative decay process. An excited molecule undergoes a collision that converts it into a new reactive state without removing much of its excitation energy. The collision may change the geometry of the reactive molecule into a form that can rearrange bonds or dissociate readily, or it may change the electronic state of the molecule through internal conversion or intersystem crossing. Changes in electronic state usually involve only the single excited electron, but sometimes the molecular geometry is also involved.

In order for any state, be it the initial excited state or any subsequent one, to play a part in a photochemical reaction, it must endure long enough for the reaction to occur. Each photochemical reaction competes with radiative and nonradiative de-excitation to consume the activated molecules. A few typical numerical examples will illustrate the way this competition goes on. Fast luminescence, called fluorescence, occurs with a half life of about 10^{-8} second; *i.e.*, roughly half of the excited molecules in any sample will radiate their excitation energy away in about 10^{-8} second provided they have an allowed path for fluorescence and provided they are not destroyed faster by some other process. Simple dissociations of excited molecules occur much faster: an excited molecule in an unstable state dissociates in about the time it takes an atom to go through one vibrational period, or about 10^{-13} second. Consequently, a molecule excited to a truly dissociating state has virtually no chance to radiate and will indeed follow the path of dissociation. Other unimolecular processes may occur with characteristic times ranging from 10^{-13} second up to a few seconds, if no other process interferes.

Radiative processes can also be slow. Slow luminescence is called phosphorescence and is the result of emission of a quantum from a molecule undergoing a relatively improbable transition. Fluorescence is the result of an "al-

Fluorescence and phosphorescence

lowed" transition. Phosphorescence frequently is characterized by a lifetime for the excited molecule between 10^{-5} and 10^{-2} second, but many cases outside these bounds are known. Fluorescence, phosphorescence, dissociation, and some rearrangements involve only a single molecule and have specific rates (rate per unit of reactant concentration) or lifetimes for reactant molecules that are independent of the number of molecules present. Processes involving two molecules have specific rates that depend on the concentration of collision partners, because these rates depend on the frequency of collisions. The maximum rate for a bimolecular reaction is just the collision rate; this rate, about 10^9 per second in the normal atmosphere, is attained only if the reaction occurs with every collision.

In solution, the maximum bimolecular rate is governed by the rate at which the reactive species can diffuse together through the liquid or solid medium. Like the collision frequency, the maximum encounter rate represents an upper limit for the rate of reaction. Rates in solution are inversely proportional to the solvent viscosity, with the result that bimolecular processes generally proceed slower in glycerin, for example, than in water. Bimolecular de-excitation processes are usually limited by the rate at which the internal electronic state can change.

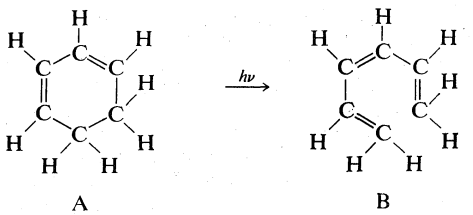
Any molecule that may decay or react in more than one way will follow the fastest of these ways. Comparison of the reaction rates shows that dissociation usually is fastest, if it can occur; fluorescence is fast in gases but is slow relative to collisions, nonradiative decay, and bimolecular reactions in solutions.

When a molecule changes its electronic state without direct absorption of light, as in nonradiative decay or internal conversion, it may go into a new excited state, which, left alone, would be short-lived (less than about 10^{-6} second); or it may go into a long-lived (metastable) excited state. Both types of reaction are known in photochemistry, but the long-lived states are particularly important in storing energy for chemical processes. The most important of the last named kind of transition is a singlet–triplet transition—singlets and triplets being electronic states that differ in having paired and unpaired electrons, respectively.

Triplet state

Triplets are the most commonly found phosphorescing states. They cannot radiate rapidly and are therefore among the most long-lived excited species. For this reason triplets are frequently the primary products of photochemical reactions. Excited singlets are sometimes found, but they more often decay, often to triplets. Triplets can be identified by their magnetic properties; by contrast with singlets, the individual electron magnets of a triplet do not cancel each other so that triplets exhibit weak magnetism (paramagnetism), which can be detected by suitable techniques.

An interesting case in which the entire photochemical process occurs in the primary reaction is that of valence tautomerization; *i.e.*, the inter-conversion of molecular structures that differ very little in the relative positions of their atoms but have different arrangements of chemical bonds, as in the photoisomerization of 1,3-cyclohexadiene shown in the structural formulas below:



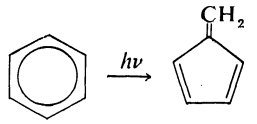
A B

Though structures A and B are nearly identical geometrically, they represent different states of the electrons.

Secondary steps. If an excited molecule does not return by a direct path to its lowest energy state, then it must dissociate, react with neighbours or itself, or revert to a new electronic state, still having some excitation energy. These are the primary steps (above). Any subsequent step is a secondary reaction. An example is the primary dissociation of an oxygen molecule (O_2) into two oxygen atoms; the only secondary reaction of importance in pure oxygen is combination of oxygen atoms with oxygen molecules to produce ozone (O_3). Both ozone and oxygen can then go on to attack hydrocarbons, in typical urban atmospheres.

Rearrangement reactions

Reactions in which atoms move from one site to another within a molecule—that is, rearrangement reactions—may be interpreted as secondary reactions following the photoexcitation process. Benzene irradiated with light between 2000 and 3000 angstroms rearranges (in low quantum yield) to the unstable hydrocarbon fulvene with the movement of a hydrogen atom, as shown in the structural formulas below:



benzene fulvene

in which the circle within the hexagon represents six electrons of benzene; the first step involves only the excitation of one of these six relatively free electrons, which do not participate in carbon–hydrogen bonds to any significant degree. Thus, only in a secondary step following photoexcitation can the atoms rearrange to the molecular structure of the product.

Frequently, the primary photochemical process is used as a tool to permit the secondary reactions to be studied. Photosensitized reactions often fall into this category. Mercury atoms can absorb light from a mercury lamp very efficiently and then, in their excited state, collide with other molecules and transfer their excitation to them. This technique has been used to dissociate nitrous oxide (N_2O) and provide free oxygen atoms. The reaction sequence proceeds as shown below:

$$Hg + h\nu \longrightarrow Hg^*$$

$$Hg^* + N_2O \longrightarrow Hg + N_2 + O.$$

(The alternative mode of decomposition, to give NO + N, has not been found to occur.) Then the oxygen atoms may react with whatever other substances are present. Measuring the amount of nitrogen produced in such a photolysis gives directly the number of oxygen atoms that were available for reaction.

Chain reactions

If a primary step involves dissociation of a molecule into two radicals (molecular fragments with single or unpaired electrons), the secondary steps are often chain reactions. About 1923 it was recognized that there were certain cyclic processes, called chain reactions, in which reaction of an active intermediate with a molecule of starting material to give a product necessarily also produced a new active intermediate. The hydrogen–chlorine reaction is an example. Light dissociates a few chlorine molecules into atoms, which then attack hydrogen molecules to give hydrogen chloride product and hydrogen atoms. The hydrogen atoms react with chlorine molecules to produce more hydrogen chloride and chlorine atoms. The reactions are:

$$h\nu + Cl_2 \longrightarrow 2Cl$$

$$Cl + H_2 \longrightarrow HCl + H$$

$$H + Cl_2 \longrightarrow HCl + Cl.$$

Reactions of this kind require an initiator (light, in this case) and a propagating mechanism and may involve a separate terminating step. Termination actually occurs either when the concentration of at least one reactant becomes very low or when another substance (*e.g.*, oxygen or iodine) is present to react more rapidly with the radicals than can the hydrogen or chlorine molecules. The reaction must also be exothermic (heat-releasing) or nearly thermoneutral (that is, neither releasing nor absorbing heat), otherwise it may cool the system so that it no longer reacts. In the reaction just cited, considerable heat is generated—usually much faster than it is radiated

or transferred to the vessel walls—and an explosion is the result. An even more unstable situation—violently explosive—arises if one radical initiator produces more than one new radical.

Chain reactions always produce many product molecules for every quantum of energy absorbed; *i.e.*, the quantum yield may be much greater than one for the reaction. Reactions involving no chain processes necessarily have quantum yields near or below one, while a chain reaction may have a quantum yield of 80,000. Even a rough measurement of quantum yield will generally show whether a reaction proceeds through a chain mechanism.

EXPERIMENTAL METHODS

The methods of studying photochemical reactions have been developed chiefly to deal with problems of three kinds, which are: (1) getting the desired light energy into the system being studied; (2) determining the progress of the reaction—getting immediate pertinent information from the experiment; and (3) determining, from the results of individual experiments, the broader inferences for which the study is carried out. Photochemists commonly try to determine from a series of individual reactions the mechanism of a reaction—*i.e.*, the behaviour of each atom throughout the progress of the reaction.

Illuminating the sample. Providing suitable illumination for photochemical reactions, the experimenter has four controls: (1) The frequency, or energy, of the light can be controlled. (2) The intensity of the light can be varied to control the rate at which photons arrive at the sample and thereby to control the number of excited molecules present at one time. (3) The time of illumination can be varied to control the total number of excited molecules produced. (4) The polarization of the light may be selected.

In the earliest photochemical observations the Sun was the source. No control could be exerted, of course, on the frequency of light coming from the source itself, but simple controls of the light reaching the sample were available through the use of sample containers of differing composition; *e.g.*, flint glass transmits very little ultraviolet radiation (of wavelengths less than about 3600 angstroms), borosilicate (Pyrex) glass transmits light of wavelengths above about 3000 angstroms, and some fused silica is transparent to wavelengths shorter than 1700 angstroms. The next simplest way to control the wavelength reaching the sample is to interpose a suitable filter in the light path. Filters may transmit broad bands of frequencies or narrow and nearly monochromatic bands. A more complex step is the use of the monochromator, a device (usually a prism or grating) that disperses light and permits the continuous variation of its frequency.

Incandescent sources, such as the Sun and ordinary electric tungsten-filament lamps, emit light of all frequencies, the total intensity and the distribution of intensity depending primarily on the source temperature. For higher intensities it is often possible to use the emission spectrum of an atomic or molecular vapour. Some common sources use mercury, cadmium, sodium, xenon, hydrogen, or deuterium in an electric-arc discharge to provide intense and steady fluxes of radiant energy. If the lamp pressures are low, the light emitted by mercury, cadmium, or xenon is restricted to a few, intense, well-separated wavelengths, from which it is easy to select one.

Conventional steady arcs, discharge tubes, and incandescent sources are inadequate for very intense illumination. High flux of quanta requires a flash tube, a pulsed X-ray or gamma source, a laser, or the radiation from an electron accelerator such as a synchrotron. The most intense monochromatic source is the pulsed laser, which can deliver as much as 10^9 watts of power for 10^{-11} second at a single wavelength.

Monitoring the reaction. The course of a photochemical reaction may be observed by many physical and chemical methods. Often it is sufficient to know what the final reaction products are, which can be determined by established methods of qualitative identification.

In other cases, the information sought from a photochemical reaction can be derived only by examining the intermediates—the species present during the course of the reaction but not at the end. Sometimes an indirect observation identifies an intermediate; methyl (CH_3), ethyl (C_2H_5), and other radicals react with metal surfaces and can be detected by their ability to remove a metal mirror from the walls of a vessel. More direct methods of studying the intermediates are available in many cases.

Use of flash-kinetic spectroscopy

Transient intermediates can be found and identified in some photochemical reactions proceeding at their natural rapid rates. One method of doing this is flash-kinetic spectroscopy: a rapid photolysis is effected by an intense flash, which is followed by a second flash from a small lamp acting as a light source for absorption spectroscopy (analysis based on absorption of radiation according to the energy content of the sample's molecules and the perturbing effects of surrounding molecules). With ultrafast laser pulses, reactions can be studied at time intervals as short as 10^{-12} second after they are initiated. The method has been particularly successful for identification of spectra of such small transient species as the amino (NH_2), aldehyde (CHO), cyanate (NCO), and methyl (CH_3) radicals and of such more complex species as the phenyl radical (C_6H_5). Gases, liquids, and solids can be examined by flash-kinetic spectroscopy to determine the time dependence in the absorption spectrum. This can be carried out in a series of photolyses, taking spectra at progressively later times after photolysis and accumulating the data, or one can move a film or an image during the event to obtain a streak photograph of the reacting system.

As a complement to the optical absorption spectrum of a rapidly reacting mixture, one can study the time dependence of the mass spectrum (the array of charged particles sorted according to their masses). The so-called time-of-flight mass spectrometer is capable of scanning the mass spectrum of a gas being sampled as often as once every 10^{-5} second. A particular mass that appears only during the middle stages of a reaction is associated with an intermediate.

Quantum yield

Possibly the most fundamental quantitative measurement of a photochemical process is that of quantum yield, which may be determined from a knowledge of the absolute emission characteristics of the exciting source and of the absolute absorbing ability of the absorbing species; or the measurement may be made as a relative one, by using as a comparison and standard a reaction the absolute quantum yield of which is precisely known. Absolute measurements are usually much more difficult than relative ones, and so the relative method is more common; but the standards must come from absolute measurements.

The standard reaction that is most reliable for solutions is the photolytic decomposition of highly purified uranyl oxalate. The quantum yield of this reaction is well established as a function of wavelength, temperature, intensity, and concentration of various ions, particularly hydrogen ions. For gas-phase reactions, other standards are available; *e.g.*, the photolysis of hydrogen iodide, the quantum yield of which is precisely two for a wide range of conditions.

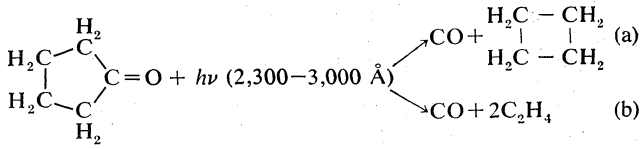
Determining the mechanisms. Determining the actual path of a photochemical reaction frequently requires setting up hypothetical models corresponding to different possible mechanisms; finding from its kinetic equations how each model system depends on concentrations, light intensity, and any other pertinent variables; and then deciding, from experimental variations, which model best describes the system. Using this method, investigators have shown how simple alkyl iodides—*e.g.*, methyl iodide (CH_3I)—decompose on photolysis and how the fragments must react. The methyl iodide system is interesting in that at least part of the energy of the absorbed photon stays in the reactive species long enough to assist in one crucial step. The steps of the mechanism and the evidence used to determine the role of each step are as shown in the Table.

Labelling. Other than direct detection of intermediates, the most powerful tool for elucidating mechanisms

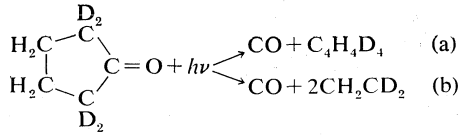
Steps in the Photolysis of Methyl Iodide

step	equation	methods of studying individual step
Primary step	$CH_3I \xrightarrow{h\nu} CH_3 + I$	brought about by near-ultraviolet radiation; quantum yield approximately 1, as shown by separate experiments using scavengers for iodine atoms and methyl radicals; presence of radicals shown by kinetic spectroscopy
Normal secondary steps	$2CH_3 + M \rightarrow C_2H_6 + M$ $2I + M \rightarrow I_2 + M$	M is an inert collision partner that carries away energy; ethane (C_2H_6) and iodine (I_2) molecules demonstrated by analysis of products
Inhibiting secondary steps (recombinations)	$CH_3 + I \rightarrow CH_3 + I$	rate determined from low-pressure, low-intensity photolysis to minimize three-body collisions
	$CH_3 + I + M \rightarrow CH_3I + M$	rate determined from high-intensity photolysis to maximize frequency of radical–radical collisions
Other secondary steps	$CH_3 + CH_3I \rightarrow CH_4 + CH_2I$	inferred from presence of methane (CH_4) in products; thermochemistry requires either reactant to have 9 kcal per mole excess energy
	$CH_2I + I + M \rightarrow CH_2I_2 + M$	inferred from presence of methylene iodide (CH_2I_2) in products
	$2CH_2I \rightarrow ICH_2CH_2I$	inferred from presence of 1,2-di-iodoethane (ICH_2CH_2I) in products

is probably the use of isotopic labelling. For example, deuterium (the hydrogen isotope that has twice the mass of ordinary hydrogen) can be used in place of hydrogen, or one of the other carbon isotopes in place of carbon-12, at selected sites in the starting material; the tagged atoms then can be located at particular positions in the product molecules. An example of the information isotopic substitution can give in a favourable case is in the photolytic decarbonylation (loss of carbon monoxide) of cyclopentanone. The process has alternative paths (a) and (b):



The use of a specific tetradeuterated species (*i.e.*, containing four deuterium atoms) shows that the paths are mutually exclusive. No ethylene of the formula C_2D_4 is formed, and therefore the ethylene cannot be a product of an intermediate in which all four carbon atoms are equivalent. The reaction is this:



Quenching. A good method, and one of the oldest in photochemistry, for determining both kinetic constants and mechanisms is the study of fluorescence quenching. If a large molecule (*e.g.*, anthracene, pyrene, or acridine) is photoexcited and does not react, in general it will fluoresce. If a species is added that reacts rapidly with the photoexcited substance, fluorescence will no longer occur; *i.e.*, it will be quenched. This method has made possible the determination of the strengths of acids in excited electronic states; of the rate of combination of two anthracene molecules (dimerization; dimer: double molecule formed by the union of two identical simpler molecules—*e.g.*, C_4H_8 is a dimer of C_2H_4); and of rates of long-range transfer of energy.

APPLICATIONS

In synthetic chemistry. Because absorption of a light quantum is frequently a property of a single group of atoms within a molecule, photochemistry sometimes provides the best means for producing a chemical reaction at a specific molecular site. This application becomes especially valuable when thermal excitation would be unselective or would destroy the reactant. Photochemical syntheses include elimination reactions and the formation of isomers and dimers.

Photolytic syntheses are sometimes most effective if the reactants are held rigidly by a solid lattice, so that a reactant molecule absorbing a quantum is fixed in a favourable orientation for reaction. Solid-state photolyses of this kind have been used in the synthesis of the dimers of maleic and fumaric acids.

In radiation chemistry. The photochemistry associated with high-energy radiation (X-rays and gamma rays) is closer to the chemistry of electron-bombarded materials than to the photochemistry of visible-ultraviolet radiation. High-energy electron or photon bombardment produces, in addition to excited neutral molecules, ionized species, which are products of very fast secondary processes, not of the primary quantum absorption process. The basic processes associated with high-energy radiation are simply too fast to lead to any chemistry in the traditional sense of bond formation or breakage. Radiation chemistry is largely concerned with determining what the first-formed new chemical species are and what reactions they undergo: it assumes a particularly important role in the understanding of how living beings react to radiation and how they can be protected against its damaging effects (see RADIATION EFFECTS ON MATTER).

Products of secondary processes

In atmospheric photochemistry. The phenomena of terrestrial and planetary atmospheres—*e.g.*, atmospheric structure, auroras, radiation shielding, and weather—depend on the chemical constituents of the atmosphere and the radiation striking it. The Earth's atmosphere is largely molecular nitrogen and molecular oxygen at the Earth's surface, but it contains quite a different set of atomic and molecular components at high altitudes, largely because of the radiation that these upper layers absorb. The photochemical processes that take place are covered in the articles ATMOSPHERE and IONOSPHERE. The very rich chemistry of polluted atmospheres is a comprehensive model that exhibits many aspects of photochemistry. Brown nitrogen dioxide absorbs sunlight and uses this energy to initiate chain reactions of hydrocarbons with oxygen. The nitrogen oxides return to pick up more sunlight, while the partially oxidized hydrocarbon molecules undergo secondary reactions and may even absorb light to start new photochemical processes.

BIBLIOGRAPHY. JACK G. CALVERT and J.N. PITTS, JR., *Photochemistry* (1966), probably the most widely used modern monograph on the subject, covering basic generalities and specific applications; W. ALBERT NOYES, JR., G.S. HAMMOND, and J.N. PITTS, JR. (eds.), *Advances in Photochemistry*, vol. 1–6 (1963–68), a serial publication, complementary to the book by Calvert and Pitts, surveying recent work in specific areas; W. ALBERT NOYES, JR., PHILIP A. LEIGHTON, *The Photochemistry of Gases* (1941, reprinted 1966), a classic, probably an easier introduction to the subject; R.M. NOYES and A. WELLER, "Photostationary Methods," and G. PORTER, "Flash Photolysis," in S.L. FRIESS, E.S. LEWIS, and A. WEISSBERGER (eds.), *Technique of Organic Chemistry*, vol. 8, pt. 2 (1963), two articles that present clear discussions of most of the experimental methods now in use, except lasers; C.B. MOORE, "Lasers in Chemistry," *A. Rev. Phys. Chem.*, 22:387–428 (1971), a brief and descriptive article on laser methods in photochemistry; V. HENRI, *Études de photochimie* (1919), important for the history of the subject.

(R.S.Be.)

Photoelectric Effect

The photoelectric effect is a phenomenon in which electric charges are liberated from a material when it absorbs radiant energy. Usually, the photoelectric effect is thought of as the ejection of electrons from the surface of a metal plate when light falls on it. In the broad sense, however, the phenomenon can take place when the radiant energy

is in the region of visible or ultraviolet light, X-rays, or gamma rays; when the material is a solid, liquid, or gas; and when the charges liberated are electrons or ions (charged atoms or molecules).

Investigations around the beginning of the 20th century on photoelectric phenomena led theoretical physicist Albert Einstein to propose the theory that light can be wavelike and yet behave like streams of discrete particles, or quanta of energy. From this beginning grew the quantum mechanics of modern physics, without which there would be no understanding of phenomena occurring on the molecular, atomic, or nuclear scale. The photoelectric effect is the basis of operation of such devices as the television camera, photographic light meters, image intensifiers for telescopes, cells for the conversion of solar energy, and detectors for all kinds of spectrometry.

Discovery and early work. The first recorded observation of the photoelectric effect, or photoemission, occurred in 1839 when it was reported that an electric current generated by a cell composed of two metallic electrodes immersed in a conducting solution, or electrolyte, will increase when the cell is illuminated and that the extent of the increase will depend on the wavelength of the illuminating light. In another experiment, an investigator a few decades later noticed that there is a change of resistance of selenium with illumination. The photoelectric effect was actually and explicitly discovered in 1887 by a German physicist, Heinrich Rudolf Hertz, who observed that ultraviolet light changes the lowest voltage at which sparking takes place between given metallic electrodes. At the close of the 19th century, it was established that a cathode ray (produced by an electric discharge in a rarefied-gas atmosphere) consists of discrete particles, called electrons, each bearing an elementary negative charge (see ELECTRON). In 1900 Philipp Edward Anton Lenard, a German physicist, studying the electrical charges liberated from a metal surface when it was illuminated, concluded that these charges were identical to the electrons observed in cathode rays. It was further discovered that the current (given the name photoelectric because it was caused by light rays), made up of electrons released from the metal, is proportional to the intensity of the light causing it for any fixed wavelength of light that is used. In 1902 it was proved that the maximum kinetic energy of an electron in the photoelectric effect is independent of the intensity of the light ray and depends on its frequency.

The observations that (1) the number of electrons released in the photoelectric effect is proportional to the intensity of the light and that (2) the frequency, or wavelength, of light determines the maximum kinetic energy of the electrons indicated a kind of interaction between light and matter that could not be explained in terms of classical physics. The search for an explanation led in 1905 to Einstein's fundamental theory that light, long thought to be wavelike, can be regarded alternatively as composed of discrete particles (now called photons), equivalent to energy quanta.

Einstein's fundamental theory and equation

In explaining the photoelectric effect, Einstein assumed that a photon could penetrate matter, where it would collide with an atom. Since all atoms have electrons, an electron would be ejected from the atom by the energy of the photon, with great velocity. The kinetic energy of the electron, as it moved through the atoms of the matter, would be diminished at each encounter. Should it reach the surface of the material, the kinetic energy of the electron would be further reduced as the electron overcame and escaped the attraction of the surface atoms. This loss in kinetic energy is called the work function, symbolized by omega (ω).

According to Einstein, each light quantum consists of an amount of energy equal to the product of Planck's universal constant (h) and the frequency of the light (indicated by the Greek nu, ν). Einstein's theory of the photoelectric effect postulates that the maximum kinetic energy of the electrons ejected from a material is equal to the frequency of the incident light times Planck's constant, less the work function. The resulting photoelectric equation of Einstein can be expressed by $E_k = h\nu - \omega$, in which E_k is the maximum kinetic energy of the ejected electron, h is a constant, later shown (see below) to be numerically the same as Planck's constant, ν is the frequency of the incident light, and ω is the work function.

The kinetic energy of an emitted electron can be measured by placing it in an electric field and measuring the potential or voltage difference (indicated as V) required to reduce its velocity to zero. This energy is equal to the product of the potential difference and an electron's charge, which is always a constant and is indicated by e; thus, $E_k = eV$.

Experimental confirmation of Einstein's equation for the photoelectric effect was provided in 1916 by the investigations of a U.S. physicist, Robert A. Millikan. Proof that Einstein's numerical constant was the same as Planck's constant gave an enormous boost to the then new developing quantum theory and gave plausibility to the Danish physicist Niels Bohr's concept of the structure of the atom in which electrons revolve in prescribed orbits about a nucleus. Einstein's theory thus stimulated a much broader area of physics than the limited field of photoemissive phenomena.

Two more milestones in the history of the photoelectric effect must be mentioned. In 1923 Arthur Holly Compton, a U.S. physicist, discovered an effect, given his name, that relates the directions taken by a photon and an electron, after they meet, with the energy of the collision (see below *Compton effect*). And in 1925, another effect, sometimes called the compound photoelectric effect, or the Auger effect, was discovered by a French physicist, Pierre-Victor Auger (see below *Auger effect*).

Photoemission. *The phenomenon.* Whenever the absorption of a photon by an atom results in the ejection of a free electron, one speaks of photoemissive phenomena. Einstein's equation predicts that the emission current composed of these electrons will be directly proportional to the intensity of the number of photons absorbed in the emitter. The equation is consistent with a three-step model of photoelectric emission from a solid. The first step consists in the absorption of a photon by an atom and the release of an electron, an event that may occur at the surface or at any distance below the surface of the solid on which the light waves fall; the penetration depends on the degree of absorption (called the absorption coefficient) of the material for the specific frequency of the incident light. If the electron is ejected below the surface, the second step consists in the movement of the electron through the layers of atoms between it and the surface. In this diffusion process the electron may either find no resistance or be hindered by collisions with atoms. In the latter case, the electron may never reach the surface, or if it does it may do so only after losing some of the kinetic energy it acquired from the photon. The third step is the escape of the electron from the surface of the solid, before which the electron must overcome an electrostatic barrier set up by the surface charges. It can do this only if its kinetic energy is high enough.

Temperature dependence. The validity of the Einstein relationship was examined by many investigators and found to be correct but not complete. In particular, it failed to account for the fact that the emitted electron's energy is influenced by the temperature of the solid. The remedy to this defect was first formulated in 1931 by a British mathematician, Ralph Howard Fowler, who, on the assumption that all electrons with energies greater than the work function would escape, established a relationship between the photoelectric current and the temperature in which the current is proportional to the product of the square of the temperature and a function of the incident photon's energy. The equation is $I = \alpha A T^2 \phi(x)$, in which I is the photoelectric current, α and A are constants, and $\phi(x)$ is an exponential series, whose numerical values have been tabulated; the dimensionless value x equals the kinetic energy of the emitted electrons divided by the product of the temperature and the Boltzmann constant of the kinetic theory: $x = (h\nu - \omega)/kT$; in which x is the argument of the exponential series, h is Planck's constant, ν is the frequency of the incident radiation, ω is the work function of the material, k is Boltzmann's constant, and T is temperature.

Fowler's equation

A great advantage of the Fowler equation is its easy comparison with experiment. By plotting the experimental results of temperature, photoelectric current, and the function argument x, the threshold frequency is easily determined. In addition to the convenience of a simple determination of the work function, the equation predicts the correct shape of curves plotted for emission current versus radiation frequency, at least in the vicinity of the threshold value.

An example of the comparison of experimental data with the Fowler theory is given in Figure 1.

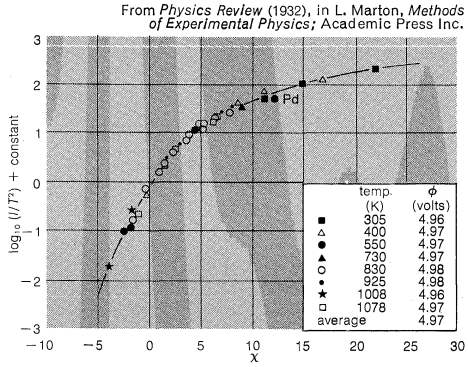


Figure 1: Spectral distribution curve of palladium, based on Fowler's theory.

Originally, most investigations of photoemission centred on the emissive properties of solid metals that are good conductors of electricity. Gradually, with the recognition of the effects of volume and surface impurities, there evolved an interest in the photoelectric emission from insulators, which are nonconductors, and from semiconductors, which conduct electricity only under certain circumstances. The energy distribution of electrons in insulators and semiconductors is different from that of metals. As a consequence, one finds that the photoelectric yield (number of electrons emitted for each incident quantum of radiation) versus energy of the incident photon can be different for materials having the same work function.

Semiconductors, metals, and insulators. The photoelectric effect in semiconductors can be understood only in terms of quantum mechanics. It is possible to think of solids as being composed of tiny crystals, each of which has an orderly array of atoms in three dimensions called a lattice structure. Every electron in the lattice, whether bound to an atom or unbound, has a definite energy, called its energy state. Energy states are influenced by the lattice structure of the crystal. Electrons can change to a higher energy state by absorbing heat or electromagnetic energy. According to the laws of quantum mechanics, only specific discrete energy states may exist, and these are grouped together in energy bands, with the regions between them called forbidden bands. The band of highest energy that is occupied when the crystal is at absolute zero temperature is the valence band. If this band contains states that are unoccupied, it is called a conduction band and the crystal is a metal. If all the states are occupied, the crystal is either an insulator or a semiconductor, depending on whether or not electrons from the valence band can be transferred to the next unoccupied band, which then becomes a conduction band. In a semiconductor the input of energy, in the form of light, can raise a valence electron into the conduction band. Electrons can also be supplied to a semiconductor by atoms having energy levels in the forbidden gap, called impurity atoms. In an insulator the forbidden gap is too large to be bridged easily.

Work function of semiconductors

The work function depends on many factors. It is affected radically by the purity and surface condition of the specimen. Even for a perfectly pure surface it is found to be strongly affected by the crystallographic orientation of the surface. The work function depends on the packing arrangement of the atoms and thus the distances between atoms in the crystal; the most closely packed surfaces have the highest work functions and the less closely packed ones (for the same substance) the lower ones. A polycrystalline surface usually represents the average work function of the most commonly occurring crystallographic orientations.

Another important distinction between pure metals and many semiconductors lies in the light absorption properties of the two. The behaviour of compound photoemitters (*i.e.*, emitters built up of layers of different materials that may or may not have diffused into each other) may be influenced by their absorptivity; *i.e.*, the fraction of light energy absorbed. The simplest case of a compound photoemitter is a metal with an absorbed layer of some chemical impurity or a metal whose surface has undergone a chemical modification. Many metals when oxidized manifest a reduced work function with respect to the pure metal by as much as 1.5 electron volts. Because the work functions of most pure metals are too high for convenient technological applications in visible light, the search for low work function surfaces has been directed toward compound photoemitters. The most important of these emitters are, in chronological order of discovery: the silver–oxygen–cesium cathode, the cesium–antimony cathode, the bismuth–silver–oxygen–cesium cathode, and the different multialkali antimonide cathodes. All modern photoemitter materials combine low work functions with high photoelectric yield.

Compound photoemitters

The time lag between the absorption of a photon and the emission of a photoelectron is short: it appears to be less then 10^{-10} second. Although in this section the discussion has been limited to photoemission from solids, it should be pointed out that much work has been done on photoionization in gases.

Photoconductivity. Many materials exhibit a marked change in electrical conductivity when irradiated. Usually the conductivity increases, although a few instances of the opposite reaction are known. In highly conductive substances, such as metals, the change in conductivity is insignificant. It is most important in semiconductors.

In a semiconductor there are a certain number of "free" electrons occupying energy levels in the conduction band and a certain number of positively charged "holes" left in the valence band by the free electrons that entered the conduction band because of the natural vibrations of atoms. During irradiation of the substance, the generation of free charge carriers (electrons and holes) increases, producing an increased conductivity, continuing until the carriers generated by the photon absorption can no longer move through the material. The fundamental electronic processes can be manifold, depending upon the band structure of the material, the imperfections in the material, and other factors. One of the most important factors in determining photosensitivity (*i.e.*, the change in conductivity per photon absorbed) is the free-carrier lifetime, which must be finite because recombination or neutralization takes place between negatively charged electrons and positively charged holes and there are thus fewer electrons available for the conduction process. This process can vary from 10^{-2} second to 10^{-12} second, not only from substance to substance but also in a single substance having varying degrees of imperfections or impurities. For a thorough study of photoconductivity it is thus important to employ materials that are well defined structurally, preferably single crystals. The lifetimes of the charge carriers determine the response time of the photoconductor to changes in the light stimulation; the observed decay time of the photoconductivity can be related to the lifetime of the free carriers by an equation in which the decay time, indicated by tau, subscript zero (τ_0), equals the lifetime, indicated by tau (τ), multiplied by the ratio of the density of trapped carriers, indicated by n, subscript t (n_t), and density of free carriers, indicated by n: $\tau_0 = \tau\ (n_t/n)$. The lifetime depends on the nature of the recombination process of the charge carriers, and many different mechanisms are possible.

Photosensitivity

Another important factor determining photosensitivity is the light absorption characteristic of the photoconductors. When the photon energy is greater than the forbidden energy gap of the solid, each photon creates a free

pair—a free electron and a free hole. If the photon energy is much larger than the band gap (as, for instance, the photon energy of X-rays), the absorption of a single photon can give rise to many free pairs. The optical absorption cross sections are incompletely known.

In practice, the ideal photoconductor with a fast and a linear, or directly proportional, response to the stimulation is almost nonexistent. Photoconductors show a number of secondary effects, a few of which may be mentioned here. Sensitive photoconductors may have response times measured in hours or days when operated at extremely high resistivities. The photocurrent may be reduced, or quenched, by the addition of further radiation, the wavelength of which is long compared with the radiation inducing photoconductivity. Similar to this infrared quenching is thermal quenching. Quenching may be explained in terms of raising an electron from the valence band to a sensitizing centre. This change can be achieved either by an infrared photon or by thermal excitation. As the free holes excited by light are caught in centres whose capture cross section for electrons is very small, moving these holes to other centres whose capture cross section for electrons is large can reduce markedly the lifetime of the free electrons. Negative photoconductivity may be due to unobserved quenching effects.

Photovoltaic effect. The photovoltaic effect consists in the generation of an electromotive force as a consequence of the absorption of radiation; that is to say, a current will flow across the junction of two dissimilar materials when light falls upon it. The primary effect is photo-ionization; *i.e.*, the production of equal numbers of positive and negative charges. One or both charges can then migrate to a region in which charge separation can occur. This charge separation happens normally at a potential barrier between two layers of the solid material.

For a large photovoltaic effect, the change of potential at the barrier should be large and the distance over which the carrier has to move should be small. This effect implies an abrupt change in some property of the material, including an abrupt change of the conductivity. The response time of the photovoltaic effect is relatively short, although not as instantaneous as photoemission.

Energy conversion device

An important characteristic of the photovoltaic effect is its relatively high efficiency as a device to convert the energy of radiation to electrical current. Extensive literature exists on conversion efficiencies for sunlight, showing experimental efficiencies reaching 14 percent in contrast to the theoretically expected efficiencies of 24 percent.

Other photoelectric effects. *Auger effect.* Photoemissive phenomena are well-known for the higher energy ranges in the X-ray and gamma regions because such penetrating radiation can release photoelectrons from inner atomic shells. If a vacancy is produced in an inner shell because of the removal of an electron through the photoelectric process, the vacancy will be filled by an electron from a less tightly bound state; the resulting excess energy may cause the simultaneous emission from the atom of a second electron, called an Auger electron, from another less tightly bound state. This is the Auger effect, which is generally, but not exclusively, observed at higher energies. A given inner vacancy can be followed by a great number of different Auger transitions. This fact, coupled with the development of high resolution spectroscopy and improved methods for the study of low-energy electrons, makes the Auger effect an important tool in the exploration of many nuclear processes such as capture and internal conversion. Also, the knowledge of the energies and intensities of Auger electrons gives information about transition probabilities, both in atoms and in molecules. Conversely, Auger spectroscopic lines with precisely known energies can be used for the energy calibration of conversion electrons—electrons that are ejected from their shells when a nucleus loses its excitation energy.

Use in energy calibration

Compton effect. In the collision between a higher energy photon and an electron initially at rest, the photon will share some of its energy and momentum with the electron, and the photon and electron will take off in divergent paths. The laws for the conservation of energy and momentum show that kinetic energy of the electron is strongly dependent on the energy of the photon, whereas the Compton shift in wavelength of the photon is independent of the energy. The Compton shift of wavelength between the incident and scattered photon, which represents an increase in wavelength before and after collision, is equal to the product of two terms: the first term is the ratio Planck's constant to the rest mass of an electron (according to relativity, a body has the least mass when at rest relative to the observer) times the velocity of light; the second term is the difference between 1 and the cosine of the angle that the scattered photon makes with the path of the incident photon. In equation form, $\Delta\lambda = (h/m_0c)(1 - \cos\theta)$, in which $\Delta\lambda$ represents the shift in wavelength, h is Planck's constant, m_0 is the rest mass of the electron, c is the velocity of light, and θ is the scattering angle of the photon. The term h/m_0c that corresponds to a 90° scattering angle of the photon is called the Compton wavelength for all electrons. It is a fundamental atomic unit of length having a numerical value of 2.4262×10^{-10} centimetre.

Atoms, electrons, and nuclei can interact with electromagnetic radiation in many possible ways. Whereas at low energies the photoelectric effect dominates, and at very high energies pair production (the conversion of a photon into an electron and a positron) is prevalent, there exists a broad range of energies in which the attenuation of radiation depends almost exclusively on Compton scattering (Figure 2).

From S. Flugge (ed.), *Encyclopedia of Physics*, in L. Marton, *Methods of Experimental Physics;* Academic Press Inc.

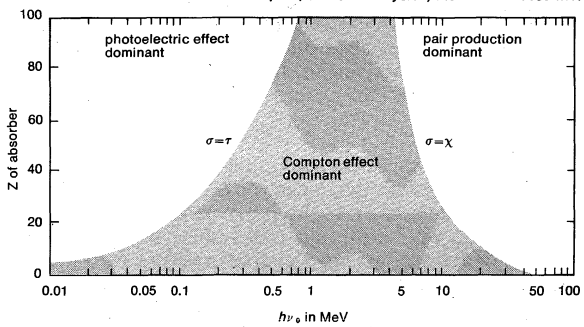


Figure 2: Locus of equal atomic cross sections for Compton and photoelectric interactions ($\sigma = \tau$) and for Compton and pair-production interactions ($\sigma = \chi$). The incident photon energy is $h\nu_0$, and Z is the atomic number of the atoms in the absorber.

Photonuclear effects. If the energy of the photon is sufficiently high, its wavelength becomes comparable with nuclear dimensions (the diameter of a nucleus) and a variety of nuclear reactions take place. The process most akin to the atomic photoelectric effect is the direct photoemission of a neutron or proton from the nucleus. More complicated interactions involve either the emission of heavier particles (alpha particles, tritium, deuterium), many particles, or photofission. Much valuable information about the nuclear photoeffect has been obtained from studies of the inverse reaction; *i.e.*, those in which the bombarding particle is captured by a nucleus and a photon is emitted. Experimental studies of elastic and inelastic photon scattering have resulted in important data, as have studies of elastic and inelastic electron scattering, with and without the detection of coincident outgoing reaction products. The cross section for these reactions shows a broad spectrometer peak in the region of 13,000,000 to 25,000,000 electron volts of photon energy, which has been called the giant dipole resonance. The width of this peak varies between 3,000,000 and 10,000,000 electron volts and is smallest for closed-shell nuclei (according to the shell model of the atomic nucleus, neutrons and protons move in concentric shells; the exclusion principle postulates that a shell is closed when it contains its quota of particles, and a closed-shell nucleus is one having its outermost shell completely occupied) and is greatest for nonspherical nuclei. At high energies (greater than 140,000,000 electron volts) photoemission processes become important.

Applications. The many technological variants of photoemissive devices take advantage of a response that is linear with light intensity, coupled with an extremely short response time. A photoemissive device (photodiode) can be coupled with a suitable amplifying system, its sensitivity thus extended to the detection of single photons. The best known amplifying system, the photomultiplier, consists of a suitable set of secondary emitters built into the same container. The disadvantages of this system are its limited spectral response: most photocells or photomultipliers are limited to the visible and ultraviolet ranges of electromagnetic radiation. Even at their peak performances, these devices give about 10 percent quantum efficiency, and these peak performances are limited to a narrow part of the spectrum. Much effort has been spent on the properties and preparation of suitable photoemissive material adapted to different applications.

The photomultiplier

Photodiodes and photomultipliers are being applied in a growing number of fields such as the control of industrial products and of manufacturing processes, the measurement of light intensities, scintillation counters, facsimile transmission, spectral analysis, and many others.

Photodiodes and photomultipliers integrate the light falling over the surface of the photoemitter to give the total amount received. Image converters, image intensifiers, television camera tubes, and image storage tubes take advantage of the point-by-point emission of the photocathode. In the first two examples, an optical image incident on a semitransparent photocathode is used to transform the light image into an "electron image." The electrons released by each element of the photoemitter are focussed by an electron-optical device onto a fluorescent screen, reconverting it in the process again into an optical image. This process can be carried out in a single step or in several steps, amplifying the image at each step.

Television camera tubes

In the most common type of television camera tubes, the first step, as with the image converter, is to transfer an electron image from the photocathode onto a thin insulating target, where, through emission of secondary electrons, a pattern of positive charges results. An independent electron beam scans the target, is influenced by the charges residing on its surface, and conveys the information to a secondary electron multiplier. The amplified signal goes to the external circuit.

Photoconductors, too, have been widely used, mostly in the form of thin film devices, for storage tubes, television pickup tubes, display devices, radiation-controlled variable resistances, electronic switches, image amplifiers, X-ray intensifiers, and, last but not least, for electrophotography.

As radiation detectors, photoconductors offer an extension of the spectral range to lower incident photon energies than obtainable with photoemitters. Infrared detectors have been described, such as indium–antimony detectors, that have a lower energy limit of 0.15 electron volt. Special detectors for the infrared are commercially available with a low energy limit corresponding to 0.030–0.040 electron volt. Germanium and silicon detectors are widely used for X-ray and gamma ray detection and spectral analysis.

Photovoltaic devices have been used extensively as photographic light meters and similar devices and have gained prominence as energy conversion devices, mainly because of their high efficiency in the visible region of the spectrum. Solar-cell arrays have been developed that can provide the necessary power for the operation of many electrical devices.

Finally, serious attempts are being made to apply the Auger effect for analytical purposes, attempts made possible because high-resolution electron-energy techniques can distinguish between Auger peaks belonging to neighbouring elements and thus can determine trace amounts of elements and their relative proportions.

BIBLIOGRAPHY. An extensive history of the photoelectric effect may be found in P. GORLICH, *Photoeffekte*, vol. 1 (1962); see also R.J. MAURER, "Photoelectric Effect," in E.U. CONDON and H. ODISHAW (eds.), *Handbook of Physics*, 2nd ed., pt. 8, ch. 5 (1967). A detailed treatment of photoelectric phenomena is given by G.L. WEISSLER, "Photoionization in Gases," and "Photoelectric Emission from Solids," in S. FLUGGE (ed.), *Encyclopedia of Physics* (*Handbuch der Physik*), vol. 21, pp. 304–382 (1956). Other works on this subject include: A.H. SOMMER and W.E. SPICER, "Photoelectric Emission," R.H. BUBE, "Photoconductors," and P. RAPPAPORT and J.J. WYSOCKI, "The Photovoltaic Effect," in S. LARACH (ed.), *Photoelectronic Materials and Devices* (1965), good summary articles; A.H. SOMMER, *Photoemissive Materials: Preparation, Properties and Uses* (1968); and M. DANOS and E.G. FULLER, "Photonuclear Reactions," *An. Rev. Nucl. Sci.*, 15:29–66 (1965).

(L.L.M.)

Photoengraving

Engraving is the broad term describing the procedure used in making printing plates, in which printing and nonprinting areas are distinguished by their height with respect to the general plane of the surface, the artistic decoration created by mechanically incising a design into a surface, and the creation of original works of art by tooling or etching an image into the surface of a metal (or plastic) plate and transferring the resultant image to paper. For detailed information on these last two subjects, see PRINTMAKING. This article is limited to consideration of the procedures whereby a printing surface useful in the production of multiple ink-on-paper images is produced.

The term photoengraving is correctly applied to the procedures discussed here, since the use of light energy, as involved in photographic processes, is essential. A distinction must be made between a relief (or letterpress) printing plate, in which the ink-carrying (or image-bearing) surface coincides with the general level of the plate surface, with nonimage portions cut below the surface, and intaglio (or gravure) printing surfaces, in which the ink-carrying image elements are incised into the plate surface. In letterpress printing, a uniform film of ink is distributed over the surface of the plate and transferred from the individual image elements to the receiving paper surface. In intaglio printing, the plate is flooded with a low-viscosity (thin) ink, then wiped with a blade (doctor blade) to remove any ink adhering to the surface. The doctoring action leaves the incised intaglio image filled with ink; later, as paper is brought into contact with this image and pressure is applied, surface-tension and capillary-action forces cause the ink to transfer from the plate to the paper.

Letterpress and intaglio printing

HISTORY

Wood engravings. The earliest engraved printing units were wood engravings, in which the nonimage areas of an illustration were removed by carving them from the surface of a flat wood block. The oldest known illustration printed from a wooden block was a Buddhist scroll discovered in 1866, in Korea. While the dating of the print is not exact, it is believed to have been prepared about AD 750. The Chinese *Diamond-Sūtra*, dated 868, incorporates a woodcut title page and text that includes numerous woodcut images.

From the 8th- and 9th-century dates above, it is clear that the use of woodcuts (images cut into a surface parallel to the wood grain) and wood-block engravings (images incised into the end grain of an assembled block) antedates the invention of movable type. The earliest extant example of a European print from a wood engraving to which a reliable date may be attributed is a print entitled "St. Christopher," dated 1423, discovered in the library of the Carthusian monastery in Buxheim, Germany. Another authenticated example of 15th-century wood-block printing is the "Apocalypse of St. John," printed in 1450, after a 14th-century manuscript.

Early etched plates. Plates engraved in wood continued to find use in printing application through the late-medieval and early-modern periods. Plates made of copper, pewter, and other metals were also produced, by a process in which an image in wax or bitumen was drawn on, or transferred to, the surface of the plate and nonimage areas removed by action of appropriate acids.

Preparation of intaglio printing plates by a procedure that involved coating a metal plate with an etchant-resis-

tant substance (ground) such as wax, bitumen, or shellac, scratching through this substance (ground) to expose the plate surface, then etching in acids is also a late-medieval European development. This process, however, developed as a medium of artistic expression, rather than a technique for the mass production of printed images.

Early use of light-sensitive materials

The first experimental application of light-sensitive materials to the production of printing surfaces was by Joseph Nicéphore Niepce, of France, an early researcher in lithography who began his experiments about 1813. He is credited with having produced the first permanent photograph. In 1826 Niepce coated a pewter or copper plate with a photosensitive asphaltum and exposed the surface to bright sunlight through an etching of a portrait, which served as a positive image. Sunlight passing through the background of the etching hardened the asphaltum, while the protected areas, under the inked portion of the etching, were developed in oil of lavender and white petroleum to create an image in exposed metal. This image was then etched into the plate, and from the intaglio image, prints were made on a copperplate press.

Though this basic discovery was of historical importance, it did not bring about the immediate use of photoengraved images for printing, and many other attempts to produce engravings by exploitation of the photosensitivity of various natural compounds were made by experimenters in Europe and the United States. The origin of the modern photoengraving process rests, however, on the report (1839) by a Scottish scientist and inventor, Mungo Ponton, of the light-sensitive properties of certain chromium salts (compounds). But Ponton, who demonstrated the chemical change that occurs when glue containing a dichromate (a compound containing two atoms of chromium) is acted upon by light, was not concerned with preparation of printing plates, and it remained for W.H. Fox Talbot, an English pioneer in photography, to propose the use of dichromated colloids such as albumin as an etchant-resistant for preparation of intaglio printing surfaces.

Early-19th-century work on production of chemically etched letterpress printing plates antedated, in many instances, the invention of photography. A researcher in Paris developed a process for the preparation of engravings on zinc. His work involved transfer of an image to the zinc plate by mechanical means, using ink or wax, and the removal of the nonprinting areas in a series of etching operations, each of which involved applying a coating of ink to the sidewalls of the etched lines by means of resilient rollers. The ink served to protect the lines of the engraving from the action of the etching acid, so that the printing area was not reduced.

Wet-collodion photography. The introduction in 1851 of a so-called wet-collodion process for photography provided a means for producing a photographic negative as the basic element in the preparation of engravings. In this process, a glass plate is coated with an alcohol–ether solution of collodion (cellulose nitrate) containing halide salts (compounds of such elements as chlorine, fluorine, etc. of the halogen family) of various alkali and alkaline-earth metals. While still wet, the plate is immersed in a silver nitrate solution, producing light-sensitive silver salts in the collodion layer. Without drying the film of collodion, the plate is placed in the camera and exposed, followed by development in ferrous sulfate solution and chemical "intensification" to produce an image of greater opacity. The image consists of deposits of metallic silver and other heavy metals imbedded in the collodion layer.

This photographic process also provided a method of stripping the photographic image from the glass plate, permitting assembly of a number of images for plate making, and also making possible the geometric reversal of the image needed in letterpress plate making to produce a right-reading print on paper. The wet-collodion process was used extensively in engraving until the 1930s, when the introduction of commercially coated stripping films led to its gradual replacement.

The halftone process. Since the letterpress printing process provides a uniform coating of ink on all printing elements, no provision can be made for reproducing tones intermediate between black and white by varying the thickness of the ink film laid down by the press. The production of shades of gray was then the role of the halftone process, in which the image is broken up into dots and variations of gray tones are obtained by varying the size of the dots, thus controlling the amount of ink laid down in a given area.

Halftone feasibility proved

The feasibility of the method was demonstrated about 1850, when a halftone image was produced by photography through a screen of loosely woven fabric. The screen was placed some distance forward of the plane of the receiving photographic surface (film or plate) and had the effect of breaking the gray tones of the subject into dots of varying sizes, through a combination of geometric and diffraction effects involving the spacing of screen from the image surface, the size of the openings in the screen, distance from lens to image plane, and the size of the aperture in the lens. It was obvious that a screen designed for this use could consist of a pattern on a glass or other firm, transparent surface.

A French patent of 1857 described a screen with parallel lines scratched in a single direction in an opaque background. As early as 1869 an image with a crossline halftone was produced in the *Canadian Illustrated News.* Later, in 1882, a crossline halftone was produced using a single-direction screen, by making half the exposure with the screen in one position and half with the screen rotated a quarter turn. Two brothers, Max and Louis Levy, of Philadelphia, in 1890 produced the first commercial halftone screens. The Levy brothers coated selected plates of high-quality optical glass with a lacquer, in which parallel lines were cut. The ruled lines were then etched with hydrofluoric acid and filled with an opaque material. Two such plates were cemented face to face with the lines at 90°, the edges sealed, and the assembly bound in a metal frame.

There has been no significant change in the methods of making halftone screens since those developed by the Levy brothers. Other screen patterns, including triangular dot patterns and the grained (mezzograph) screen, have been proposed, but none has produced consistently satisfactory results. For special effects, screens having straight or wavy line patterns and screens that produce a pattern of circular lines, concentric about a point that is chosen to be the focal point of the readers' interest, are in use. These screens are generally produced photographically from hand- or machine-drawn patterns and are used in the form of contact screens.

Basis for selection of screen ruling. Halftone screens may be obtained with line frequencies of 50 to 400 lines per inch (one inch equals 25.4 millimetres). The coarser screens are used for reproductions printed on coarse papers, the fine screens for higher quality reproductions on highly finished and coated papers. Screens in the 50–85-line frequency range are used primarily in newspaper illustration, while 100-, 110-, and 120-line halftones are suitable for highly polished papers and for some magazines, where single-colour and some multicolor work is involved. The 120-, 133-, and 150-lines-per-inch screens are generally used for colour illustrations in magazines and books printed on coated papers, when picture detail is important. Screens of 175 and more lines per inch are seldom used in letterpress printing, since the inks tend to fill the screens, causing difficulties in the press run. Such screens, however, do have some use in printing by offset lithography. In general, where paper quality permits, the finer screens are used when reproduction of fine detail is important. But since the letterpress process requires that the diameter of the finest highlight dot should not be less than 0.0015—0.002 inch, the use of very fine screens will lead to loss of image contrast, since some 3 to 5 percent of the picture area, in highlights, will be ink covered.

"Altone Gradar Screen"

An interesting development in glass screens was the "Altone Gradar Screen," manufactured in Germany. These are glass screens, ruled and etched in the usual manner, but with the rulings of the two glass elements filled with a transparent magenta lacquer of two different optical densities. When the screens are assembled, lines in one direction exhibit a density different from that of lines in

the perpendicular direction, and the intersections have a density equal to the sum of the densities of the two lacquers. The effect is to provide elongated halftone dots, with improved tonal reproduction in intermediate gray tones on coarse paper such as newspaper stock.

Contact screens. Perhaps the most significant recent advance in the halftone process has been the use of contact screens—films bearing a gray or magenta-dyed image of the light-distribution pattern behind a conventional halftone screen. The screen is placed in contact with the surface of a high-contrast film, in the image plane. The image, as recorded on the film surface, has the characteristic of a halftone exposed through a glass screen, with significant improvements in rendition of detail of the subject. Though first proposed in 1855 and developed by a number of later investigators, this technique was not fully exploited commercially until the 1940s. The contact screen eliminates certain diffraction effects inherent in glass screens, frees the operator from some of the lens-diaphragm restrictions imposed by the glass screen, and eliminates the necessity that lens opening, bellows extension, screen distance from the focal plane, and screen ruling all be in a particular relationship.

Contact screens are made with a silver (gray) screen pattern image and with a magenta dye image. The dyed screen gives some additional control over halftone negative quality through use of colour filters on the camera.

The "benday" process. An entirely mechanical procedure for production of a halftone image on a metal printing plate is the "benday" process (1879), named after its inventor, a New York newspaper engraver. This process utilizes a series of celluloid screens bearing raised images of dot and line patterns. The screen surface is covered with a waxy ink and the ink transferred, by pressure and rolling, to prepared portions of a metal plate. By selecting different screen patterns for transfer to different parts of the image, a mechanically produced halftone image is rendered. The ink image is reinforced with powdered resins and the plate etched. This process has been supplanted by completely photomechanical techniques.

Special effects. Such techniques as dropping out the highlights from the halftone negative (*i.e.*, eliminating the dots in these areas) in order to achieve increased contrast in illustrations were studied and introduced by several individuals. Such a method was patented in 1893, and in 1925 a camera attachment was introduced making it possible to impart a slight motion to the image on the film and thus reduce exposure to the point at which small highlight halftone dots were not printed or developed.

The most successful of the highlighting methods were those employing fluorescence phenomena, in which an object produces visible light when illuminated with ultraviolet light. In 1938, for example, the fluorographic process, in which fluorescing materials were incorporated in the artist's pigments, was patented. Similar pigments, designed for colour correction in watercolour illustrations, were patented in 1935 and 1938. Another process introduced shortly thereafter utilized a fluorescing paperboard. All of these processes were based on the same procedure: making an exposure under normal lighting for overall reproduction and then making an additional correcting exposure under ultraviolet light. The fluorescence produced by the ultraviolet illumination provided additional exposure in the affected areas that gave the necessary correction for highlighting or colour correction, by eliminating the screen pattern from "white" areas, in the case of monochrome, or reducing printing dot sizes, in critical areas of colour work.

Process developments. The discovery of the halftone screen was primarily responsible for the development and growth of photoengraving; further growth was related to other developments in the printing and allied industries.

Effect of colour film on printing developments

The introduction in 1935 of the first practical colour film for amateur and professional use probably did more to accelerate printing developments than any single invention. By making bulky studio-type colour cameras obsolete and permitting the use of readily portable camera equipment, previously limited to monochrome photography, for the production of colour images, on-the-spot colour photography became possible, greatly increasing the use of coloured illustrations. At approximately the same time, the commercial production of coated paper and heat-drying printing inks for letterpress printing began. Many colour developments for films, printing processes, and materials followed.

Electrolytic etching. The electrolytic principle, in which the plate being etched is an anode (positive electrode) in an electrolytic bath, was first applied to producing printing plates in 1912. The passage of direct current through the bath causes metal ions to leave the positively charged copper plate, resulting in an image in relief. The technique never achieved widespread popularity and has been supplanted today by new chemical etching processes except for specialized uses outside the graphic arts.

Chemical etching—traditional and powderless processes. Early methods of etching zinc and copper, methods that have persisted in some areas to the present day, were tedious and inexact and could be learned only through trial-and-error training. The principal difficulty stemmed from the fact that the chemical removal of metal from nonimage areas proceeds in all directions. Thus, etching done perpendicular to the plate surface to achieve the depth required for satisfactory printing also proceeds sideways, causing reduction in width of lines and dots of the printing image and also undercutting halftone dots—producing a below-surface dimension smaller than the printing surface. This leads to a mechanically weakened dot that may collapse under printing pressure.

Some success in overcoming this problem has been achieved by depositing an etchant-resistant material about the sidewalls of etched lines and dots, thus preventing lateral etching. The method of rolling a waxy ink onto sidewalls of lines and dots, called gillotage, has found wide use among European engravers. The "powdering" process, most widely used in the United States, involves brushing a resinous powder (dragons' blood) against the sides of partially etched lines and dots and fusing, with heat, to provide an etchant-resistant coating. A number of repetitions of the operation—etching, application of the protective material, and etching again, are needed before sufficient depth is attained. Results of this process are dependent upon the skill of the operator and on such ambient conditions as temperature and relative humidity, since these affect the performance of the powder. A major step toward solving the problem—in fact, the most important development in the field of etching since photoengraving was invented—came with introduction of a process of etching a magnesium plate without the use of powder. Experimenters found that by adding an oily material and a surface-active (wetting) agent to the nitric acid bath and controlling the conditions under which the plate was etched, they could produce characters in relief with adequate etching depth and virtually no printing-area loss during the etching. Later adapted to the etching of zinc, the process was quickly adopted by engravers in all parts of the world.

Powderless etching of copper

With this major hurdle in the etching of zinc and magnesium overcome, attention turned to copper, and in 1954 it was found that a powderless etching process for copper was possible if an organic compound (thiourea) was added to the iron chloride etching bath. Further refinements in the process and the introduction of new compounds to add to the etching bath followed.

Electromechanical plate making. While these developments in chemical etching were taking place, other experiments were being conducted to assess the feasibility of replacing traditional methods with the techniques of electronics, optics, and mechanics. The first successful result of these efforts was a device, introduced in 1947, that optically scanned a picture and simultaneously reproduced it as a relief printing plate on a plastic sheet. This device found wide application, particularly in newspaper plants, where photoengraving procedures were time-consuming and therefore costly. Within a short time, machines were developed that were capable of making etched plates in metals.

Meanwhile, investigators in the United States discovered about 1950 that some methacrylate compounds could be

quickly polymerized (turned into a polymer, or organic molecule with repeating chains of similar atoms) by exposure to light. Nylon was also found to have photosensitive properties, and by 1958 both materials were being offered for use in printing plates. Another plate-making system, reportedly based on light-sensitive urethane resins, was introduced in 1968.

Colour scanners. Paralleling the development of the electromechanical engraving machine, experimenters in the United States and Europe independently devised a number of electromechanical devices that automatically produce, from a colour-transparency image, corrected film negatives from which the four printing plates used in full-colour reproduction can be prepared. In one of these, a photographic transparency, wrapped around a glass cylinder, is scanned by a narrow beam of light. After passing through the transparency, the light continues through a colour splitter, and the blue, green, and red components are directed onto the sensitive surfaces of photocells. The electronic signals thus generated are modified and amplified in a computer that functions as an electronic analogue of photographic colour-separation processes. The computer activates a series of lamps, which expose the colour-corrected images onto photographic films mounted on another drum, attached to the same shaft as the transparency holder. This development was based on initial experimentation in commercial laboratories in the late 1930s and early 1940s. Other units, based on similar principles but differing in some details of structure and operating procedures, have been manufactured in the United States, West Germany, and Great Britain.

MODERN ENGRAVING TECHNIQUES

In terms of cost, engraving methods range in ascending order as follows: line engravings, halftone engravings, combination line-and-halftone engravings, single-colour, two-colour, and duotone engravings, and process colourplates. Each of the types may be produced in any of the customary metals or plastics. Process colourplates are usually made of copper in the United States and United Kingdom and of zinc elsewhere.

Basic production processes. The essential operations for the production of all types of photoengravings are similar. They include photography, photomechanical operations, etching, finishing, routing, blocking, and proofing.

The process camera

Camera and darkroom equipment. The engravers' camera, called a process camera, is a rigidly built machine designed to allow precise positioning of the lens and copyboard so as to provide control over the enlargement or reduction in size of the copy. It has a colour-corrected lens designed to give the sharpest possible overall image when focussed on a plane surface, without the distortions commonly found (though usually unnoticed) in the average portrait or amateur camera lens. Process cameras are designated as gallery or darkroom types. The gallery camera is freestanding and may be installed in any convenient location, but film must be removed in a light-tight cassette and processed in a separate darkroom. The darkroom camera is installed with its film holder as an integral part of the darkroom wall, giving easy access to the darkroom facilities.

The material to be reproduced, called copy, is mounted on a board or glass-covered copyholder, carried on the bed of the camera. Illumination for exposure is provided by arc lamps or high-intensity gas-discharge lamps. The most common camera lamp systems in late years have involved pulsed-xenon lamps, in which a high-voltage alternating current, passing through a glass tube containing the rare gas xenon, causes the emission of a light rich in the ultraviolet wavelengths.

Virtually all photographic work is done on film coated with high-contrast emulsions especially developed for graphic arts work. The introduction of dimensionally stable film bases has nearly eliminated the use of glass plates. Film emulsions used for halftones yield the extremely high contrast needed for halftone or line reproduction. Stripping film, a laminated film with a soft adhesive layer between the base and the emulsion layer, is widely used to permit images to be removed from the base and properly oriented on the glass or film flat through which the metal plate will be exposed.

In the early days of photoengraving, with wet-plate images on a glass support, it was impossible to process photographic images by any means other than immersion in solutions contained in a shallow pan or tray or by dipping into a tank of solution. Such tank and tray processing remains important but is now being supplanted by the use of automatic film-processing machines. Derived from equipment originally designed for processing of motion-picture film or photostat prints, these consist of belt- or roller-driven apparatus that carries the film through developer, fixing, and washing solutions, and, in most cases, through a drier, permitting delivery of a processed, dried film within three to five minutes after insertion into the machine. Such machines, with different processing solutions, may be used for continuous-tone or lith-type films.

Automatic film processors

Plate coating and printing. Photomechanical operations include cleaning the metal plate surfaces, coating with a light-sensitive solution, drying the coating (known as the top or enamel), and making the exposure on this coating through the negative prepared in the photographic step. Throughout these operations care is required to prevent imperfections such as bubbles, dirt, or scratches in the light-sensitive coating. The zinc, magnesium, or copper is prepared by careful cleaning with pumice and water. The light-sensitive coatings are usually poured over the surface and the plate, held flat, is whirled to ensure uniform coverage by the solution.

Light-sensitive coatings are usually a dichromated colloid material, but light-sensitive resins are also used. "Cold top" enamels are used on zinc and magnesium, which cannot be heated; these are usually slightly alkaline solutions of shellac or polyvinyl alcohol to which a dichromate is added. "Hot top" enamels nearly always contain fish glue as well as some egg albumin, to which is added a dichromate sensitizer. Mixtures of glue and albumin are used when it is necessary to control the etch resistance and the ease with which the edges of the enamel break away during the etching process. Hot top enamels must be set at temperatures of 550°–650° F (285°–345° C) and are used mainly on copper, the crystal structure of which is not altered at these temperatures. Polyvinyl alcohol and shellac resistants are set at temperatures of 350° and 220° F (175° and 105° C) respectively; therefore they are used on zinc and magnesium.

The tops are high-contrast materials that, when exposed to strong ultraviolet light, harden where the light has struck them and lose their solubility in water. Development in water then removes the coating from the unwanted areas of metal, exposing the metal for the etching process. Photosensitive resinous materials find wide application in electronic circuit printing, an operation analogous to photoengraving. They have more limited applications in the making of photoengraved letterpress plates, where they are used especially on zinc and magnesium and where their excellent storage properties permit their application in the metal-finishing plant, obviating the necessity for coating of the resist onto the metal in the photoengraving shop. These resinous materials are developed in organic solvents.

Etching and finishing. Nitric acid is commonly used in etching zinc and magnesium, the strength varying from 6 to 15 percent, depending on the metal. Copper is more readily attacked by ferric chloride (iron chloride), which is commonly used in concentrations of 28–45 percent. The etching may be done in an open tub or tray, though this method does not give the control needed for economical operation and is employed only where control is not critical. Most quality work is carried out in etching machines provided with impellers, devices that break up the etchant into a spray and force it against the plate.

In the conventional etching processes, the acid or iron chloride is used without modification, although great care is needed to prevent overetching. In many cases, especially when making line plates, etchers powder to protect the upper printing areas from attack while continuing to etch

in depth. The powderless etching processes, described earlier, are now almost universally in use, obsoleting the powdering technique. Line plates are usually etched to depths of 0.010 to 0.045 inch. Halftones may be etched to depths of 0.0023 to 0.009 inch, depending on the fineness of the screen. Coarser screens are etched deeper.

Photosensitive plastic plates are not etched in the ordinary sense. Unexposed resins, from nonprinting areas, are washed out with either dilute alkali or alcohol. Overetching is not a problem with this type of plate.

Finishing includes hand operations with engravers' tools, to remove imperfections in the image area of the plate and to improve its outward appearance. In colourplates, finishing also includes colour correction, a process of further etching or burnishing selected areas to improve the fidelity of reproduction. Finally, unwanted metal in the nonprinting areas of the plate is removed by a mechanical routing machine.

Blocking and proofing. Blocking consists of adhering the plates to cherry wood, plywood, or metal blocks to bring the printing surface to type height, which is 0.918 inch. Until the development of thermoplastic adhesives in the 1940s, blocking was always done by nailing the plates to wooden blocks. This tedious and costly operation has been largely replaced by hot mounting, in which process the plate is placed on a block of wood precoated with adhesive and this sandwich is subjected to heat and pressure. Upon cooling, the plate adheres firmly to the block.

Proofing consists in placing the plates on a precision press and taking sample impressions, or proofs that show how the plates will print during a regular press run.

Colourplate production. The first printed colour work was produced manually; artists painted in the necessary colours on black-and-white printed sheets. Later, stencils were used to speed this work, and in a further development, colours were printed, either as solids or tints, from hand-engraved plates. All of the work was crude by modern standards, however, and nothing approaching four-colour process printing was possible.

Modern colour printing

Modern colour printing, done with either three or four plates, each using a different colour of ink and overprinting the others, is based on a subtractive system of colours in which intermediate hues are obtained by some combination of two or more of the subtractive, or secondary, colours. The best colour printing is usually done with four process colours: yellow, magenta (blue red), cyan (blue green), and black.

The black plate is used to provide added uniformity of colour reproduction, since it will overcome changes in hue of critical neutral tones that could occur with random or cyclic variations in the amount of ink being transferred to the plate from the press inking system. Further, the use of a black plate aids in maintaining sharpness of picture detail.

In theory, black should result whenever the three subtractive colours are superimposed. Thus, it should be possible to produce black wherever all three of the secondary colours are present without affecting reproduction. Further, any colour that is within the range of colours reproducible with inks on paper can theoretically be obtained by using only black plus the proper pair of the secondary colours. But this has not been found practical because of the nature of printing ink pigments and the lack of total precision in the printing operation. Consequently, it is common practice to use the black plate to supplement the colourplates, portions of which are allowed to print in all except pure white areas of an illustration. The colourplates and the black plate must all be printed in register; *i.e.*, they must be superimposed so that identical portions of the image in each plate colour overprint each other.

Colour separation. In manufacture, the production of an individual colourplate involves the same steps used in producing an ordinary black-and-white engraving, once the etchant-resisting image has been printed on the metal. Prior to this, the only differences lie in the use of colour filters on the engraver's camera and in steps to reduce the range of colour contrast of the copy. Negatives representing the images to be printed with each of the coloured inks are obtained by photographing the colour copy through colour filters. These filters, usually used in the form of thin sheets of dyed gelatin inserted into the lens, are complementary in colour to the coloured printing inks used.

Masking is the use of positive or negative images, taken from one or more of the set of colour-separation negatives and used in register with a given negative, to correct for the deficiencies in printing inks and colour of the copy. Common colour errors corrected by masking include the removal of excessive yellow values and magenta values from the blue (yellow printer) and green (magenta printer) negatives.

Colourplates may be made by the use of two general photographic methods—one indirect and one direct. The indirect method produces either continuous-tone negative images, from which halftone negatives are made, or continuous-tone negatives, from which continuous-tone positives are prepared. In the direct method, screen negatives are prepared directly from the copy through the colour-separation filters and a halftone screen onto a high-contrast panchromatic film or plate to produce a negative ready for transfer to the metal plate.

Proofing of colourplates

The proofing of halftone colourplates for wet printing on high-speed presses (when one colour does not have time to dry before the next is laid down) is a critical operation, for the proofing must be carried out under conditions simulating as closely as possible those that will be encountered on the production press. Specially built proof presses make this possible. In appearance they resemble four conventional press units placed end-to-end, and the sheet of paper is passed in turn over the four plates. However, because the production press employs not the original flat plates but curved duplicates made from them, and because ink and paper specifications are highly variable, exact duplication of production results in a proofing operation is difficult.

Elimination of moiré. A serious problem in colour reproduction is the occurrence of an interference pattern, or moiré, caused by the overprinting of the screens in the colourplates (a similar effect can be obtained by superimposing two pieces of window screening or fine net cloth). Because it is impossible to maintain printing register within the degree necessary to avoid such an effect, it is common practice to rotate the halftone screen when making the negatives so that each of the four plates has its screen pattern in a different position.

Electromechanical engraving machines—colour scanners. Reference has been made to devices for the electromechanical production of relief printing plates. The first of these utilized a heated pyramidal stylus, the motion of which was controlled by an electrical signal from a scanning photocell, to penetrate a plastic plate to a distance inversely proportional to the optical density of copy, thus burning out varying areas from the plate surface. In another machine of the same general type, an oscillating gouge cuts a halftone pattern in a flat plastic or metal plate, under control of a signal from a scanning photocell; in yet another a spiral groove, of varying width, is cut into the surface of a plastic plate wrapped on a rotating cylinder.

The colour scanner has been described elsewhere in this article. The first such units were capable only of producing colour-separation negatives of the same size as the copy that was scanned. In later developments, circuits were provided to produce positive images, and mechanical or electronic devices were developed to allow enlargement or reduction of the size of the final image as compared with size of the original copy. When scanners were first made available, it was believed that their cost would limit their use to a few large plate-making establishments. In the early 1970s, however, acceptance was exceeding expectations.

Production specifications. These include specifications for line plates, halftone specifications, and combination plates.

Line plates. In line illustrations all of the image areas are either black or white, and hence no halftone screen is

required to copy them for use in making a printing plate. Suitable copy consists of line drawings, etchings, etc. The negative as it comes from the process camera is suitable to transfer the line image onto the metal.

Plate preparation, coating, burning in, etching, and finishing are essentially the same as for halftone plates. Certain specifications must be met, however. The nonprinting areas must be etched sufficiently deep to prevent the ink rollers from touching them on the press, and to prevent them from rubbing on the surface of the paper during wet colour printing. For presses with accurately adjustable ink rollers, the etch depth may be as little as 0.01 inch. The same depth is permissible in thin, wrap-around press plates. For conventional printing presses, the minimum etch depth is about twice this. Plates that are to be duplicated by electrotype or stereotype processes may require slightly greater depths, although normal etching ordinarily is sufficient to produce good duplicates. Plates from which rubber duplicates are to be made will require etch depths as great as 0.045 inch.

Halftone specifications. Etch depths in halftone plates need not be as great as those in line plates, but the contour of the halftone dot and the depth of the etched areas are very important. Etch depth in highlight areas, the most critical portions of halftones, varies from 0.006 inch in a 65-line halftone to 0.002 inch in a 133-line.

Combination line-and-halftone plates. These plates must be prepared by assembling, in the negative form, the halftone and the line portions of the illustration and then, after transferring them onto the metal, etching them in two operations, so as to attain the best results for both portions. The powderless etching processes, however, permit easier etching of coarse-screen combination plates for use in newspapers. Combination line-and-halftone plates may also be produced by making two plates in separate operations and mounting them on a single block in proper position with respect to each other.

Engraving techniques applied to intaglio printing processes. Procedures similar to those described for production of letterpress printing surfaces are applicable to the production of intaglio printing surfaces. In intaglio, or gravure, printing, the image to be transferred to paper is etched or incised into the surface of the printing plate or cylinder. The entire surface is covered with ink, and by means of a doctoring, or wiping, operation, excess ink is removed from the surface, leaving only that which is retained in the image areas. Paper is brought into contact with the surface, and, under high mechanical pressures, the ink transfers from plate to paper.

Intaglio printing surfaces

Intaglio printing surfaces are of two general types: line intaglio (sometimes referred to as copperplate gravure), in which the ink-retaining image consists of discrete lines that may vary in width and depth; and gravure (also known as rotogravure), in which both continuous-tone and line copy are reproduced as a series of tiny cells etched into the printing surface. These cells commonly vary in depth, and hence in the volume of ink they will retain. Variations in density are produced on paper by the different amounts of ink that the cells transfer to paper.

In the rotogravure printing process, the walls surrounding each cell act as a support for the doctor blade that removes ink from the printing cylinder or plate surface.

Line intaglio. This process is widely used in the production of bank notes, securities, stamps, and engraved documents. The distinctive sharpness of fine lines and readily discernible differences in ink thickness that the process produces make it a preferred technique for production of bank notes and securities. These appearance characteristics cannot be readily counterfeited by photomechanical processes.

The printing surface is created either by mechanically scratching an acid-resistant ground from the plate surface, as described above, or by use of a photographic positive of the desired line pattern to prepare a photoresist image on the metal. The image is etched into the plate, using the techniques of letterpress line etching, with maximum depth of etch usually less than 0.007 inch. Metals commonly used may be steel, brass, or copper.

When a mechanical engraver is used to expose the metal for etching, a pointer or stylus is used to follow a usually enlarged pattern in a metal or plastic master stencil, causing a diamond stylus, in contact with the lacquer-covered plate surface, to remove the lacquer in a sharply defined pattern. The intaglio image is then prepared by etching the exposed metal with appropriate chemicals.

In printing from intaglio forms, the plate is flooded with an ink of medium viscosity and the surface of the plate wiped clean with either a metal doctor blade or a piece of hard-surfaced paper. To overcome wear of the plate from the abrasion of the wiping mechanism, the surface is frequently protected by an electroplated chromium layer.

Wiped free of excess ink, the plate is brought into contact with the paper surface. A roughly outlined relief image (counter) of the printing pattern is often used to provide high local pressures, forcing the paper into the ink-filled intaglio image. As the paper is pulled from the plate, capillary-attraction and surface-tension forces act to pull the ink from the plate. After drying, the image has a distinctive appearance in which the ink has appreciable thickness, and thin lines have less thickness than wider lines.

Gravure (rotogravure). The gravure printing process is one of the three major processes used for catalogs, magazines, newspaper supplements, cartons, floor and wall coverings, textiles, and plastics. Gravure printing is done with flat plates or, more commonly, with cylindrical surfaces. A screen pattern is superimposed over all image areas; thus the edges of type or lines printed by gravure will have a rough, or sawtoothed, appearance. This does not detract from readability.

Early gravure work

Early work described above formed the basis for modern gravure engraving and printing. Karl Klič (Klietsch) of Bohemia, who was instrumental in making photogravure a practical commercial process, in 1878 exposed a positive transparency over carbon tissue, a film made of coloured gelatin sensitized with potassium dichromate and backed by a sheet of paper. The exposed film was pressed down on a copper plate coated with an even layer of resin or asphalt powder. The carbon tissue was developed in water, making the gelatin swell in inverse proportion to the exposure it had received. The plate was etched with ferric chloride in successive baths of varying strengths. Penetration of the tissue by the etchant, and hence the resulting depth of etch in the metal surface, was controlled by the degree of swelling of the gelatin. Klič's process produced sure and predictable results and became the preferred method for later workers.

In later developments, the irregular grain pattern, produced by use of resinous powders, was replaced by a regular overall pattern of intersecting lines, produced by exposing the carbon tissue to a glass screen bearing an overall pattern of clear lines, intersecting at right angles.

In rotogravure, separate negatives of type matter, other line copy, and continuous-tone copy are assembled and positioned according to a prepared layout. After the negatives are retouched, a continuous-tone positive is made and retouched to ensure desired density values.

A sheet of carbon tissue is next exposed under a gravure screen. This screen is a film or sheet of glass on which fine transparent lines, usually 150–175 per inch, cross at right angles to form opaque squares. The lines of the screen are positioned at an angle of 45° to the axis of the printing cylinder. Their purpose is to provide a support for the doctor blade, which wipes ink from nonprinting areas. The lines of the screen allow the light to penetrate to the film and harden the gelatin. The square "islands" remain soft.

Next, the continuous-tone positive is placed in contact with the carbon tissue and exposed under an arc light. The soft squares are hardened in proportion to the amount of light that penetrates the varying grays of the positive. The carbon tissue is pressed with a rubber roller to a cylinder that has a polished electroplated copper surface. After adhering the exposed carbon tissue to the cylinder surface, it is "developed" with warm water, which has the effect of swelling and removing unhardened gelatin. The result is an image in hardened gelatin, which

varies inversely in thickness according to the density of the photographic positive. The cylinder is rotated in a tray of ferric chloride, producing an etched image in the copper surface. The squares are etched to varying depths, depending on the degree to which they were hardened. The crosslines of the screen, which were entirely light-hardened, are not etched at all. In this way, pits or wells of different depths are etched into the copper. For very long press runs, the cylinder can be plated with nickel or chromium.

Rotogravure printing

In rotogravure printing, the cylinder usually is arranged so that during its rotary movement it passes through a trough filled with a thin solution of fast-drying ink. A thin steel doctor blade moves across the cylinder with a slight oscillating action and removes the ink from the surface, but not from the wells beneath. The cylinder then comes in contact with the paper, and the paper draws the ink out of the wells in the plate. After being printed, the paper shows through thin deposits of translucent ink, thus creating pale grays; heavier ink deposits from the more deeply etched wells appear correspondingly opaque. Thus a full range of tonal values can be printed. Since the ink used in rotogravure printing is quite fluid, it penetrates through the pores of the paper surface, obliterating the screen pattern. In reproducing illustrations, gravure comes closest to simulating continuous-tone copy. In colour printing, a separate cylinder is prepared for each colour.

Other methods. In the so-called Dultgen halftone intaglio process, which is widely used in colour work, two positives are made from the continuous-tone copy, one through a halftone screen or a special contact screen and the other without a screen. The carbon tissue is first exposed to the screened positive, which produces an image of dots of varying sizes, then to the continuous-tone positive, which produces differing degrees of hardening of the dot image. When etched, the dots are of differing sizes and of differing depths. This method thus uses two methods for controlling tonal values.

The Henderson process, sometimes referred to as "direct transfer," or "inverse halftone," gravure, has won some acceptance in the printing of packaging materials. Retouched continuous-tone positives are used in preparation of halftone negatives and, by a contact-printing operation, halftone positives. These positives show dot size variations proportional to the desired print density. The cylinder is coated with a cold-top photoresist, as in letterpress engraving. This resist is then exposed to ultraviolet light, through the positives, and the image developed. The cylinder may then be etched either in ferric chloride solution or in a powderless etching bath, similar to that used for letterpress photoengraving. Tonal variation in the resulting print on paper is caused almost entirely by variation in the area of dots.

BIBLIOGRAPHY. J.S. MERTLE and G.L. MONSEN, *Photomechanics and Printing* (1957), a good basic text, emphasizing equipment and techniques of use, now obsolete in many technical details; W.J. SMITH, E.L. TURNER, and C.D. HALLAM, *Photo-engraving in Relief*, 3rd ed. (1951), basic information, should be supplemented with more recent texts; R.W.G. HUNT, *Reproduction of Colour*, 2nd ed. (1967), an authoritative presentation of fundamentals of colour printing and platemaking; F.G. WALLIS and R.V. CANNON, *Letterpress Platemaking* (1969), a recent general work on the platemaking process, recommended as both comprehensive and current; H.M. CARTWRIGHT, *Ilford Graphic Arts Manual*, vol. 1, *Photo-engraving* (1962), an excellent basic text, combining historical data and practical information, with bibliographies for further study; *Photogravure*, 2nd ed. (1939), a good presentation of principles and basic technology; and with R. MACKAY, *Rotogravure* (1956), a recent work on the technology of gravure; H. DENISON, *A Treatise on Photogravure* (1894), a classic volume dealing with this process.

(P.F.B.)

Photography, Art of

Photography is essentially the production of a permanent record of an image by the combined action of light and chemical processing. This article is concerned with still photography as an art medium. The technology of such areas as lenses, cameras, films, and processing is treated in PHOTOGRAPHY, TECHNOLOGY OF. Motion pictures are covered in MOTION PICTURES, ART OF and MOTION PICTURES, TECHNOLOGY OF.

As a means of visual communication and expression, photography has marked aesthetic capabilities. In order to understand them, the characteristics of the process itself must first be understood. Of these the first is immediacy. Usually, but not necessarily, the image that is recorded is formed by a lens in a camera. Upon exposure to the light forming the image, the sensitive material undergoes changes in its structure; a latent image is formed, which becomes visible by development and permanent by fixing. With modern materials, the processing may take place immediately or may be delayed for weeks or months. But, either way, the elements of the final image are determined at the time of exposure. This characteristic is unique to photography and sets it apart from other ways of picture making. Although the photographer can control the character of the original image that he captured upon film by the way he develops the negative and prints it, he cannot alter it except by manual techniques.

Immediacy of the photograph

A second characteristic of the photograph is that it can contain more than the photographer can see. The first daguerreotypes, shown to an astounded public in Paris in the winter of 1838–39 by the inventor Louis-Jacques-Mandé Daguerre, were praised because of the amount of detail recorded by them; looking at one with a magnifying glass, it was said, was like looking at nature with a telescope. The rival inventor of photography, William Henry Fox Talbot, after noting this characteristic, commented:

> It frequently happens, moreover—and this is one of the charms of photography—that the operator himself discovers on examination, perhaps long afterwards, that he has depicted many things he had no notion of at the time. Sometimes inscriptions and dates are found upon the buildings, or printed placards most irrelevant, are discovered upon their walls: sometimes a distant dial-plate is seen, and upon it—unconsciously recorded—the hour of the day at which the view was taken.

As photography has become more sophisticated in its technique, its scope has expanded enormously. High-speed photography has made visible motion man has never before seen; with material sensitive to invisible radiation, hidden aspects of nature can be revealed; and, by a combination of photographic, electronic, and space technology, even the planets can be observed in new ways.

Authenticity of the photograph

The seemingly automatic recording of an image by photography has given the process a sense of authenticity shared by no other picture-making technique. The fact that the photograph can show more than the eye can see and that the image is not filtered through the brain of a man and put down by the skill of his hand has given it value as evidence. Ralph Waldo Emerson wrote in his *Journals* of the daguerreotype process: "Tis the true Republican style of painting. The artist stands aside, and lets you paint yourself." The photograph has become, in the popular mind, so much a substitute for reality and of such apparent accuracy that the adage "The camera does not lie" has become a cliché. The tattered photograph a traveller takes out of his wallet to show someone is not, to him, a picture of his child: it is a substitute. Even people who do not travel know the distant parts of the Earth from photographs. And when a person does travel, he often seeks that which he knows from photographs and brings back his own photographs, in which he and his family may sometimes appear, as evidence that he has been somewhere.

This intrinsic characteristic is of such strength that it has dominated the evaluation of photography as an art form. Indeed, photography has been used over and over again as a foil by art critics to denounce paintings that rely heavily upon exact representation of subject matter.

Technical controls

In truth, photography is not the quasi-automatic process that is implied in a technical definition. The photographer has at his command a wide variety of controls. The image can be modified by the optical characteristics of the lens. The type of sensitive material used to record the image is a further control, and the contrast between highlight and

shadow can be changed by variations in development. In printing the negative, the photographer has a wide choice in the physical surface of the paper, the tonal contrast, and the image colour.

The most important control is, of course, the photographer's vision. He chooses the vantage point and the exact moment of exposure. Through experience he knows how the camera will record what he sees. He learns to previsualize the final print. If he has visual imagination and perception, he can make more than a passive record. He can express universal qualities. He can extend the vision of the viewer.

So facile a medium is photography that it is difficult to grasp its aesthetic capabilities and accomplishments. Of the billions of photographs that are taken every year, few can be considered art. Only a relatively small number of camera users are deliberately concerned with the production of photographs to be judged as art. A far greater number look upon photography as a means of communication. While the aim of the commercial photographer, the photojournalist, and the scientist may not primarily be aesthetic, it is significant and remarkably characteristic of the medium that often in their work can be found memorable pictures that reach beyond the particular to the universal. Recognition plays an overwhelming role in photography: recognition by the creative photographer of the picture possibilities presented to him and recognition by the creative editor of aesthetic qualities in photographs that he sees.

This article is divided into the following sections:

I. Equipment, materials, and technique
  Lenses
  Cameras
  Controls in exposure and processing
II. Stylistic and historical development
  The pioneers
  Developments to World War I
  Developments since World War I

I. Equipment, materials, and technique

Creative photography demands a thorough knowledge of photographic technique. By experience and discipline the master photographer works out a personal technique that is adapted to his aesthetic purposes. To assume that there is a standard universal technique is to deny the basic controls that every artist in every medium depends upon. On the other hand, to overemphasize technique to such an extent that it becomes the master, rather than the servant, of the photographer is as great a mistake as to underestimate its importance.

LENSES

Creating the image with the lens

The lens, the most vital part of the photographer's equipment, is the extension of his own eye: through it he creates the image. Perspective, which from the days of the Renaissance to the dawn of the photographic age in the 19th century was a baffling problem to the artist, is automatically created by a lens. So accurate is its image that the rules of perspective can be used in reverse: ground plans of architecture and contour maps of the Earth can be drawn from photographs. When a photographic image looks unnatural, it is likely to be because of the viewer's preconceived notions of "normal vision." Human vision embraces a restricted angle of view. Furthermore, eyes are never static: they scan the field continuously, viewing details in the distance at one instant, those close to at another, and those to the right or to the left at yet another. This visual report is instantly synthesized by the brain. The eye of the camera, its lens, is static. Its field of view is determined by its design. An extreme wide-angle lens can embrace a field of 180°. A long focus telephoto lens may limit the field to but a few degrees.

The size of the photographic image as well as the field of view depends on the focal length of the lens—*i.e.*, the distance between the optical centre of the lens and its image when the camera is focussed at infinity. The longer the focal length is, the smaller is the angle of view and the larger the image. The choice of focal length is a valuable control for aesthetic effect. With a wide-angle lens, the photographer can make foreground objects appear large. With a long-focal-length lens, he can visually compress the distance from foreground to background. Most portrait photographers prefer a lens that has a longer-than-normal focal length to avoid distortion problems.

Controlling the depth of field

Depth of field—or, as it is more commonly but incorrectly called, depth of focus—is a critical control for the photographer. He can render a scene sharp from the immediate foreground to distances as great as infinity, or he can put everything out of focus except what lies in a narrow zone. One of the factors in depth of field is the aperture or lens opening. Decreasing the aperture increases the depth of field. The focal length of the lens is also a factor: the shorter the focal length, the greater the depth. A third factor, distance to the subject, also enters into the depth of field. The closer the distance, the narrower the depth. Thus, the photographer has a number of ways of achieving the depth of field that expresses his intention.

The modern camera lens may reveal vastly more detail than was visible to the naked eye. Lenses, in the past, were deliberately designed to produce soft images. These lenses were particularly popular at the turn of the century, when the ability to render a great amount of detail was considered by many critics to be an aesthetic disadvantage. Today, when the photographer wishes an unsharp image, he uses techniques that allow him to alter the sharp image that his camera lens normally produces.

CAMERAS

The choice of camera depends upon what the photographer wishes to accomplish. If he seeks maximum precision of image, with relatively static subject matter, he may find the classical view camera mounted on a sturdy tripod best suited to his purpose. On the other hand, for swift action and work under low levels of illumination and when complete mobility is essential, a lightweight hand-held camera is preferable.

View cameras. The view camera takes large negatives on glass plates or sheet film. It is of simple construction: a bellows with a ground-glass viewing screen at one end and a board at the other end to which any one of a wide variety of lenses can be mounted. A special feature of many view cameras of great importance to the creative photographer is that the back of the camera can be swung on both the vertical and horizontal axes and its lensboard raised, lowered, or tipped. This flexibility enables the photographer to control the image formed by the lens. This control can be used, for instance, to solve the common problem of photographing a tall building: if the camera is tilted upward, all vertical lines will converge. By swinging the back so that the film plane is exactly perpendicular, the verticals are imaged as parallel lines. If there is too much foreground, the lens board can be raised to eliminate it, without altering the perspective. These controls can be used to create an image that conforms to conventional perspective or can be used for deliberate distortion.

View-camera movements

The view camera gives the photographer maximum control over his image and the large negative size gives an image of maximum definition. It is ideal for landscape, architecture, and formal portraiture—subjects that can be studied before exposure and that will remain relatively static while preparations are made for exposing the film.

Hand cameras. The modern hand camera differs dramatically from the view camera in terms of mechanical complexity, automatic features, and convenience of use. The operation of hand cameras is largely automatic. After one exposure the camera can be reset by moving a single lever. Consequently, photographs can be taken in rapid sequence. Although the negatives are smaller than those produced by a view camera, with modern precision optics, excellent prints can be made from them.

The role of the hand camera

The hand camera seems best suited for recording the stream of life. Its ease of operation allows the photographer to make many exposures, from which at some later date he may choose the image that best expresses his vision.

Hand cameras usually accommodate either roll film or 35-mm film. The negative size with roll film is usually 2¼ by 2¼ inches; with 35-mm film it is usually one by 1½ inches. The types most commonly used for creative photography are single-lens reflex and twin-lens reflex cameras in the roll-film size and single-lens reflex and rangefinder cameras in the 35-mm size.

Roll-film cameras. The single-lens reflex roll-film camera is fitted with ground glass on its top onto which the image formed by the lens is reflected by a mirror set at an angle of 45°. An instant before exposure, the lens is automatically set to the desired taking aperture, or stop, and the mirror moves out of the way to allow the image to fall directly on the film. This arrangement enables the photographer to view the exact image right up to the moment he presses the shutter release. By simply turning a crank, fresh film is brought into the focal plane, the mirror is put in its 45° position, the shutter is cocked, and the diaphragm is opened.

The twin-lens reflex is of simpler construction. It has two lenses of identical focal length. The upper lens, which has neither shutter nor adjustable diaphragm, focusses the image on the ground glass by a mirror permanently fixed at an angle of 45°. The lower lens focusses its image directly on the sensitized film and is fitted with a shutter and adjustable diaphragm. Although of much simpler construction than the single-lens reflex, it has the disadvantage that the user does not see the exact image of the picture he is taking but an approximate one, from a viewpoint slightly higher.

35-mm cameras. The 35-mm single-lens reflex camera operates on the same principle as the larger roll-film version, but the ground glass image is not usually viewed from above but at eye level, through a prism.

The 35-mm rangefinder camera does not permit the photographer to view the image of the lens. Instead, the camera is fitted with an optical finder that shows the field of view and also permits exact focussing by a rangefinder. The photographer focusses the lens by superimposing two images that are visible in the viewfinder.

CONTROLS IN EXPOSURE AND PROCESSING

The operation of the camera gives the photographer one kind of creative control. Another kind is offered by the selection of the film, the processing of the film, and the printing and processing of the print.

Film characteristics

Films. The photographer has a wide choice of negative films, varying widely in light sensitivity, grain, and contrast, among other characteristics. Films that are highly sensitive to light (so-called fast films) tend to be grainy. Processed film consists of a myriad of silver grains. The size of these grains directly affects the quality of the image, particularly when enlargements are made from the negatives. There are occasions on which a photographer may deliberately induce grain for aesthetic purposes, but usually it is desirable to hold grain to a minimum. Films that are less sensitive to light (the so-called slow films) tend to be less grainy. In some films the contrast range can be so narrow that there are no middle tones. Such films, called process films, can be used to produce poster-like effects from scenes having wide tonal ranges.

Modern panchromatic film is sensitive to all the wavelengths of the visible spectrum and even beyond the visible. This range of sensitivity gives the photographer a way of precisely and often dramatically controlling the tones of pictures. Blue sky can be rendered a dark gray or black by putting an orange or red filter over the lens during exposure. Since the red absorbs both blue and green, none of the blue rays present in the sky fall upon the film. In general, a filter lightens objects of its own colour.

With infrared film—that is, film that is particularly sensitive to infrared radiation—startling effects are possible, such as jet black skies, penetration of atmospheric haze, and a stark white rendering of some foliage.

Film processing. The processing of the negative offers the photographer further controls. The purpose of development is to convert to silver those silver halide crystals suspended in the gelatin emulsion that have received light. The purpose of fixing is to remove the silver halide crystals not affected by light. The rendering of tones, which is the very essence of the photographic process, is a function of both exposure and development. Between highlights and shadows an infinite number of tones can be recorded on the film, providing that the brilliance of the highlights does not exceed more than approximately 128 times that of the shadows. If the range is greater, the photographer must sacrifice detail either in the shadows or in the highlights. If the brilliance range is less than one to 128, he can increase the contrast between highlights and shadows. There is no fixed optimum exposure time or developing procedure; the combination of these factors that will produce a print that expresses what the photographer visualized just before releasing the shutter is determined by experience or by calculation.

Printing. While the photographer has many technical controls over his negative, it is in the printing of the negative that he has the greatest control. The U.S. photographer Ansel Adams has used a musical simile: the negative is the score, the print is the performance.

One of the chief characteristics of photography is that the final picture can be made of almost any size. The photographer's first decision, then, is to determine the format of the print. He may wish to retain as much detail as possible and will make his print the exact size of the negative by contact printing, so called because paper is laid in contact with the negative in a frame and then exposed to light. By using an enlarger, which projects an image of the negative on printing paper, he can produce a picture of mural size. How far a negative can be blown up depends on its size, its grain, and the amount of loss of definition that can be accepted.

A second characteristic of photography is the ease with which a portion of the negative can be used for the final picture. This is, of course, not unique to photography. Although it rarely happens, an artist using traditional mediums can present only a part of his original concept. It is common, however, for photographers to trim, or crop, their pictures either after they have printed them or in the process of doing so, by choosing small areas of the negative for enlargement. In photographic aesthetics there are two schools of thought: those who believe that the photographer should previsualize the final print at the moment of exposing the negative and those who believe that the making of the negative is but the first step and that the creative process of selection, emphasis, and composition extends into the darkroom.

Different views of photographic aesthetics

The surfaces of printing papers, their colour and their response to light (sensitometric characteristics), vary widely. The photographer who seeks maximum rendition of detail will usually prefer glossy paper. For subjects with bold, linear compositions, a rough-textured paper may be more suitable. By a simple chemical treatment the blue-black tones of modern printing papers can be converted to a rich brown. If he wishes, the photographer can make his print almost any colour.

As noted earlier, the contrast range of the negative can be controlled by exposure and development. Contrast can also be controlled by the selection of the paper. From the same negative the photographer can produce prints that are "flat," with shadows and highlights rendered as gray tones, or brilliant, with shadows approaching black and highlights white.

Photographers of the 19th century did not limit their printing papers to the gelatin and silver bromide or silver chloride emulsions that are almost universal in the 20th century. They also used papers made light sensitive with iron, platinum, and palladium, or with a mixture of a colloid, such as gum arabic, with potassium bichromate. Often they sensitized the paper themselves, which gave them a wide choice of surfaces. Some of these processes allowed the photographer to control development by manual intervention. Many of these long-obsolete processes are being revived by photographers who seek a greater control over the image than straight photography permits.

There is also a recent tendency to combine photography

Combining photography with other artistic mediums

with other artistic mediums. An artist may introduce a print into a painted or drawn composition. Photomontage, the presentation of two or more photographic images together as a single composition, can be done by pasting cutout photographic prints on a support or by projecting separate negatives on photographic paper that has been masked so that only portions are exposed at a time. By the latter technique, double exposures can be made and thus a wide variety of effects, unique to photography, can be produced.

Colour photography. Many of the creative controls characteristic of black-and-white photography can be used in colour photography. During the original photography, pale-colour filters can be used to alter colour rendition. Different films have different colour characteristics, some emphasizing warm tones, others cool tones. Throughout most of the history of colour photography, the processing of colour film and the making of colour prints has been such an exacting and complex matter that photographers left it to laboratories that specialized in such work. In recent years, however, the processes have been simplified and are being carried out by more and more individual photographers, a change that has opened the door to creative experimentation.

While most colour photographers strive for the reproduction of the colours of nature as they appear to the eye, startling effects can be obtained by departing from reality. Existing colour processes are based on securing records of the red, the green, and the blue rays in negative form and then converting them to the respective complementary colours: cyan, magenta, and yellow. If this order is deliberately changed, completely unreal images are obtained. By the use of infrared emulsion together with the colour records, even greater changes can be made.

II. Stylistic and historical development

THE PIONEERS

Photography came into being through the artistic aspirations of two Frenchmen, Joseph-Nicéphore Niépce and Louis-Jacques-Mandé Daguerre, and two Englishmen, Thomas Wedgwood and William Henry Fox Talbot, at the beginning of the 19th century.

Contributions of Niépce and Daguerre. Niépce, an amateur inventor living in Chalon-sur-Saône, a city 189 miles southeast of Paris, made his contribution to photography through his interest in lithography. In this process drawings were copied by hand onto the lithographic stone. To make the drawings Niépce relied upon his son's artistic skill, but, when his son entered military service, he was left without a draftsman. Not artistically trained, he began to work out a way by which light could draw the pictures he needed. Using a type of asphalt, bitumen of Judea, which changes its solubility in oil of lavender according to its exposure to light, he succeeded, in 1826, in making a crude photograph of the barnyard from an upper window of his country estate. The exposure was so great—a matter of several hours—that the shadows were obliterated as the sun moved from east to west. But the world's first photograph, now preserved in the Gernsheim Collection at the University of Texas, showed the path that others were to follow with more success.

Daguerre was a professional scene painter. Between 1822 and 1839 he was co-proprietor of the Diorama in Paris, an auditorium in which he and his partner Charles-Marie Bouton displayed immense paintings, 45½ by 71½ feet (14 by 22 metres) in size, of famous places and historical scenes, done in a style that emphasized illusion.

Like so many other artists, Daguerre made his preliminary sketches by tracing the images produced by a camera obscura—a box with a lens at one end, which threw an image on a piece of ground glass. At some time, the precise date is unknown, Daguerre began experiments in recording the camera image "by the spontaneous action of light," to use his own words. Learning of Niépce's work, he wrote to him, and in 1829 the two men formed a partnership. On Niépce's death in 1833, Daguerre continued the research alone, and by 1837 he had produced a photograph of his studio on a silvered copper plate, a photograph that was remarkable for its fidelity and detail.

In 1839 Daguerre and Niépce's son sold full rights to the daguerreotype and the heliograph (as Niépce had named his process) to the French government, in return for annuities for life. On August 19 full working details were published.

The daguerreotype process

To make a daguerreotype, a silver-plated sheet of copper was made light sensitive by exposing the silvered side to the fumes of iodine, which produced silver iodide. After exposure to light in a camera, the plate was exposed to fumes of mercury, which formed an amalgam with the nascent silver formed in the areas of the plate that had been exposed to light. The picture was fixed (made permanent) by removing the unaffected silver iodide with sodium thiosulfate, then called the hyposulphite of soda. Daguerre wrote a booklet describing the process, *An Historical and Descriptive Account of the Various Processes of the Daguerreotype and the Diorama*, which at once became a best seller: more than thirty editions and translations appeared before the end of 1839.

The contributions of Wedgwood and Talbot. In 1802 Thomas Wedgwood, son of the famous potter Josiah Wedgwood, reported his experiments in recording images on paper or leather sensitized with silver nitrate. Although he could record silhouettes of objects placed on the paper, he was not able to make them permanent and, to his disappointment, he failed to record a camera image.

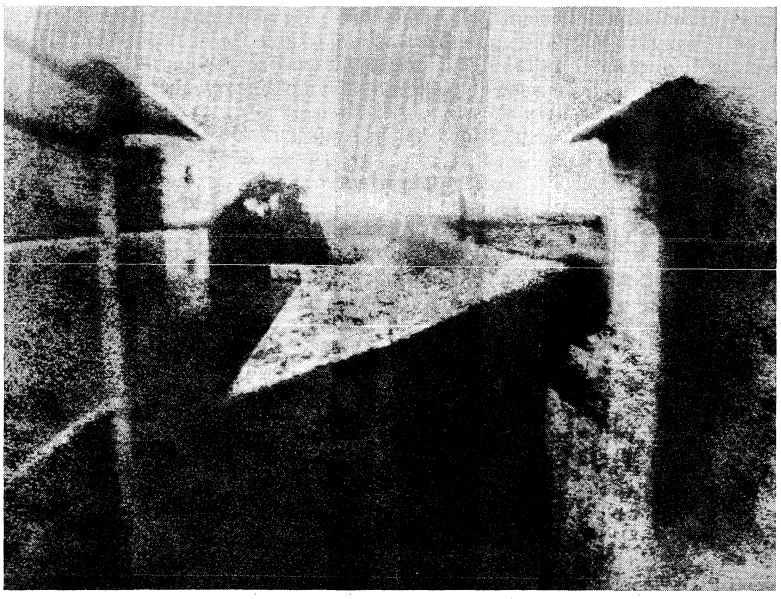
Unaware of the work of Wedgwood and the French pioneers, Talbot, trained as a scientist at Cambridge University, was led to invent a photographic process because of his inability to draw landscapes. On a holiday trip to Italy in 1833, the idea came to him of recording by chemical means the images he observed in his camera obscura. By 1835 he had a workable technique: he made paper light sensitive by soaking it alternately in solutions of common salt (sodium chloride) and silver nitrate. Silver chloride was thus produced in the fibres of the paper. On exposure to light the silver chloride became finely divided silver, dark in tone. He decreased the sensitivity of the unexposed silver chloride by washing it in a strong solution of sodium chloride. The result was a negative, from which he could make any number of positives simply by putting fresh sensitized paper in contact with the negative and exposing it to light. Later, upon the suggestion of the scientist Sir John F.W. Herschel, he fixed his photographs with sodium hyposulphite (now called sodium thiosulfate).

When news of Daguerre's process reached England in January 1839, Talbot rushed publication of his "photogenic drawing" process and subsequently explained his technique in full detail to the members of the Royal Society—six months before the French government divulged working directions for the daguerreotype. There were many others who had similar techniques and who were to claim priority, but to Talbot and Daguerre are owed the two basic processes that were to sweep across the civilized world and establish photography as the most facile and convincing way to produce pictures.

Differences between the two pioneer processes

First criticism. The two pioneer processes were different in several ways. Daguerreotypes were on metal; photogenic drawings were on paper. Each daguerreotype was unique; photogenic drawings could be duplicated endlessly. The aesthetic as well as physical character differed markedly. The daguerreotype rendered detail to a degree that was remarkable; the photogenic drawing, because of the fibrous structure of the paper supporting the silver image, gave a broader, somewhat diffused effect.

The first criticism of photography was necessarily based on a comparison with painting or drawing, since no other standards of picture making existed. The Parisian painter Paul Delaroche, when questioned about the value of the daguerreotype to artists, went so far as to say that "the process completely satisfies art's every need." Photography's remarkable ability to record a seemingly inexhaustible amount of detail was marvelled at again and again. The critics regretted that, because of the great length of exposure, moving objects were not recorded or were ren-



First successful photograph from nature, heliograph on pewter by Joseph-Nicéphore Niépce, 1826.
By courtesy of the Gernsheim Collection, The University of Texas at Austin

dered blurry and indistinct. The inability of the first processes to record colours was disappointing, but since the critics were already conditioned to black-and-white prints and drawings, this was not as serious a drawback as the harshness of the tonal scale. The technique of photography was at once recognized as a shortcut to art. No longer was it necessary to spend years in art school drawing from sculpture and from life, mastering the laws of linear perspective and chiaroscuro. As Daguerre boasted in a broadsheet in 1838, "with this technique, without any knowledge of chemistry or physics, one will be able to make in a few minutes the most detailed views . . ."

DEVELOPMENTS TO WORLD WAR I

Spread of the daguerreotype. Daguerre's process rapidly spread throughout the world. Before the end of 1839, travellers were bringing back to Paris daguerreotypes of famous monuments in Egypt, Israel, Greece, and Spain; from them engravings were made that were published in two volumes as *Excursions daguerriennes* between 1841 and 1843. Although his process was published "free to the world" by the French government, Daguerre took out a patent for it in England; the first licensee was Antoine-François-Jean Claudet. The first daguerreotypes in America were made on September 16, 1839, just four weeks after the announcement of the process. Exposures were at first of excessive length—a daguerreotype of King's Chapel, Boston, in the International Museum of Photography, collection in Rochester, New York, bears a label stating that it was made between 4:40 and 5:30 PM on April 19, 1840. At such exposures moving objects could not be recorded, and portraiture was impractical. Even in blazing sunlight and with the face whitened by flour, a person had to sit immobile for minutes on end. Within a few years, however, such improvements had been made in lenses and sensitizing techniques that exposure times were reduced to approximately 20 to 40 seconds. At once daguerreotyping became a flourishing industry. It was generally conceded that the United States led the world in the production of daguerreotypes. In the late 1840s every city had its "daguerrian artist," and villages and towns were served by travelling photographers who had fitted up wagons as studios. In New York City alone there were 77 galleries in 1850. Of these, the most celebrated was that of Mathew B. Brady, who began in 1844 to form a "Gallery of Illustrious Americans," and to that end collected portraits of notables taken by his own and other cameramen. Twelve of the portraits were published by lithography in a folio volume. In Boston the firm of Albert Sands Southworth and Josiah Johnson Hawes produced the finest portraits ever made by the daguerreotype process. They avoided the stereotyped lighting and posing formulas of the average daguerreotypist and did not hesitate to portray their sitters unprettified and "as they were." Lemuel Shaw, a judge of the Supreme Court of

Early daguerrotypes in America

By courtesy of the Metropolitan Museum of Art, New York, gift of Edward S. Hawes, Alice Mary Hawes, Marion A. Hawes, 1938



"Chief Justice Lemuel Shaw," daguerreotype by Albert Sands Southworth and Josiah Johnson Hawes, 1851.

Massachusetts, stands with crumpled coat and unruly locks of hair under a glare of sunshine; Lola Montez, adventuress, dancer, actress, lolls over the back of a chair, a cigarette between her gloved fingers. Cities and towns, as well as their inhabitants, were photographed by American daguerreotypists: the rapid growth of San Francisco was documented month by month and the first history of the city, published in 1855, was illustrated by engravings made from daguerreotypes.

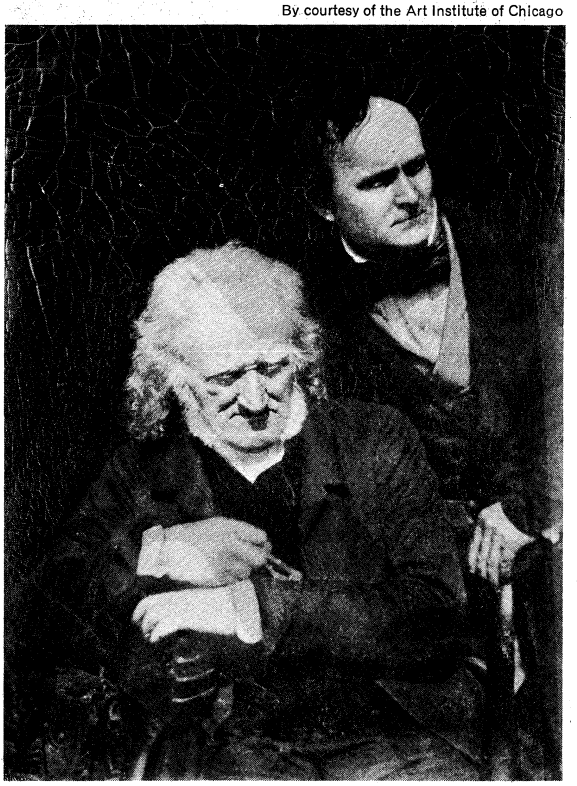
Development and use of the calotype process. By 1840 Talbot had greatly improved his photogenic drawing process by increasing the sensitivity of his prepared paper. At

first this silver chloride paper was allowed to remain in the camera (or beneath a negative) a matter of minutes, until a visible image appeared. Talbot then discovered that a much shorter exposure, which he estimated at 1/60 of the previous exposure, was sufficient to create an invisible (latent) image that could later be changed into a visible image by a chemical process now known as development. The discovery of the theory of the latent image was Talbot's great contribution. He named his improved negative process the calotype, from the Greek meaning "beautiful picture."

Discovery of the latent image

The first and most aesthetically satisfying use made of this improved process was in the work of David Octavius Hill, a Scottish landscape painter, and his partner, Robert Adamson, an Edinburgh photographer. In 1843 Hill decided to paint a group portrait of the ministers who in that year formed the Free Church of Scotland. There were over 400 figures to be painted. Sir David Brewster, who knew of Talbot's process from the inventor himself, suggested to Hill that he make use of this new technique. Hill then enlisted the aid of Adamson, and together they made hundreds of photographs, not only of the members of the church meeting but also of people from all walks of life. Although their sitters were posed outdoors in glaring sunlight and had to endure exposures of upward of a minute, Hill and Adamson managed to retain spontaneity. Hill's vision was dominated by the painting style of the period in lighting and posing, particularly in the placement of the hands. Many of the calotypes are strikingly reminiscent of canvases by Sir Henry Raeburn and other contemporary artists. Indeed, William Etty, a Royal Academician, copied in oils the calotype Hill and Adamson made of him in 1844 and exhibited it as a self-portrait. In addition to their formal portraiture, the partners made a series of photographs of fishermen and their wives at Newhaven, in Edinburgh, and architectural studies.

By courtesy of the Art Institute of Chicago



"John Henning and Alexander Handyside Ritchie," calotype by David Octavius Hill and Robert Adamson, c. 1845.

Calotypes of architectural monuments

The potential of the calotype for recording great monuments of architecture was shown by a number of Frenchmen, many of whom were trained as painters. In the 1850s they began to photograph historical buildings for the government. Working with cameras making photographs as large as 20 by 29 inches (51 by 74 centimetres), Henri Le Secq, Charles Marville, and Charles Nègre produced remarkable calotypes of the cathedrals of Notre-Dame in Paris, Chartres, and Amiens, as well as other structures that were being restored after centuries of neglect. An establishment was set up in Lille, France, by Louis-Désiré Blanquart-Evrard for bulk printing these paper negatives. Among the products of this firm was a superb volume of photographs by the Parisian writer Maxime Du Camp taken during his travels with the writer Gustave Flaubert in Egypt, Israel, and Syria, from 1849 to 1851.

Development and use of the collodion process. Photography was revolutionized in 1851 by the introduction of the collodion process for making glass negatives. This new technique, invented by the English sculptor Frederick Scott Archer, combined the advantages of both the daguerreotype and the calotype processes: the negatives were of glass and recorded detail in a way that rivalled the daguerreotype, and from them quantities of paper prints could be made. The process almost at once became universal, and the popularity of photography increased. The process had one serious drawback: the photographer had to sensitize his plate almost immediately before exposure and expose it and process it while the emulsion was moist. Collodion is a solution of nitrocellulose (gun cotton) in alcohol and ether; when the solvents evaporate, a clear plastic-like film is formed. Since it is then impervious to water, the chemicals used for developing the exposed silver halides and removing the unexposed salts cannot penetrate to them. Despite the drawback that the photographer had to have a darkroom of sorts always available, some of the most memorable photographs of all time were produced by this wet-plate process.

Curiously, in the United States, the new collodion plates were at first used to produce imitation daguerreotypes called ambrotypes. They were simply negatives with a dark backing to make the clear areas of shadows appear dark. In posing and lighting, these popular portraits were identical to daguerreotypes; they were of the same standard sizes, and they were enclosed in the same type of case. They did not approach the brilliancy of the daguerreotype, however. Tintypes, first known as ferrotypes or melainotypes, were merely variations of the ambrotype. Instead of glass the collodion emulsion was coated on thin iron sheets enamelled black. At first they were presented in cases, surrounded by narrow gilt frames, but by the 1860s this elaborate presentation had been abandoned, and the metal sheets were simply inserted in paper envelopes, each with a cutout window the size of the image. Easy to make, inexpensive to purchase, tintypes remained a kind of folk art through the 19th century. Poses were often informal, if not humorous.

Development of the tintype

Portraiture. A new style of portrait, introduced in Paris by André-Adolphe-Eugène Disdéri in 1854, came to be called the *carte-de-visite* because the size of the mounted photograph (four by 2½ inches [ten by six centimetres]) corresponded to that of a calling card. Disdéri used a multilens camera to produce eight negatives on a single glass plate. Each picture could be separately posed, or several exposures could be made at once. The principal advantage of the system was its economy: to make eight portraits the photographer needed to sensitize only a single sheet of glass and make one print, which he then cut up into separate pictures. At first *cartes-de-visite* almost invariably showed the subjects standing. Backgrounds, almost invariably plain in the days of the daguerreotype, became ornate: furniture and such architectural fragments as papier-mâché columns and arches were introduced, and heavy-fringed velvet drapes were hung within range of the camera. With the advent of the cabinet-size (6½ by four inches [17 by ten centimetres]) picture in 1866 the baroque tendencies of the photographer became yet more audacious, so that in 1871 a photographer wrote: "One good, plain background, disrobed of castles, piazzas, columns, curtains and what not, well worked, will suit every condition of life." It was at this period that retouching was introduced, as was the practice of painting over the photograph in oil colours.

Contributions of Nadar, Carjat, and Cameron

In contrast to the excessive reliance on accessories and retouching shown by the popular portrait photographers of Europe and America, the work of two Frenchmen and one Englishwoman stand apart. In their portraiture they reached a level unsurpassed since the daguerreotypes of Southworth and Hawes and the calotypes of Hill and Adamson. These photographers were Gaspard-Félix Tournachon, a Parisian writer, editor, and caricaturist who used the pseudonym of Nadar; Étienne Carjat, likewise a Parisian caricaturist; and Julia Margaret Cameron, wife of a British civil servant.

Nadar took up photography in 1853 as a means of making studies of the features of prominent Frenchmen for inclusion in a large caricature lithograph, the "Panthéon Nadar." He posed his sitters against plain backgrounds and bathed them with diffused light, which brought out every detail of face and dress. He knew most of them, and the powers of observation he had developed as a caricaturist led him to recognize their salient features, which he recorded directly, without the exaggeration that he put in his drawings. When Nadar's photographs were first exhibited, they won great praise in the *Gazette des Beaux Arts*, then the leading art magazine in Europe. Nadar was a colourful man. He had a passion for balloons, and in 1855 he combined his interests by taking the world's first aerial photograph.

Carjat's portraits, less intense perhaps than Nadar's, do have the dignity and distinction of those of his contemporary and rival.

Mrs. Cameron took up photography as a pastime in 1864. Awkward as it was, she used the wet-plate process and began to take portraits of such celebrated Victorians of her acquaintance as Alfred, Lord Tennyson, Thomas Carlyle, Charles Darwin, and Sir John F.W. Herschel. A number of her portraits were shown at the Paris International Exhibition of 1867. Mrs. Cameron used a lens of the extreme focal length of 30 inches (76 centimetres) to obtain a large head size. This lens required such long exposures that the subjects frequently moved. The lack of optical definition plus this accidental blurring was universally criticized, yet the very power of her work won her international praise. This can only be explained by the intensity of her vision. "When I have had these men before my camera," she wrote about her portraits of great men,

> my whole soul has endeavoured to do its duty toward them in recording faithfully the greatness of the inner man as well as the features of the outer man. The photograph thus obtained has almost been the embodiment of a prayer.

Besides these memorable portraits, Mrs. Cameron produced a large number of allegorical studies, of children and young women in costume, acting out biblical scenes or themes based on the poetry of her hero, Tennyson. In making these pictures—which today seem weak and sentimental—she was much influenced by the Pre-Raphaelite painters and the photographic work of Oscar G. Rejlander and Henry Peach Robinson (see below), leaders in the production of photographs that emulated paintings.

Influence of painting. Consideration of photography as an aesthetic medium was given impetus by the formation of photographic societies, made up of both professionals and amateurs, who had been attracted to the camera by the popularity of the collodion process. In 1853 the Photographic Society, parent of the present Royal Photographic Society, was formed in London, and in the following year the Société Française de Photographie was founded in Paris.

At the first meeting of the Photographic Society the president, Sir Charles Eastlake (who was then also president of the Royal Academy) invited the miniature painter, Sir William Newton, to read a paper "Upon Photography in an Artistic View" (*Journal of the Photographic Society*, i, 1853). His argument was that photographs could be useful to the painter so long as they were taken "in accordance (as far as it is possible) with the acknowledged principles of Fine Art." One way by which the photographer could make his results more like works of art, Sir William pointed out, was to throw the subject slightly out of focus. The extreme blue-sensitivity of wet collodion plates rendered the sky areas so dense that they appeared as blank paper on the print. To overcome this failure of the photographic process he recommended liberal retouching, the free use of handwork on the negative.



"The Mountain Nymph, Sweet Liberty" by Julia Margaret Cameron, 1867.
George Eastman House Collection

Combination printing

An outcome of the urge to create photographs that would fit a priori concepts of what "art" should be was the technique of making one print from several negatives. A famous example was by Oscar G. Rejlander, a Swede who had studied art in Rome and was practicing photography in England and used 30 negatives to produce a 31- by 16-inch (79- by 41-centimetre) print titled "The Two Ways of Life," showing, in allegory as obvious as it was sentimental, that the way of the blessed led through good works and the way of the damned through vice. Rejlander, who described the technique in detail in photographic journals, stated that his purpose was to prove to artists the aesthetic possibilities of photography, which they had generally denied. The photograph was shown in the Manchester Art Treasures Exhibition of 1857 and was purchased by Queen Victoria for Prince Albert.

Rejlander's technique stimulated Henry Peach Robinson, a professional photographer who had been trained as an artist, to produce similar combination prints. He achieved fame with a five-negative print, "Fading Away," produced in 1858. The subject, a dying girl, was considered by critics as too painful a subject to be represented by photography. Perhaps the implied authenticity of the camera bothered them, for painters had long presented subjects of far more painful nature.

Robinson became a most articulate member of the Photographic Society, and his teaching was even more influential than his photography. In 1869 appeared the first of many editions and translations of his book, *Pictorial Effect in Photography*. Robinson went a step further than Newton; he offered photographers a set of rules. From an outmoded handbook on painting, he borrowed compositional formulas the use of which, he claimed, would bring artistic success. The student was instructed to arrange his models in pyramidal compositions, and examples from the work of Sir David Wilkie, a popular painter of genre scenes, were given to show how this artist used triangular

Robinson's rules for photography



"The Two Ways of Life," a composite photograph made from 30 negatives taken and assembled by Oscar G. Rejlander, 1857. The tableau is allegorical, showing a young man torn between a life of debauchery (left) and virtue (right).
George Eastman House Collection

compositional schemes. The importance of balance was stressed, and the opposition of light against dark was made clear. The fault of the book lay not only in the assumption that rules set up for one art form can be applied to another but also in its intellectual and academic approach to art.

Robinson's work is weak and artificial by present standards of taste. Not only did he practice combination printing—itself a denial of photography's power of conviction—but he preferred to work in the studio, against painted backdrops and with props of natural objects, even foliage, mounted on casters. When he did photograph the real world, he took models with him, dressing them up to play the part of country girls.

So long as photographers maintained that the way to art was by the emulation of painting, critics were reluctant to admit the new medium to an independent aesthetic position. Portraits, when done as sensitively and as directly as those produced by Hill and Adamson, Nadar, and Mrs. Cameron, won the praise of art critics. But sentimental genre scenes, posed and arranged for the camera, lacked that sharp objective truth that is a characteristic of photography. Other photographers, less concerned about artistic issues, were making photographs of the world and man's activities upon it with such extraordinary perception and understanding of the medium that often their work surpassed record making. These men took their cameras to battlefields and to faraway places, often at the risk of their lives.

Combat photography. In 1855 Roger Fenton went from London to the Crimea to photograph the war. He developed his wet plates in a darkroom on wheels, at times under enemy fire. When the Civil War broke out in America, Mathew Brady, a New York daguerreotypist who had been among the first to adopt the wet-plate process, conceived the bold plan of making a photographic record of the hostilities. The *New York Times* on July 21, 1862, commented that Brady was

> the first to make Photography the Clio of the war. . . . His artists have accompanied the army in nearly all its marches, planting their sun batteries by the side of our General's more deathful ones, and "taking" towns and cities, forts and redans, with much less noise, and vastly more expedition. The result is a series of pictures christened "Incidents of the War," and nearly as interesting as the war itself; for they constitute the history of it, and appeal directly to the great throbbing hearts of the North.

Brady's "Incidents of the War"

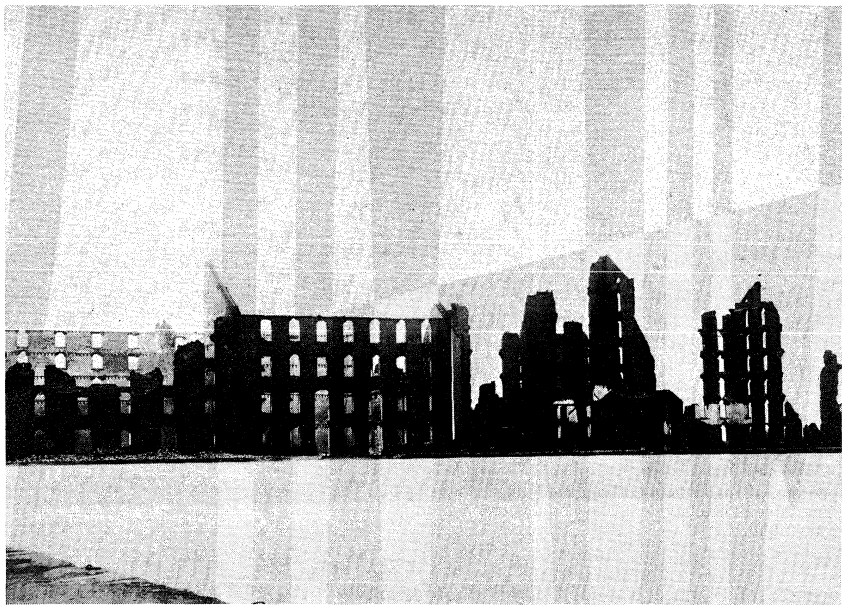
Brady and his photographers—notably Alexander Gardner and Timothy H. O'Sullivan, who left his employ in the midst of hostilities—produced an amazing record of the battlefield. At his New York gallery, Brady showed pictures of the dead at Antietam. The *New York Times* reported on October 20, 1862:

> Mr. Brady has done something to bring home to us the terrible reality and earnestness of war. If he has not brought bodies and laid them on our dooryards and along the streets, he has done something very like it. . . . It seems somewhat singular that the same sun that looked down on the faces of the slain blistering them, blotting out from the

George Eastman House Collection



"Fading Away," a composite photograph made from five negatives by Henry Peach Robinson, 1858.



"Ruins of the Gallego Flour Mills, Richmond, Virginia" by Mathew B. Brady, 1865.
By courtesy of the Museum of Modern Art, New York

bodies all semblance to humanity, and hastening corruption, should have thus caught their features upon canvas, and given them perpetuity for ever. But it is so.

Oliver Wendell Holmes, the American physician and writer, was a great lover of photography and appreciated its specific aesthetic qualities. Of the Civil War photographs he wrote:

> It was so nearly like visiting the battlefield to look over these views, that all the emotions excited by the actual sight of the stained and sordid scene, strewed with rags and wrecks, came back to us, and we buried them in the recesses of our cabinet as we would have buried the mutilated remains of the dead they too vividly represented. (*The Atlantic Monthly*, July 1863.)

To Holmes, the power of the photograph to convince, to carry a sense of authenticity, set it apart from other mediums.

> The very things which an artist would leave out, or render imperfectly, the photograph takes infinite care with, and so renders its illusions perfect. What is the picture of a drum without the marks on its head where the beating of the sticks has darkened the parchment? (*The Atlantic Monthly*, June 1859.)

The camera records the very texture of the scene. The strange becomes familiar. The viewer knows the roughness of the boulders in the memorable photograph "'Home of a Rebel Sharpshooter" by Alexander Gardner. The soldier's gleaming rifle, his crumpled uniform, his pale face—these can be grasped in an instant, without a shadow of doubt. One is transported to the scene, unconsciously convinced of the accuracy of the picture. The response is utterly different from that provoked by a painting or a drawing. For a moment the viewer sees through the eyes of the cameraman.

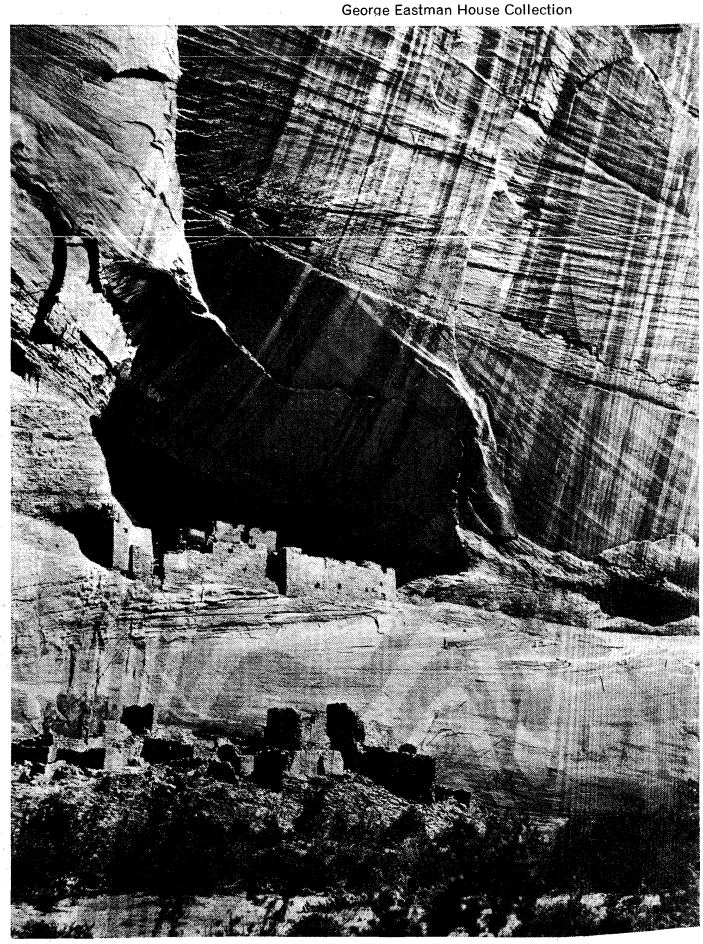
Long prized for their value as historical documents, the Civil War photographs are now valued for their aesthetic qualities as well. The change came about through a recognition that photography had its own particular aesthetic and that photographs should be seen directly, without comparing them to paintings or to drawings.

Field photography. The last half of the 19th century saw the rapid rise of transportation and travel. The Suez Canal and fast ships brought colonies in the Southern Hemisphere closer to Europe; the transcontinental railroad joined the East and West coasts of the United States, and vast regions beyond the Mississippi became more accessible. Photographers were important members of government surveys and were also commissioned by railroad companies to make publicity pictures of track laying, bridge building, and spectacular scenery through which the new lines ran.

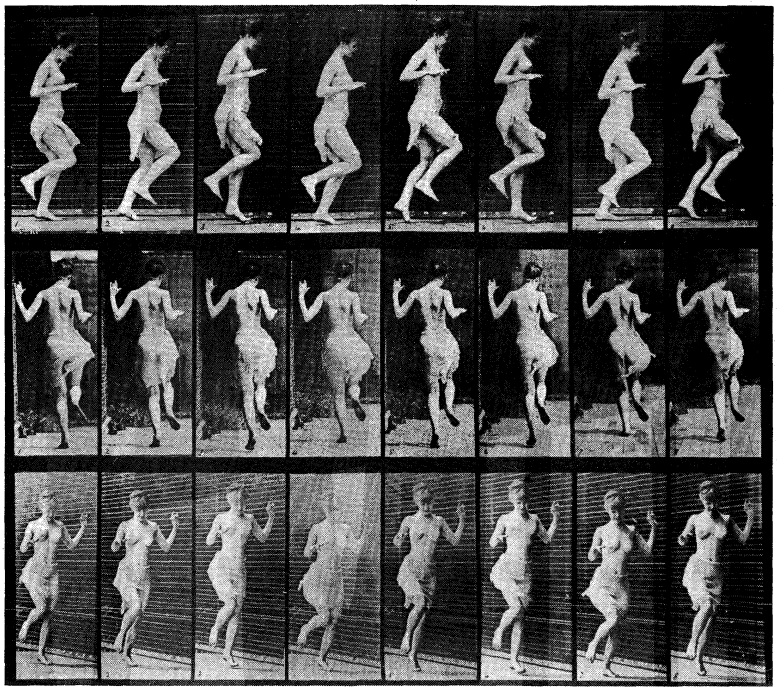
Photography on the American frontier

Of the photographers of the American frontier, two stand out: Timothy H. O'Sullivan, of Civil War fame, and William Henry Jackson. O'Sullivan's photographs of the Southwest are of great beauty, particularly his views of Indian cliff dwellings in the Canyon de Chelly, Arizona, made in 1873. Jackson, self-trained as a painter in Vermont, crossed the plains as a wagon driver. In 1868 he opened a photographic gallery in Omaha, Nebraska. The sedentary life of a portraitist and colourist did not please him, and he began to go into the field to photograph

George Eastman House Collection



"Canyon de Chelly, Arizona" by Timothy H. O'Sullivan, 1873.



"Figure Hopping" by Eadweard Muybridge, 1887. Sequence of eight stages of movement, simultaneously photographed by multiple cameras at three different positions.
By courtesy of the Cooper-Hewitt Museum of Decorative Arts and Design, Smithsonian Institution

Indians. The Union Pacific Railroad was under construction, and he received an order to produce 10,000 stereographs. The excellence of his work led F.V. Hayden, a geologist, to hire him to photograph the Yellowstone as part of Hayden's government-financed expedition there in 1871. The photographs Jackson took were influential in the decision by Congress to create the Yellowstone National Park. Later, in 1875, he recorded the immensity of the western landscape, using glass plates as large as 20 by 24 inches (51 by 61 centimetres). "They convey," he wrote in the catalog, "an impression of the real grandeur and the magnitude of mountain scenery that the smaller views cannot possibly impart."

The far corners of the world were reached by intrepid photographers. Francis Frith, an Englishman, exposed negatives in the Egyptian desert of a size rivalling Jackson's; Felice Beato, another Englishman, went to India, China, and Japan; and the Americans John Dunmore and Critcherton went to Labrador, where they made large photographs of icebergs and Eskimo life.

Stereoscopic photography. All of these field photographers also took stereographs. These double pictures, taken with twin-lens cameras, produce a remarkable effect of three dimensions when viewed through a stereoscope. First described in 1832 by an English physicist, Charles Wheatstone, the stereoscope did not become practical until Sir David Brewster, in Scotland, designed a simplified instrument for viewing them, which was exhibited at the 1851 Great Exhibition in the Crystal Palace, London. Queen Victoria was entranced by the stereographs she saw there, and instantly three-dimensional photography became a popular craze. Stereography is uniquely photographic, since no artist could draw two scenes in exact perspective from viewpoints separated only 2½ inches (six centimetres)—the normal distance between human eyes.

The illusion of looking at reality in a well-photographed stereoscopic view is remarkable. The aesthetics of two-dimensional picture making cannot be applied to these views. Indeed, many art critics hold that stereography is so rooted in illusion that such pictures can never be of lasting value. The millions of stereographs that were circulated in the years before newspaper reproduction of photographs did, however, have an enormous impact.

Instantaneous photography and the dry plate. The technology of photography took an important step forward as a result of the development of the dry plate coated with gelatin containing silver salts. By 1880 it was no longer necessary for the photographer to sensitize his own plates: he could buy them ready-made and could process them long after exposure. There was a further advantage: gelatin plates were about 60 times more sensitive than collodion plates. The simplicity and flexibility of the new process not only attracted an ever-increasing number of people to photography, but it led to new aesthetic considerations.

The development of the dry plate

Earlier, collodion plates had been used to make instantaneous photographs. Street scenes in London, Paris, and New York showed pedestrians and vehicles stopped in their tracks. Oliver Wendell Holmes found these records invaluable in his study of the human gait—a field in which as a physician he was deeply concerned. He noted that the camera shows the attitudes of legs and arms in ways different from the visual impression of them. "No artist," he wrote in 1863, "would have dared to draw a walking figure in attitudes like some of these."

Fifteen years later, the artistic world was amazed at the photographs of horses taken by Eadweard Muybridge in California. Using a series of cameras ranged side by side opposite a reflecting screen, with their shutters released by the breaking of threads as the horse dashed by, Muybridge secured sets of sequence photographs of successive phases of the gallop, the trot, the pace, and the walk.

In 1878, when the pictures were published internationally in the popular and scientific press, it was difficult to believe that they were accurate. To prove that his photographs were correct, Muybridge threw them upon a screen one after the other with a lantern-slide projector he had built for the purpose; the result was the world's first motion-picture presentation. This memorable event took place at the San Francisco Art Association in 1880.

These famous photographs were taken with wet plates. Later, with the new gelatin plates, Muybridge was able to improve his technique greatly, and, at the invitation of the University of Pennsylvania, he produced 781 sequence photographs of many kinds of animals as well as men and women, both clothed and nude, engaged in a wide variety of activities.

Growth in amateur photography. With the increased speed of negative materials, the camera was freed from the tripod. This not only simplified picture taking but was to open up a wholly new vision. With the hand camera



"Pond in Winter" by Peter Henry Emerson, 1888.
George Eastman House Collection

the very flux and flow of life could be captured, and an aesthetic based upon this capability was to be developed; but the first impact of the new, simplified gelatin process was its role in the growth of amateur photography. A great variety of hand cameras became available at relatively low cost; photography became available to all. Box cameras predominated. At first they were called "detective cameras" because they were inconspicuous and pictures could be taken with them without attracting the subject's attention.

Invention of the Kodak camera

Of these, the most famous was the Kodak camera, invented by the American George Eastman in 1888. In place of glass plates, it contained a roll of negative material sufficient for taking 100 pictures, each 2½ inches (six centimetres) in diameter. After exposing the last negative, the entire camera was sent to one of the Eastman factories (Rochester, New York, or Harrow, Middlesex), where the roll was processed and printed. "You Press the Button, We Do the Rest" was Eastman's description of the Kodak system. At first Eastman's so-called "American film" was used in the camera. This film was paper based, and the gelatin layer containing the image was stripped away after development and fixing and transferred to a transparent support. In 1889 this troublesome manipulation was eliminated with the introduction of film on a transparent plastic base of nitrocellulose. With the Kodak camera and the myriad of other similar cameras that followed in its wake, photographs of the most casual sort were taken by the million.

Concern with photography as an art form. The snapshot era made everybody an artist—or so thought the multitude. But art is not so lightly won, and those amateurs who were seriously concerned with photography as an art banded together in societies. At a meeting of the Camera Club of London in 1886, Peter Henry Emerson, physician by profession and an ardent amateur photographer, gave a lecture on "Photography, a Pictorial Art," in which he attacked the artificiality of the photographs of Henry Peach Robinson, then still highly popular, and quite generally accepted by the public as outstanding examples of the artistic use of the camera. *The Illustrated London News*, for example, featured full page wood-engravings of Robinson's best picture of the year. Despite his failings, Emerson's vehement and passionate plea for the return to natural subjects was indeed salutary. Of greater importance was his advice that photographers should respect the photographic process, and limit their controls to those that were inherent.

In his book, *Naturalistic Photography for Students of the Art* (1889), he further developed his theories. He believed that photography was potentially a great creative art, and that it did not need to be bound to rules set up for painting. "*Naturalism*," he wrote,

> is an *impersonal* method of expression, a more or less correct *reflection* of nature wherein (1) truth of sentiment, (2) illusion of truth of appearance (so far as is possible) and (3) decoration are of first and supreme importance.

Emerson sought a scientific basis for pictorial photography. He analyzed the physiology of vision in order to find out how the camera can simulate human perception. He reasoned that since the eye sees distinctly only a small part of what lies before it, while the surrounding area appears indistinct, it should be possible to reproduce this phenomenon by control of focussing. Only a small part of the photograph, he argued, should be sharp; the rest should be slightly out of focus. The fallacy in this theory of vision is that the eye, unlike the lens of a camera, is a living organism, which scans the field of view constantly so that the mind receives a composite report and the impression of full detail. Emerson's naturalistic theory was an advance in that he believed the photographer should rely entirely on controls that were themselves part of the photographic process; in addition to control through differential focussing, he advised the careful use of developing and printing methods.

Emerson's photographs of natural subjects

Emerson took a large number of photographs of the life of the simple country folk on the Norfolk Broads. These photographs were far removed from the artificial genre studies of Robinson and close to the graphic work of the French painter J.-F. Millet, which Emerson greatly admired. Imbued with a deep artistic feeling, Emerson's photographs, published in limited editions in handsome folio volumes, inspired other amateurs to seek inspiration in nature.

If Emerson was not the first to use the newly invented platinotype, he was the first to popularize its use for artistic ends. In this printing paper, the salts of iron and platinum replaced those of silver as the light-sensitive material. Platinotypes had a long and delicate tonal scale, and the choice of paper surface ranged from smooth to the rough texture of watercolour paper.

In 1891 Emerson renounced the art claims he had made for photography. A great artist, he said, had shown him that the mere reproduction or translation of nature was not art. The fallacy of this argument lay, of course, in the definition of art: to Emerson the words "art" and "paint-



"The Big White Cloud, Lake George, New York" by Edward Steichen, 1903.
By courtesy of the Museum of Modern Art, New York

ing" were almost synonymous. He dramatically published his retraction in a black-bordered pamphlet, "The Death of Naturalistic Photography."

In his writings, Emerson had much to say about photographic exhibitions. He felt that they were poorly organized. He recommended that photographs be framed in simple white molding and hung on one line, not plastered over gallery walls with little or no space between them. Especially he felt that photographs produced as art objects should be shown apart from work done for scientific, technical, and other nonaesthetic purposes.

Dissatisfaction with the scientific bias of the Photographic Society under the presidency of Sir William Abney caused a split in the society in 1892, when a group of pictorially minded members, under the leadership of H.P. Robinson and George Davison, formed a new society for the exclusive purpose of furthering "the development of the highest form of art of which photography is capable." They named the society the Linked Ring and held annual exhibitions, for which they chose the word salon, to show their artistic allegiance. By 1901 it was their proud boast that "through the *Salon* the Linked Ring has clearly demonstrated that pictorial photography is able to stand alone and that it has a future entirely apart from that which is purely mechanical."

In 1887 Emerson, judging a competition for *The Amateur Photographer*, had awarded a prize to Alfred Stieglitz, a young American who was studying photography in Germany. It was the first award that Stieglitz had received, and the first recognition of one who was to carry pictorial photography to a high level and be the constant champion of all that was progressive in photography and other forms of modern art.

Stieglitz's contributions to photography

When Stieglitz returned to the United States in 1890 after his student years in Berlin, he found that photography was a passing craze and that the New York Society of Amateur Photographers had threatened to change their organization into a bicycling club. After serving as editor of the *American Amateur Photographer* for a number of years, Stieglitz organized a sympathetic group in the Camera Club, New York City, and founded a handsome periodical, *Camera Notes and Proceedings*, which was outstanding for the unusually high quality of its reproductions of work not only by the members but by leaders in the pictorial movement in Europe and America. Stieglitz was one of the first to demonstrate the pictorial possibilities of the hand camera with his photographs of New York under all weather conditions. In an article, "The Hand Camera—Its Present Importance," after pointing out the dangers of casual snapshooting, he advised the reader to

> choose your subject, regardless of figures, and carefully study the lines and lighting. After having determined upon these watch the passing figures and await the moment in which everything is in balance; that is, satisfies your eye. This often means hours of patient waiting. My picture, "Fifth Avenue, Winter," is the result of a three hours' stand during a fierce snow-storm on February 22d, 1893, awaiting the proper moment.

Stieglitz championed many photographers who shared his conviction in the artistic capabilities of the camera. Three stand out: Edward Steichen, Clarence H. White, and Gertrude Käsebier. Steichen, trained as a painter, took up photography first as an aid in making drawings for the lithographic firm that employed him. From photographic magazines, he saw the creative possibilities of the camera and sent pictures to the Philadelphia Salon in 1899, where they instantly won praise. Passing through New York on his way to Paris in the following year, Steichen visited Stieglitz, who bought some of his photographs and encouraged him to continue with the camera as well as with brush and canvas. The first large showing of his photographs was in London, in the "New American School" exhibition of 1900. Steichen's soft-focus landscapes and portraits, reminiscent of Whistler in their use of flat tones and bold cutting of edges, were very controversial. In Paris he met the sculptor Auguste Rodin and produced a series of portraits of him and interpretive photographs of his work. He mastered the newly introduced gum print process. This technique gave the photographer the utmost manual control. He coated paper with watercolour pigment of any desired tint mixed with gum arabic and potassium bichromate. On exposure to light beneath a negative, the pigment became insoluble according to the amount of light received. The print was "developed" simply by bathing it in water. If desired, areas could be eliminated by brushing them with hot water.

Steichen's gum prints

Clarence H. White, of Newark, Ohio, and later New York, had a delicate sense of light. He preferred the platinotype process for printing, which gave delicate tonal rendering.

Gertrude H. Käsebier was a professional portraitist

who used unusual poses and had an especially fine feeling for children.

Formation of the Photo-Secession

In 1902 Stieglitz, with Steichen, White, Käsebier, and other friends, disturbed by the lack of standards and the general conduct of photographic exhibitions in the United States, formed a society, the Photo-Secession. They chose the word "secession" to show their split with the status quo, just as avant-garde German and Austrian painters had used the same word to make manifest their independence of officialdom. The record of the Photo-Secession is contained in 50 issues of *Camera Work*, published by Stieglitz between 1903 and 1917; this quarterly publication contained superb reproductions of photographs and extensive reviews and articles. The Photo-Secession had a gallery, which came to be known as 291 from the street number on Fifth Avenue, New York City, where it was located. There Stieglitz showed not only pictorial photographs but also modern art selected at first by Steichen in Paris. At 291 Americans saw, long before the Armory Show of 1913 made them popular, lithographs by Cézanne, drawings by Picasso, and sculpture by Matisse. If critics were amazed that Stieglitz, the champion of photography, was showing avant-garde modern art, he would reply that 291 was an experimental gallery, where photography would take its natural place along with the more traditional forms of picture making.

By courtesy of the Art Institute of Chicago, Alfred Stieglitz Collection



"The Steerage" by Alfred Stieglitz, 1907.

In 1915 Stieglitz and his friends published the periodical *291*, which printed a handsome 11- by 14-inch photogravure of his photograph "The Steerage," with essays praising Stieglitz for his triumph over the machine.

The last two issues of *Camera Work* contained photographs by Paul Strand, which showed a new approach. There were views looking down from unusual angles; there were bowls in quasi-abstract arrangements; there was a group of powerful open-air portraits taken in the streets of the lower East Side of New York. These portraits had been taken with a 45° prism fitted over the lens, so that the subjects were unaware that they were being photographed. Alvin Langdon Coburn, one of the first of the Photo-Secessionists, made a series of photographs in 1912 looking down from tall buildings, which he exhibited as "New York from Its Pinnacles"; they were remarkable for the way in which emphasis was placed on form. He pushed this interest in abstraction to the total elimination of recognizable subject matter in his "Vortographs," some of which were published in *Photograms of the Year* in 1917.

DEVELOPMENTS SINCE WORLD WAR I

The new vision of the '20s. At the close of World War I, Steichen, who had been in command of aerial photography for the American Expeditionary Forces, abandoned the broad, quasi-impressionistic style for which he was so famous for sharply focussed photographs taken with lighting more familiar to the stage than to the portrait studio. His portraits of celebrities, taken for *Vanity Fair* and *Vogue* magazines, are a monumental record of the actors and actresses, artists and writers, thinkers and teachers of the '20s and '30s.

Steichen's new style

The painter Charles Sheeler used the camera to record the stark beauty of Pennsylvania barns and the forms of industrial structures.

To Stieglitz this revaluation of photographic aesthetics was in fact a strengthening of the beliefs he had held since his student days. He was to produce his most powerful work in the 1920s—his series of portraits of his associates at the galleries that succeeded 291 after 1917, when the building in which it was located was torn down; his views of New York, taken from the windows of his apartment or gallery; and especially photographs of cloud forms that he called "Equivalents" because through them he could express his intense emotion. Ill health forced him to abandon the camera in 1936, but he continued to maintain an art gallery until his death in 1946.

In California, Edward Weston, already honoured for his flatly composed, softly lit, and often sentimental landscapes and figure pieces, turned around 1920 to a more direct and realistic use of the camera, with a certain emphasis upon abstract form that he never allowed to detract from the recognition of subject matter. "The camera should be used for a recording of *life*," he wrote in 1924, "for rendering the very substance and quintessence of the *thing itself*, whether it be polished steel or palpitating flesh." He formulated a working method based on the use of the classic view camera for 8- by 10-inch (20- by 25-centimetre) negatives, a sharp-cutting lens, well stopped down to ensure great depth of field, and the complete avoidance of artificial lighting. He printed the negatives by contact on glossy-surfaced paper to secure maximum detail. This technique was a logical outgrowth of his theory of previsualization. In 1934 he wrote,

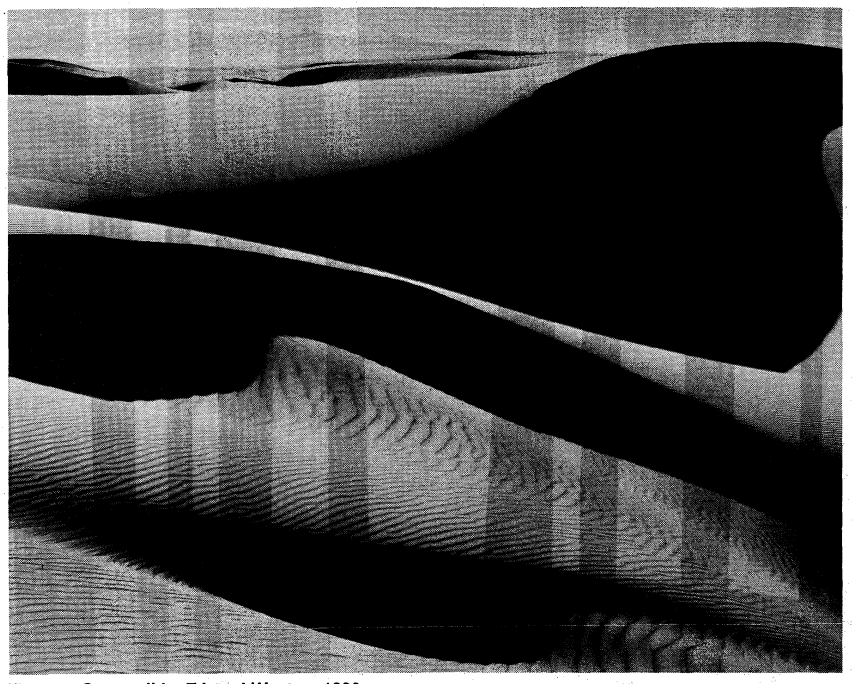
> The conception must be seen and *felt* on the camera groundglass complete in every detail; all values, textures, exact dimensions must be considered once for all, for with the shutter's release the isolated image becomes unalterably fixed. Developing the negative and making the final print, completes the original conception. This is the procedure in straight, *real* photography.

Weston's theory of previsualization

Retouching of any kind was inadmissable.

Paul Strand, whose striking close-ups and semi-abstract photographs Stieglitz had first exhibited, developed a style equally rigorous and self-disciplined during the same period. Working in Maine, the Gaspé, the Southwest, and Mexico, he produced powerful landscapes, direct portraits, and minute details of driftwood and plant life. Throughout his long career, Strand has emphasized the feeling of place in a way unequalled by any other photographer. In 1950 Nancy Newhall collaborated with him in editing *Time in New England*, a collection of his photographs with a selection of writings by New Englanders from the first settlers to the present day. When Strand moved to France, he wrote a series of similar publications on France, Italy, the Hebrides, and Egypt. None celebrate well-known capital cities; they concentrate upon life in small communities. Asked to define his sphere of interest, Strand once replied that he considered himself a photographer of people.

Viewing the negatives of Strand led Ansel Adams in 1930 to make photography his career. He was then studying the piano and photographing for his own sat-



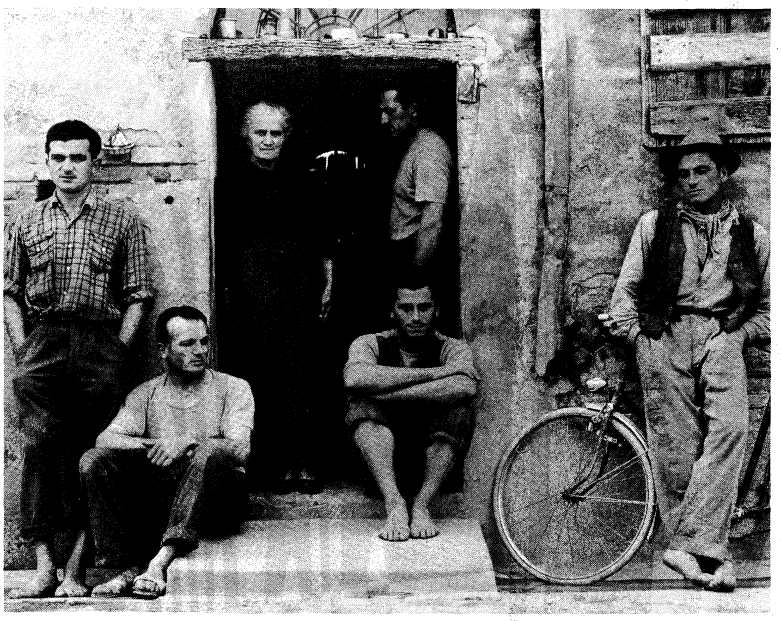
"Dunes, Oceano" by Edward Weston, 1936.
Cole Weston

tisfaction. He now determined to put his whole effort into photography. Long interested in nature—his first photographs were of the Sierra Nevada Range and Yosemite National Park—he refined and sharpened his technique. In 1932, with Willard Van Dyke, he founded an informal society, Group f.64, so named for the setting of the aperture of the lens that gave maximum depth of field. Edward Weston was an honoured member of the group and wrote the foreword to Adams' book *Making a Photograph*, published by the London Studio in 1935. This was the first of his many technical manuals to stress the concept of previsualization and to define in precise terms a functional aesthetic based upon the characteristics of the photographic medium. Adams' great contribution has been in what he calls "the interpretation of the natural scene." His photographs "Mount Williamson—Clearing Storm" (1944) and "Sierra Nevada from Lone Pine" (1944) are classics.

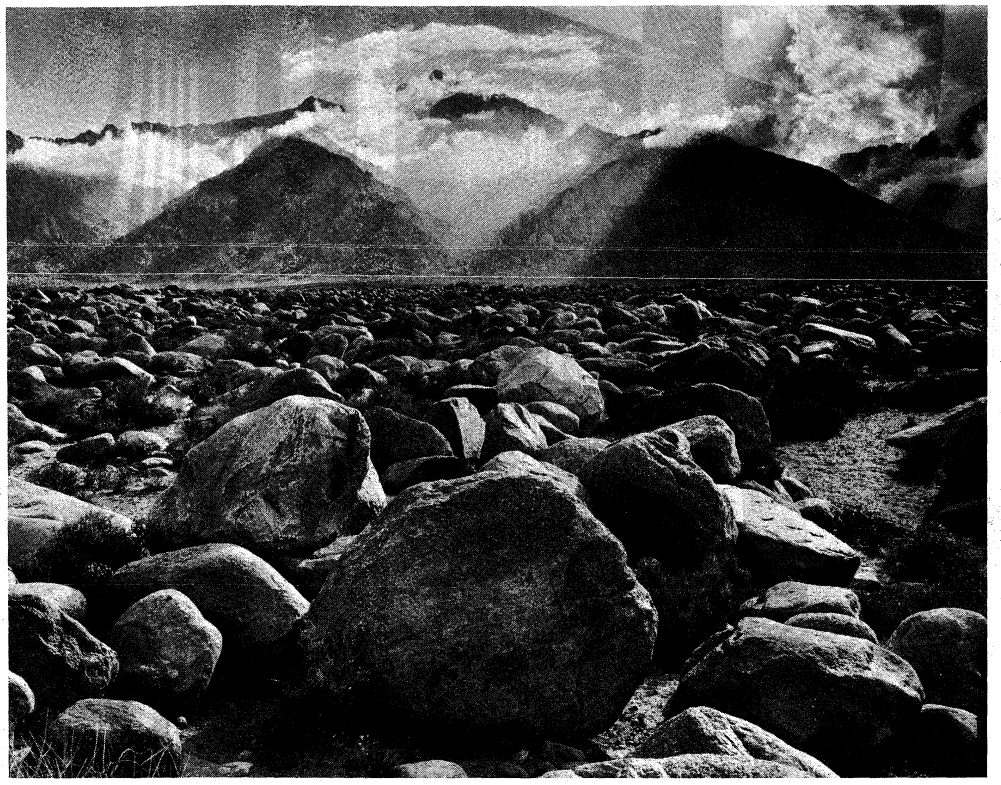
An approach somewhat similar to that of Weston, Strand, and Adams was to be found also in the work of Albert Renger-Patzsch, a professional photographer who spent most of his life in Essen, Germany. Using large tripod cameras, composing with deliberation and care on ground glass, he approached the world in a direct manner. He abhorred the casual and the accidental, and like Weston he insisted that the final image should exist, in all its completeness, before the exposure was made and that the print should directly record this image in full detail. His work was first exhibited in 1925, and four years later his most famous book appeared, *Die Welt ist schön* ("The World is Beautiful"). The very title set the work apart in a period when artists were more concerned with creating abstractions than interpreting the environment. Renger-Patzsch looked hard at the world. The photographs are stark, easily read; their challenge is their presentation of beauty in the ordinary.

Eugène Atget. At this same period, the photographs of Eugène Atget first became appreciated. Beginning

© 1953, Paul Strand



"The Family" by Paul Strand, 1953. One of a series of photographs comprising a profile of the small Italian town of Luzzara.



"Mount Williamson—Clearing Storm" by Ansel Adams, 1944.
Ansel Adams

around 1898, this French photographer produced a series of photographs of Paris and its environs that were direct, straightforward, and poetic in their sympathetic rendering of the very fabric of the city. He photographed shop fronts, buildings, carrousels, carriages of all kinds, and the people who earned their living in the streets. Unknown to the photographic world, Atget worked alone, earning a meagre living selling prints to architects, painters, and historians. The beauty of his photographs attracted the attention of Man Ray, an American painter living in Paris and a member of the Surrealist group, who published a few of them in the periodical *La Revolution surréaliste* in 1926. Upon Atget's death in 1927 his entire collection of prints and negatives was saved for posterity by the U.S. photographer Berenice Abbott; they are now in the collection of the Museum of Modern Art, New York.

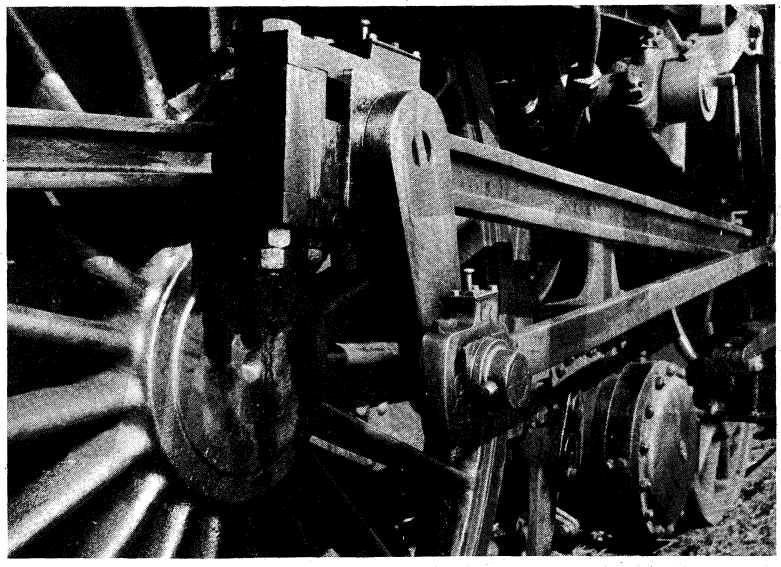
Influence of abstract art. In 1918 Christian Schad, a member of the Dada group of modern artists in Zürich, arranged small miscellaneous objects directly on photographic paper. Upon exposure to light, the paper darkened more or less or not at all, according to the opacity or transparency of the objects. These "Schadographs," as they came to be called, were minor contributions to the Dada movement and would be forgotten except that Man Ray and László Moholy-Nagy used like techniques.

Contributions of Man Ray

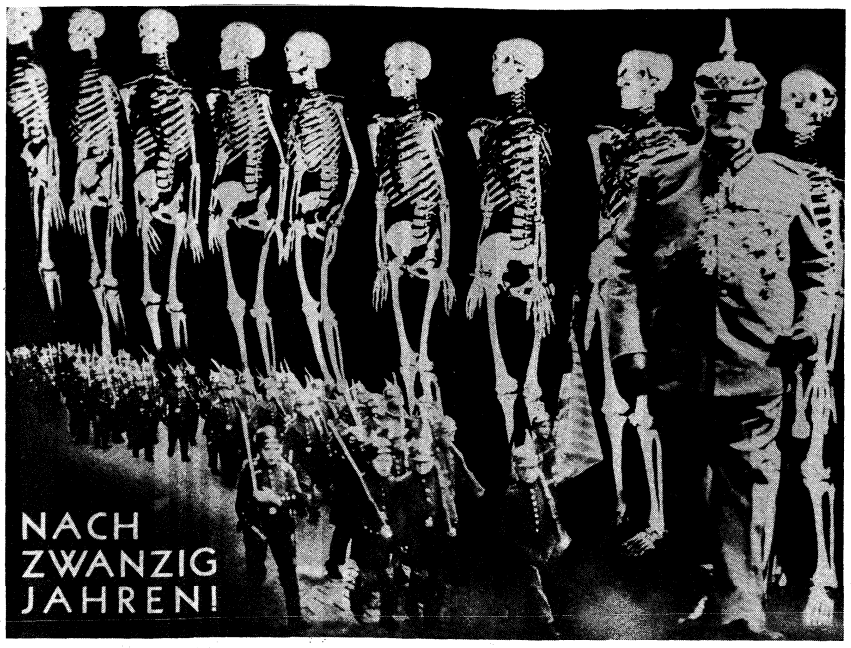
Man Ray was supporting himself by making fashion photographs in 1922 when he accidentally set a glass funnel, a graduate, and a thermometer on a piece of photographic paper, thus producing "Rayographs."

> I turned on the light; before my eyes an image began to form, not quite a simple silhouette of the objects as in a straight photograph, but distorted and refracted by the glass more or less in contact with the paper.

By courtesy of the Gernsheim Collection, The University of Texas at Austin



"Driving Shaft of a Locomotive" by Albert Renger-Patzsch, 1923.

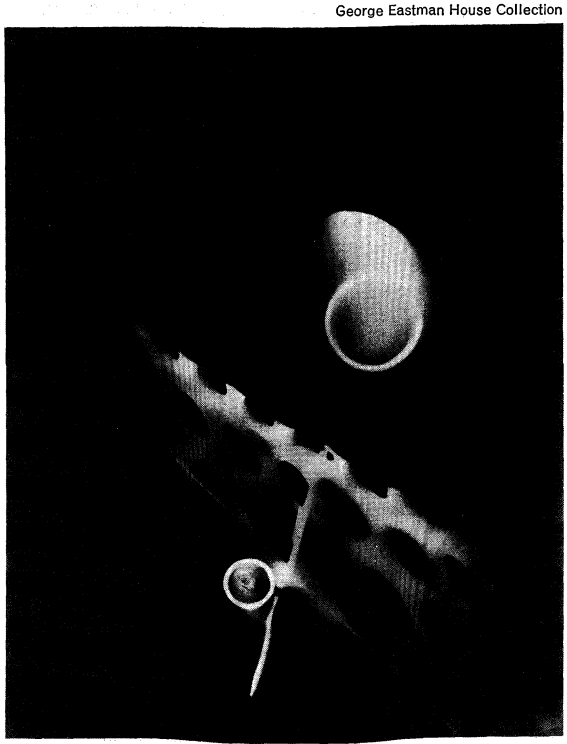


"Twenty Years Later," photomontage by John Heartfield, 1934.
By courtesy of Mrs. Gertrud Heartfield

At the same time, in Berlin, Moholy-Nagy began to make similar cameraless photographs. Although identical in technique, the work of each was quite different. Man Ray emphasized the distorted but recognizable object; Moholy-Nagy the play of light, no matter how abstract. The former brought to photography the vision of the Surrealists; the latter, that of the Constructivists.

As a teacher in the influential Bauhaus art school, Moholy-Nagy explored the potential of the unconventional use of the camera as a means of discovering form. He delighted in the worm's-eye and the bird's-eye view. He considered the negative an end as well as a means. Amazed at the power of the medium to make visible the invisible, he collected X-rays, photomicrographs, astronomical, and ultrahigh-speed photographs. When in 1925 he put together a book of these pictures titled *Painting, Photography, Film,* the eyes of the world were opened to the scope and breadth of photography as a tool for vision.

George Eastman House Collection



"Photogram," photogram by László Moholy-Nagy, 1926.

Photomontage. A marked characteristic of the new art that was evolving in the first decades of the 20th century was the free combining of mediums. Paintings could no longer be classified as oil, watercolour, or tempera; sculpture was not limited to wood, bronze, or marble. The Cubist painters began to glue on their abstract canvases words clipped from newspapers, labels from bottles, and even actual objects. The critic Guillaume Apollinaire wrote in 1913: "You may paint with whatever material you please, with pipes, postage stamps, postcards or playing cards, candelabra, pieces of oil cloth, collars, wall paper or newspaper." The extension of this collage technique (from the French *coller*, meaning "to glue") to photography was logical. The Austrian Raoul Hausmann, one of the pioneers, has said that the idea of photomontage came to him in 1918 when he chanced to see a colour lithograph of a grenadier with a portrait photograph pasted on the place of the head. The German John Heartfield, the greatest master of photomontage, claimed that the origin lay in postcards that he and his friend the German artist George Grosz sent to friends at the front during World War I. These were

The photomontage technique

> a mischmasch of advertisements for hernia belts, student song books and dog food, labels from schnaps and wine bottles, and photographs from picture papers, cut up at will in such a way as to say, in pictures, what would have been banned by the censors if we had said it in words.

The photomontages of Hausmann, Grosz, and Heartfield have a savage quality, reflecting their protest against the upheaval of German society following World War I. Violent contrasts of the scale and perspective of the image elements, the ruthless cropping of heads and bodies, the substitution of machine parts for vital organs, and other seeming illogical juxtapositions, were carefully calculated to have a shock effect. During the Third Reich, Heartfield's anti-Fascist montages were among the strongest protests made by any visual artist.

The montages of the Constructivists are more architectural: space is created with the purely photographic self-portrait of the Russian artist El Lissitzky (1924) and a fantasy world is built by Moholy-Nagy in his "Leda and the Swan." The power of montage lies in the tensions set up by the juxtaposition of disparate visual elements.

Instantaneous photography. In contrast to the production of such basically synthetic compositions as those produced by the masters of photomontage, and the analytical use of the camera to reveal aspects of the world previous-



'Children in Seville, Spain" by Henri Cartier-Bresson, 1933.
Henri Cartier-Bresson—Magnum

ly unknown or unseen, was the formation of an aesthetic based on the power of the camera to record life in the fraction of a second. Throughout its history, one of the characteristics of photography that most impressed the public was the relatively short length of time required to make pictures. Not only was the photographer seemingly more efficient than artists using other mediums, but he could reveal action hardly visible to the unaided eye. Throughout the 19th century this essential time factor was not considered an aesthetic one: indeed, the revelation by Muybridge of the true attitudes of figures and animals in motion had been scorned by artists—even though many of them made use of his findings. One of the earliest recognitions of the aesthetic value of the instantaneous image came with Moholy-Nagy's investigations of the manifold power of the camera to reveal the unseen and thus to expand vision. Most of the pictures that he reproduced in his influential book, *Painting, Photography, Film*, were not the work of conscious artists, but of press photographers, scientists, and amateurs.

To make a photograph of rapid action and then from scores of photographs to discover in some of them a new vision of great beauty is to use photography as a tool rather than a medium. The ability to train oneself to perceive the significant in the fraction of a second and to use the camera with such speed and precision that the instantaneous perception is preserved forever is a great creative gift. The gift is evident in the work of the Hungarian André Kertész as early as 1915 and in his later work in Paris during the 1920s. The Frenchman Henri Cartier-Bresson began around 1930 to develop the style that he later called the search for the "decisive moment." To him the camera is an "extension of the eye." With extraordinary precision he perceives a fully composed picture of the most fleeting scenes. Unlike many other photographers, he does not crop or recompose his pictures after they have been taken: every detail of a Cartier-Bresson photograph is present in the negative. He prefers the miniature 35-mm-film camera. When he finds a picture possibility he stalks his prey unobtrusively, working with his camera to a visual climax.

Cartier-Bresson's instantaneous image

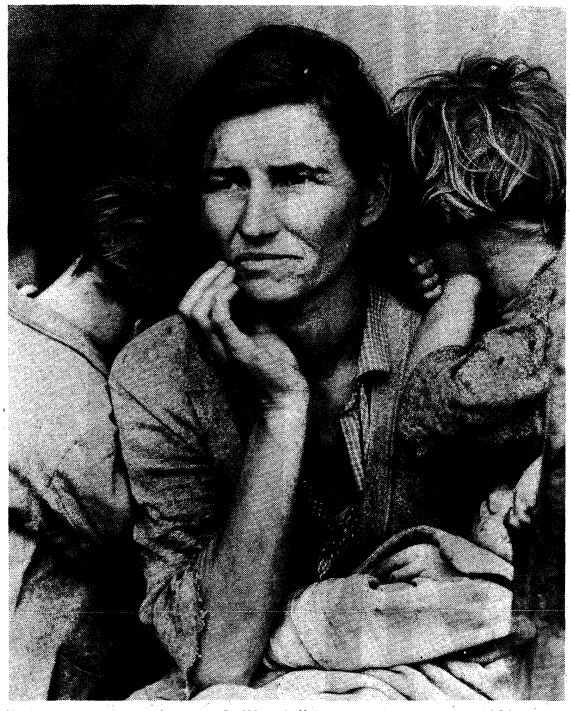
Documentary photography. The value of the photograph as a record of the world and man's achievements had been established with the very publication of the first two successful techniques in 1839. William Henry Fox Talbot foresaw at once that his "photogenic drawing process" could be used to record ancient buildings, to inventory collections of artifacts, to duplicate manuscripts, and to reproduce botanical specimens. The first critics of Daguerre's daguerreotype process saw it as a means of bringing to Paris "the most remarkable views of the world," to quote the title-page of the two-volume album of engravings, *Excursions daguerriennes*, publication of which began in 1841. The recognition of the power of photography to persuade as well as to inform, however, came somewhat later. The classic sociological study *London Labour and the London Poor*, by Henry Mayhew, John Binny, and others (1851–62), was illustrated with drawings partly credited to daguerreotypes taken by Antoine Claudet; a sequel, patterned upon it, appeared in 1877—*Street Life in London*, by Adolphe Smith and John Thomson. The photographs, taken by Thomson, were moving, straightforward, unretouched pictures of chimney sweeps, flower sellers, bargemen, and other tradesmen. They were reproduced by the woodburytype process, a technique now long obsolete, which gave almost exact facsimiles of the original prints. The intent of the publication was to show—as Charles Dickens had shown in his novels—the hardships and problems faced by the ever-growing underprivileged population of London. Each of the photographs was accompanied by a detailed explicative text by Smith. Oscar G. Rejlander photographed orphan children in the streets of London performing such humble tasks as cleaning boots and opening carriage doors.

Thomson's London scenes

To Jacob A. Riis, a police reporter in New York City in the 1880s, the camera became the ally of his pen in the personal crusade he was waging to better the lot of the immigrants who then lived and worked in wretched conditions in teeming tenements in Manhattan's lower East Side. Himself an immigrant—he arrived in New York from Denmark in 1870—he knew at first hand the conditions he was seeking to eradicate. To supplement his written descriptions, he turned to photography. When the cameraman he employed proved unsatisfactory, he learned the process himself and was one of the first to use flash powder, a German invention recently introduced to America. His photographs, published in crude facsimile in newspapers and in his now-famous book *How the Other Half Lives* (1890), were instrumental in legislative reform.

Beginning in 1905, Lewis W. Hine, a teacher at the Ethical Culture School in New York and a trained sociologist, began to make a pictorial record of immigrants passing through Ellis Island. He followed them in New York City, photographing them in their living quarters, at work, and in the streets. Later Hine travelled to textile mills, mines, and other places where young children were employed; largely through the evidence of his photographs, legislation was eventually passed against child-labour abuses.

This use of photography came to be called documentary in the 1920s by a group of motion-picture producers in Great Britain, headed by John Grierson. They saw the great potential of the photographic image as believable evidence. But the simple record photograph is not enough



"Migrant Mother, Nipomo, California" by Dorothea Lange, 1936.
By courtesy of the Library of Congress, Washington, D.C.

in itself; it must be endowed, as were those of Thomson, Riis, and Hine, with compassion for the subject and a sense of beauty. Seemingly, the use of photography as a tool for the betterment of mankind would belong more properly in a discussion of sociology rather than art, but some of the greatest examples of the art of photography have been produced under the banner of the "documentary movement."

Documenting the Depression in America

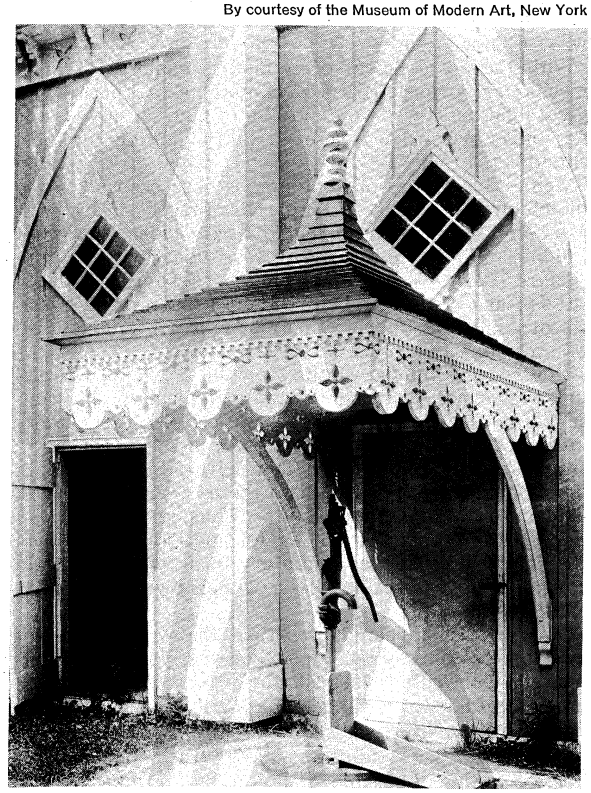
This is especially true of the body of photographs produced in America during the Great Depression by a group of dedicated photographers, under the direction of Roy E. Stryker, for the Farm Security Administration. The pictures covered the period from 1935 until the outbreak of World War II. Stryker, a professor of economics at Columbia University who had used photography in his studies, was invited by the Secretary of Agriculture to form a historical section of a newly formed division of the department that was concerned with the plight of farmers driven from their land in the dust bowl and who were migrating to the West. This at once took the form of a photographic project. The first photographer to be hired was Arthur Rothstein, a student of Stryker's. From California came Dorothea Lange, who had photographed migratory workers there. Her photographs are notable for their compassionate attitude toward people. In her "Migratory Pea Picker," a destitute young mother, surrounded by her children, peers at the camera with determination and courage. Walker Evans, with a direct uncompromising sense of environment and the beauty of everyday architecture, contributed a notable series. With the writer James Agee he documented the lives of sharecroppers in the book *Let Us Now Praise Famous Men* (1941). Others were documenting America at this time: Berenice Abbott produced a notable series of photographs of New York City, and Margaret Bourke-White with her husband Erskine Caldwell did a sensitive survey of the South, which appeared in book form as *You Have Seen Their Faces* (1937).

Photojournalism. From the outset, photography has served the press. Within weeks after the French government's announcement of the process in 1839, a daguerreotype was reproduced in the magazine *Le Lithographe*, and soon the by-line "from a photograph" appeared with increasing frequency in the captions of wood engravings in the popular illustrated newspapers and magazines. These reproductions were generally crude and carried little of the conviction of the original photograph. Regular use of photographs began with the perfection of the half-tone process for facsimile reproduction in the late 1880s. Magazines and newspapers now turned to photography for reporting topical events, and the profession of illustrator, which so many noted artists in Europe and America had followed, gradually became obsolete. With a few notable exceptions, such as William Warneke titled "Shooting of Mayor William J. Gaynor of New York," which appeared in the New York *World* in 1910, these early news photographs were unimaginative and dull.

After World War I, two of the largest picture magazines in Europe, the *Münchner Illustrierte Zeitung* and the *Berliner Illustrirte Zeitung*, began to develop a new approach to pictorial journalism. Using miniature cameras fitted with wide-aperture lenses—at first the Ermanox and then the Leica—Erich Salomon and Felix Man (pseudonym of Hans Baumann) took photographs indoors by available light, without flash. This technique led to informality of pose and a sense of presence that was remarkable. Stefan Lorant, editor of the *Münchner Illustrierte*, encouraged Man to take sequences of photographs, which he then laid out in imaginative picture essays. The example of the German picture magazines was followed in other parts of Europe and in America. One was the short-lived *Vu*, published in Paris. An issue of *Vu* devoted entirely to the Spanish Civil War contained memorable photographs by Robert Capa. In 1936 both *Life* and *Look* were conceived in America, and a formula was evolved in which the picture editor, photographer, researcher, and writer constituted a team. The result was the creation of a definite photographic style. *Life*'s first photographers were Margaret Bourke-White, already famous for her industrial photographs made largely for the magazine *Fortune*; Alfred Eisenstaedt, an experienced photo reporter for the German illustrated weeklies; Peter Stackpole, whose photographs of the Golden Gate Bridge in San Francisco attracted much attention; and Thomas D. McAvoy, a news photographer who had pioneered in photographing interior scenes by available light. The concept of *Life* from the start was summed up in the words of the founder, Henry Luce, who proposed to replace haphazard picture taking and editing with the

Life's first photographers

By courtesy of the Museum of Modern Art, New York



"Maine Pump" by Walker Evans, 1933.

"mind-guided camera." Photographers were briefed for their assignments and encouraged to take great quantities of photographs, in order that the editors might have a large selection. The visual organization of the picture story was carefully planned for maximum reader impact. The opening photograph of the picture essay, as this type of reportage came to be known, established the situation, and like written narration there was a visual climax and a definite conclusion. Usually the photographs were chosen and arranged on the pages before the accompanying text was written. Unlike the illustrated article, the picture essay quite logically is based upon the photographs, and the text is devoted to information that cannot be expressed visually: names, dates, places.

In the early days of *Life* and *Look*, photographers made great use of so-called synchroflash; *i.e.*, flash that was synchronized with the camera shutter, as distinct from open flash, which obliged the photographer to open the shutter, fire the flash, and close the shutter manually. The next step was the multiple flash, which made possible more sophisticated and pleasing lighting effects. The aim of *Life* and *Look* was to stress the human aspects of each situation as much as possible. By duplicating the existing illumination with the synchroflash lighting, photographers could await the moment when people were at their most natural and then make the exposure without the need for posing. This style was used brilliantly by Bourke-White. She wrote:

> Frequently I have set up a camera with remote control in the corner of a room, seated myself at a table some little distance away, and released the flash, possibly an hour later, when everybody had forgotten about the camera.

This way of photographing was soon challenged. Stefan Lorant, formerly editor of the *Münchner Illustrierte Zeitung*, left Germany during the Third Reich for London and established *Picture Post* in 1938. Working with him was Felix Man, who had pioneered with Erich Salomon in the use of the miniature camera. Lorant encouraged Man and other *Picture Post* photographers to develop the technique and pictorial style of available-light photographs so brilliantly begun in the 1920s. Their pictures, though lacking in sharpness and depth, had a remarkable naturalness that brought great reader appeal—so much so that *Life* began to publish similar photographs and hired a former *Picture Post* photographer, Leonard McCombe, with an extraordinary clause in his contract: he was forbidden to use flash.

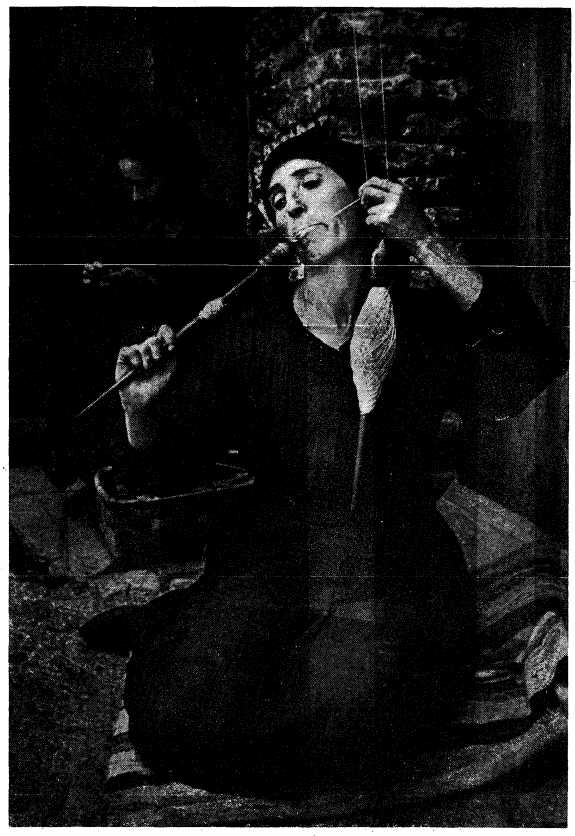
Available-light photography

With the development of miniature cameras of great mechanical precision, lenses combining sharpness with great light-passing power, and film of high sensitivity, it is no longer necessary to use auxiliary lighting for most situations.

Memorable groups of photographs have been taken for the picture magazines. Examples are Man's "A Day with Mussolini," first published in the *Münchner Illustrierte Zeitung* (1931), and then, with a brilliant new layout, in *Picture Post;* W. Eugene Smith's "Spanish Village" (1951) and "Nurse Midwife" (1951) in *Life;* and Eisenstaedt's informal, penetrating portraits of famous Britons, also in *Life*.

Photographic books. Following the trend of the picture magazines, publishers began to replace the illustrated book, in which photographs were often haphazardly introduced to illustrate the text, with layouts designed to emphasize the photographs themselves. In these books the emphasis was reversed, and the text served rather to explain or illuminate the pictures. Thus, in *An American Exodus* (1939), by Paul S. Taylor and Dorothea Lange, striking photographs of migratory farm workers are accompanied by short quotations of their own words spoken at the time.

Other works appeared, in which the writing did not explain the photographs nor did the photographs illustrate the writing. Instead, the images and text were used together to produce what Wilson Hicks, *Life's* first picture editor, called the "third effect." Recently some magazine publishers have turned to book production, using the same techniques of group journalism in layout, writing, and production that had proved so successful in their magazines. To many photographers, books have become a chief method of the presentation of their work. In 1960 the Sierra Club began publication of handsome folio size books with *This is the American Earth* by Ansel Adams and Nancy Newhall. This book was followed by other publications, notably a series by Eliot F. Porter, entirely in colour.



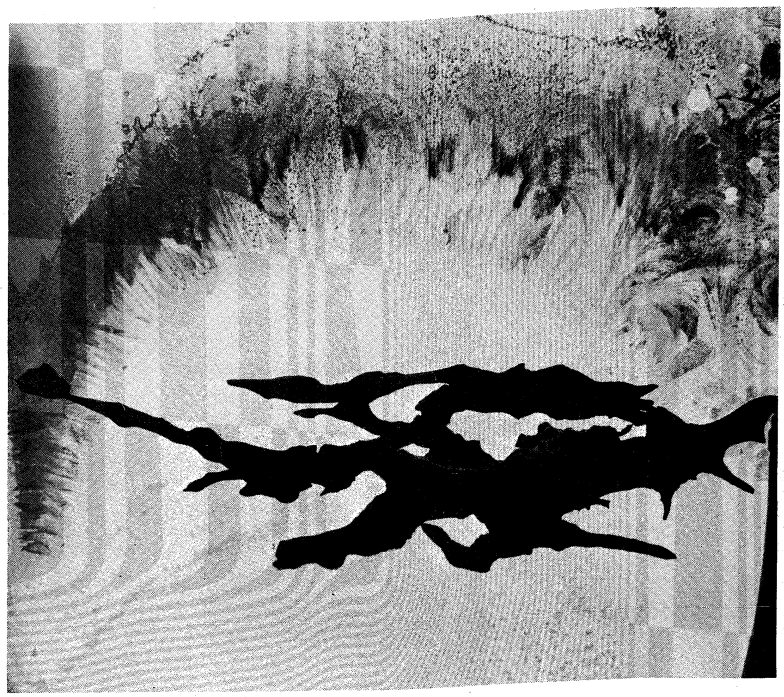
"Spanish Village" by W. Eugene Smith, 1951.
W. Eugene Smith

Recent trends. Throughout its history, there have been two complementary yet distinct aesthetic approaches to photography. On the one hand, there has been the recognition of the basic qualities of photography and the desire to make use of them in a functional way. On the other hand, there have been those who believe that the most aesthetic use of photography is to relate it to other mediums. Since 1950 these trends have been pursued with equal vigour. To them has been added a third approach, which may be summed up in the word Alfred Stieglitz used to describe his late work: "the equivalent." To him photographs could be an equivalent of his thoughts and emotions.

In the United States, Minor White, through his long and distinguished career, his writing, his teaching, and his founding and editing of the influential magazine *Aperture,* has developed the Stieglitz approach to a highly sophisticated level. Metamorphosis to him is one of the paradoxes and virtues of photography. Its very reality and believability allow the sensitive artist to endow his photographs with overtones that may verge upon the mystical. The photograph must be transformed in such a manner that the viewer can read an inner message, which is not visible upon the surface, but which is carried by it. White's book *Mirrors, Messages, Manifestations* (1970) is a collection of superb photographs that present his spiritual biography.

Influence of Minor White on photography

To Aaron Siskind, who has worked with wall scrawls, weathered wood and plaster, torn billboards, and what he calls "the detritus of our world," the photograph must communicate more than the subject itself. The photographs of Harry Callahan express his highly developed sense of linear form, often by means of sharp contrasts of black and white and multiple images.



"Ritual Branch" by Minor White, 1958.
George Eastman House Collection

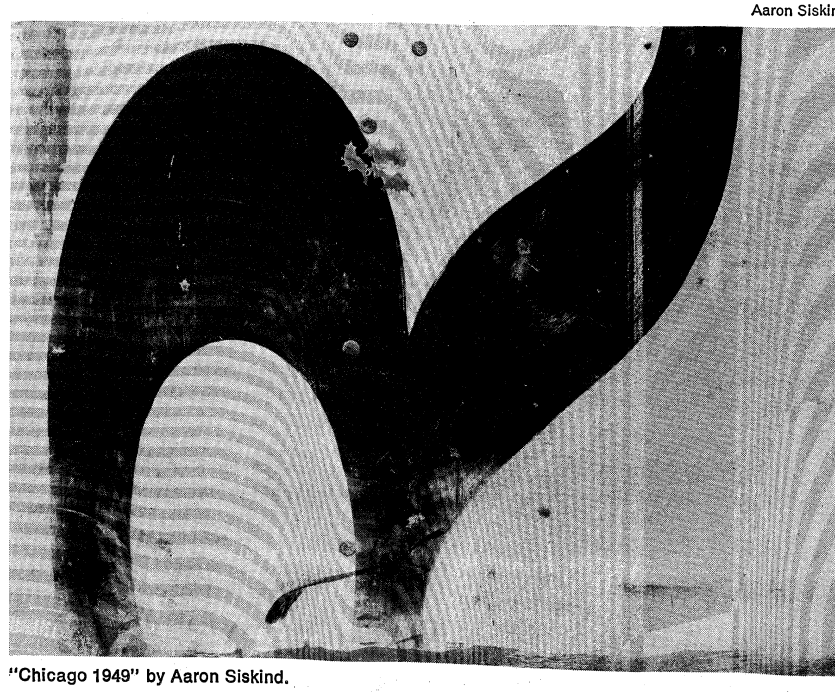
The American scene has been a subject of much interest to photographers: often called the "Social Landscape," the work of such photographers as Robert Frank, Bruce Davidson, and Lee Friedlander takes the form of penetrating sociological observations, somewhat reminiscent of the documentary photography of the 1930s. The approach differs in that rather than presenting problems faced by a certain level of American society, emphasis is placed on the effects of the urban environment upon people.

Colour photography has become increasingly popular within the ranks of the amateur. Statistics on home processing, however, show a reverse trend, a reflection in part of the more difficult techniques demanded by existing colour film and print material. Many professionals, however, prefer black and white to colour film for aesthetic, rather than technical reasons, and only a few have explored the artistic possibilities of colour photography. Among professional colour photographers, Eliot F. Porter and Marie Cosindas are the leaders in America: both prefer a somewhat realistic approach.

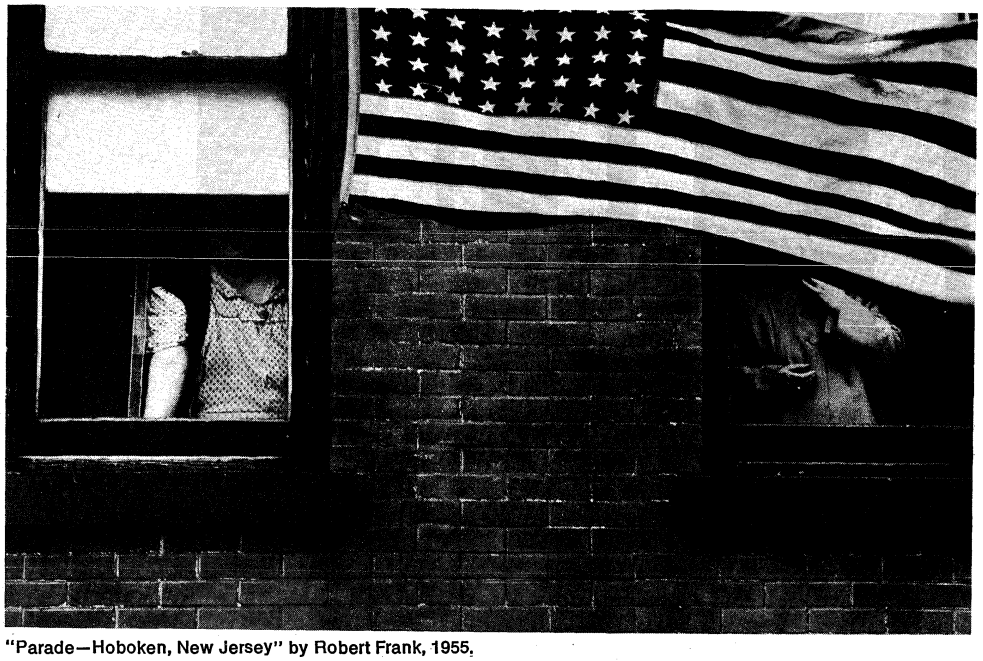
In England and Europe, trends closely parallel those in the United States. Bill Brandt, whose documentation of London began with the book *The English at Home* (1936) and continued during World War II with trenchant photographs of the bombing of that city, changed his style during the 1950s: with the use of an extremely wide-angle lens he created highly abstract studies of nudes. The "social landscape" school is well represented in the work of Tony Ray-Jones, who studied and worked in the United States before returning to Great Britain in 1966.

On the Continent, the most imaginative photographs are mostly made for publication, rather than for exhibition or hanging in galleries and homes. Europeans, in general, do not consider the photographic print as an end in it-

Aaron Siskind



"Chicago 1949" by Aaron Siskind.



"Parade—Hoboken, New Jersey" by Robert Frank, 1955.
Robert Frank

self, but as a step toward reproduction in periodicals and books. Brassaï (the pseudonym of Gyulas Halász), made his name in photography with the publication of *Paris de Nuit* (1933), intimate and sympathetic photographs of night life in the more humble quarters of Paris. Recently, like Aaron Siskind, he has been fascinated with wall scrawls and *graffiti;* a selection of them was published in 1960. Robert Doisneau is a master of humour, catching moments of absurdity in everyday life. Lucien Clergue has turned to the natural scene, and to the nude in surf. The tradition of the cameraless abstraction has been enriched by the Belgian Pierre Cordier with the introduction in 1958 of his "chimigrammes"—images made, not by light, but chemical action on photographic paper.

In Germany, the greatest influence has been the teaching of Otto Steinert at Saarbrücken and, since 1959, at the Folkwangschule in Essen. Almost single-handedly he brought back to Germany that spirit of experimentation and boldness of concept which had been suppressed during the Third Reich. An excellent photographer, Steinert was the founder in 1951, together with the art historian J.A. Schmoll gennant Eisenweth, of a movement they named "Subjective Photography." Chargesheimer, who made extraordinary abstractions, showed a different approach in his book *Cologne Before 5:30 A.M.:* a hard, almost brutal look at the skeleton of a city, with not a human to be seen. Horst Baumann specialized in illustration, particularly in colour. The Austrian, Ernst Haas is a master of colour photography, turning toward the abstract in his remarkable photographs of blurred action and bold compositions. A member of the Magnum group, a cooperative formed by Cartier-Bresson and others, Haas has produced work that is international in scope. The leading Swiss photographer is Emil Schulthess. In Italy the photographer's concern is chiefly centred upon the recording of the social scene: Mario Giacomelli's series of photographs of nuns and village life is typical and outstanding.

In Japan the extraordinary photographs of the landscape, particularly of Fuji-yama (Mt. Fuji) by Hiroshi Hamaya, parallel the straight photographic approach. Superb photographs have also been produced by Takayuki Ogawa, the social realist.

A lively interest in expanding the medium of photography beyond the straight approach is characteristic of much recent work. The experimentation takes many forms, including the revival of long-obsolete printing techniques such as the gum bichromate process and the platinotype. Perhaps the most startling change that has come about since 1960 is an interchange of mediums between photographers and painters. Many photographers make liberal use of manual techniques, such as negative and print retouching and the addition of colour. Simultaneously, painters, who have long utilized photographs as tools for observation, boldly imitate the very quality of photographic vision and sometimes introduce unaltered photographs by collage techniques or silk-screen reproduction directly into their canvases.

Interchange of mediums between photographers and painters

Collections. With the general acceptance of photography as an artistic medium, most museums today house collections of fine prints. The Metropolitan Museum of Art in New York City has long considered photography a branch of the graphic arts and has a notable collection in its Department of Prints and Photographs, including originals of the photographs reproduced by Stieglitz in *Camera Work*. The Museum of Modern Art, also in New York City, pioneered through the formation in 1940 of a separate Department of Photography. The International Museum of Photography at George Eastman House in Rochester, New York, is the largest museum of photography in the world; its rich collection covers production since the days of Daguerre. Photographs are also to be found in the Art Institute of Chicago, the San Francisco Museum of Art, the Oakland (California) Museum, and the Amon Carter Museum of Western Art in Fort Worth, Texas. The collection formed by the historians Helmut and Alison Gernsheim is now in the University of Texas at Austin. Both the Smithsonian Institution and the Library of Congress, Washington, D.C., have large collections of fine photographs.

In Great Britain the largest collections are those of the Royal Photographic Society, the Science Museum, and the Victoria and Albert Museum, all in London, and the Kodak Museum in Harrow, Middlesex. The Société Française de Photographie in Paris has a superb collection, dating from the formation of the society in 1854, and the Cabinet des Estampes of the Bibliothèque Nationale possesses large quantities of photographs. The largest collection in Germany, based on materials collected by the historian Erich Stenger, is owned by Agfa-Gevaert A.G., the photographic-materials manufacturers in Leverkusen. Collections are also in the Folkwang Museum, Essen, and the Staatliche Landesbildstelle in Hamburg. In Italy the Museo del Cinema in Turin, in addition to material on the history of motion pictures, has a collection of photographs.

BIBLIOGRAPHY. ALBERT BONI (ed.), *Photographic Literature,* 2 vol. (1962–72), an exhaustive and valuable bibliog-

raphy, listing books and articles, with detailed references to bibliographies.

General works: *The Encyclopedia of Photography*, ed. by WILLIARD D. MORGAN, 20 vol. (1963–64), richly illustrated, with contributions by world leaders in the art, science, and technique of photography; *Focal Encyclopedia of Photography*, rev. ed., 2 vol. (1965), especially strong on technique and scientific theory; *Life Library of Photography*, 17 vol. (1970–72), an indispensable series of well-illustrated books, covering all aspects of photography, with special emphasis on its use as a medium of communication and expression, unique in that the history of the medium is integrated with technical exposition.

Equipment, materials, and techniques: ANSEL ADAMS, *Camera and Lens: The Creative Approach*, rev. ed. (1970), vol. 1 of the series "Basic Photo," gives a detailed description of the aesthetic capabilities of the medium with precise technological instructions and illustrations. See also the earlier volumes in the series: *The Negative; The Print; Natural-Light Photography; Artificial-Light Photography* (1948–56); and ANDREAS FEININGER, *Total Picture Control* (1961), a textbook of advanced photographic technique. (*Colour photography*): E.J. WALL, *The History of Three-Color Photography* (1925, reprinted 1970), a basic technological survey of all known methods of synthesizing colours by recording the red, green, and blue rays photographically; J.S. FRIEDMAN, *History of Color Photography*, 2nd ed. (1968), a continuation of Wall's *History;* RALPH M. EVANS, W.T. HANSON, JR., and W. LYLE BREWSTER, *Principles of Color Photography* (1953), a detailed discussion of the visual perception of colour and the various systems by which it can be recorded photographically; LEWIS WALTON SIPLEY, *A Half Century of Color* (1951), a good survey of the medium, less technical than the encyclopaedic works of Wall and Friedman, discussing the use of colour photography, especially in the graphic arts.

Stylistic and historical development: JOSEF MARIA EDER, *Geschichte der Photographie*, 4th ed. (1932; Eng. trans., *History of Photography*, 1945), a pioneer Austrian work that deals primarily with the scientific and technological development of photography, but also discusses the work of artist-photographers (most useful for detailed citations of original sources); NATHAN LYONS (ed.), *Photographers on Photography* (1966), anthology of essays by photographers from H.P. Robinson (1892) to Wynn Bullock (1962); BEAUMONT NEWHALL (ed.), *On Photography* (1956), anthology of the inventors' own accounts of various processes, photographers' opinions of their art, and criticism; BEAUMONT and NANCY NEWHALL (eds.), *Masters of Photography* (1958), short, yet detailed, biographical sketches of 19 photographers, from David Octavious Hill to the present, with representative photographs of each; BEAUMONT NEWHALL, *The History of Photography from 1839 to the Present Day*, rev. ed. (1964), the stylistic development of the art of photography as related to the technological and scientific characteristics of the medium; and *Latent Image: The Discovery of Photography* (1967), although written from a scientific point of view, helpful for the understanding of the establishment of the medium; PETER POLLACK, *The Picture History of Photography*, rev. and enl. ed. (1969), especially valuable for its wealth of illustrations; *French Primitive Photography* (1969), a catalog of an exhibition, mainly from the collection of Andre Jammes and George Eastman House, with superb reproductions; HELMUT and ALISON GERNSHEIM, *The History of Photography from the Camera Obscura to the Beginning of the Modern Era*, 2nd rev. ed. (1969), an encyclopaedic general history of picture making by photography, with particular emphasis on European work; HELMUT GERNSHEIM, *Creative Photography: Aesthetic Trends, 1839–1960* (1962), complements the Gernsheim's *History* by bringing aesthetic developments up to 1960; GEORGES POTONNIEE, *Histoire de la decouverte de la photographie* (1925; Eng. trans., *History of the Discovery of Photography*, 1936), the first detailed account of the early days of photography; JOHN SZARKOWSKI, *The Photographer's Eye* (1966), a penetrating examination of photographic aesthetics covering photography's development; ROBERT TAFT, *Photography and the American Scene* (1938, reprinted 1964), the first history of photography in America in the 19th century. (*English pioneers*): D.B. THOMAS, *The First Negatives: An Account of the Discovery and Early Use of the Negative-Positive Photographic Process* (1964), a well-documented account of William Henry Fox Talbot's photographic work; WILLIAM HENRY FOX TALBOT, *The Pencil of Nature* (1844–46, reprinted 1969), the inventor's account, illustrated with 24 actual calotypes; see also BEAUMONT NEWHALL, *Latent Image* (1967). (*The Daguerreotype process*): BEAUMONT NEWHALL, *The Daguerreotype in America*, rev. ed. (1968), a study of the industry as well as the art of daguerreotyping, with a biographical index, a detailed discussion of the technique, and bibliography; RICHARD RUDISILL, *Mirror Image: The Influence of the Daguerreotype on American Society* (1971), a thorough survey, with emphasis upon the impact of the daguerreotype on American culture. (*The Collodion process*): FREDERICK SCOTT ARCHER, "On the Use of Collodion in Photography," *The Chemist*, Series 2, 2:257–258 (1851), the inventor's account, reprinted in BEAUMONT NEWHALL, *On Photography*, pp. 79–80 (1956); JOHN TOWLER, *The Silver Sunbeam* (1864, reprinted 1969), the most complete and popular textbook on photography during the collodion era. (*Pictorial effect and combination printing*): HENRY PEACH ROBINSON, *Pictorial Effect in Photography*, 3rd ed. (1886), the first book to provide directions for the production of photographs in the manner of paintings. (*Combat photography*): FRANCIS TREVELYAN MILLER (ed.), *The Photographic History of the Civil War*, 10 vol. (1911; reprinted in 5 vol., 1957), the first major publication of the immense photographic documentation of the Civil War by M.B. Brady, Alexander Gardner, and others. (*Stereoscopic photography*): WILLIAM CULP DARRAH, *Stereo Views: A History of Stereographs in America and Their Collection* (1964), a collector's account of the development of the industry, with useful checklists of photographs and publishers of views in all the U.S. states and considerable information on European production; BEAUMONT NEWHALL, "Looking Back at Stereo," in WILLIARD D. MORGAN and HENRY M. LESTER, *Stereo Realist Manual*, pp. 375–394 (1954), a comprehensive essay on the invention of the stereoscope; DAVID BREWSTER, *The Stereoscope: Its History, Theory and Construction* (1856), the first book on the application of stereo vision to photography, by the inventor of the lenticular, box-form stereoscope. (*The dry plate*): WILLIAM JEROME HARRISON, *A History of Photography* (1887), a detailed account of the perfection of the gelatino-silver-bromide process, with a letter from the inventor, R.L. Maddox; C.E. KENNETH MEES, *From Dry Plates to Ektachrome Film* (1961), the former director of research at Eastman Kodak recounts the invention of the dry plate and subsequent improvements, for many of which he was responsible; BEAUMONT NEWHALL, "Muybridge and the First Motion Picture," *Image*, 5:4–11 (1956), a biographical essay.

The Linked Ring and the Photo-Secession: JOSEPH T. KEILEY, "The Linked Ring: Its Position and Origin, and What It Stands for in the Photographic World," *Camera Notes*, 5:111–120 (1901), an account of the establishment of this influential society, its membership, and its exhibitions; *Camera Work: A Photographic Quarterly*, published and edited by ALFRED STIEGLITZ, 50 vol. (1903–17, reprinted 1969) magnificently illustrated quarterly essential for the understanding of the contribution of Stieglitz to photography and the introduction of modern art to America (for excellent reproductions of a selection of the plates, see *Life Library of Photography: The Print*, pp. 25–49); ROBERT DOTY, *Photo Secession: Photography As a Fine Art* (1960), an account of events leading up to the founding by Alfred Stieglitz of the society, with a detailed chronicle and interpretation of its activities.

Photojournalism: KEN BAYNES (ed.), *Scoop, Scandal and Strife: A Study of Photography in Newspapers* (1971), the first history of news photography to discuss the perceptual and editorial aspects as well as the technical; WILSON HICKS, *Words and Pictures: An Introduction to Photojournalism* (1952), basic description by the pioneer picture editor of *Life* magazine of problems and solutions of the photographic essay; see also, NANCY NEWHALL, "The Caption," *Aperture*, 1:17–29 (1952), a study of the mutual relationship of words and photographs; STANLEY RAYFIELD, *How Life Gets the Story* (1955), field experience of photographers of *Life*, presented in the format and photographic essay style of the magazine; NATHAN LYONS (ed.), *Toward a Social Landscape: Bruce Davidson, Lee Friedlander, Garry Winogrand, Danny Lyon, Duane Michals* (1966), photojournalistic techniques applied to observation of the urban scene; *The Persistence of Vision: Donald Blumberg, Charles Gill, Robert Heinecken, Ray K. Metzker, Jerry N. Uelsmann, John Wood* (1967), mixed media, biographical and bibliographical data on each photographer; *Vision and Expression* (1969), catalog of an exhibition; JERRY N. UELSMANN (1970), chronology, bibliography, and list of exhibitions.

Collections: LOUIS WALTON SIPLEY, "Museums and Collections of Photographs," in *Encyclopedia of Photography*, vol. 13, pp. 2468–2477 (1963); JACOB DESCHIN'S article "Historic Photo Collections on the Move," *Popular Photography*, 69:102–105, 138–139 (1971), brings Sipley's report up to date. See also "Museums and Collections," in the *Focal Encyclopedia.*

(Be.N.)

Photography, Technology of

Photography is the recording of visible images by light action on light-sensitive materials. The name is derived from the Greek *photos* (light) and *graphos* (writing). Usually photography refers to the formation of optical images projected by a lens in a camera onto a film or other material carrying a layer of light-sensitive silver salts; in an extended sense, the term also covers images formed by certain invisible radiations (ultraviolet and infrared rays) and images recorded in other sensitive materials not containing silver by means of chemical or physical processes or both. Related to photography is image recording by X-rays, electron beams, and nuclear radiations (radiography) and the recording and transmission of light images in the form of electromagnetic signals (television and video tape).

Photography is either still photography or motion-picture photography. Still photography records single self-contained images on separate sheets of film or other sensitized material. For convenience in camera use, the film may be a short strip carrying a number of such images. In cinematography or motion-picture photography a continuous strip of film records sequences of images representing successive movement phases of a subject in motion. Suitable projection then re-creates the illusion of movement (see MOTION PICTURES, TECHNOLOGY OF).

Still photography covers principally the recording of optical images formed by a lens in a camera and the duplication and reproduction of such images by light action (printing) to yield final pictures for various purposes. Generally, still photography is considered under two aspects: (1) pictorial and record photography, the aim of which is to yield a more or less accurate representation of an object photographed as an end in itself and (2) applied photography, in which the photographic image is an intermediate stage of, or serves as a tool for, other processes and procedures. Examples of applied photography are photomechanical reproduction for the graphic-arts field, making printed circuits, photographic templates, and the use of photography in measurements. Such uses often overlap applications that concern techniques of photographing special objects or objects under special conditions —*e.g.*, high-speed photography and photomicrography.

This article is divided into the following sections:

I. History and basic technology
  History
  Basic technology
II. Cameras and lenses
  Camera technology
  Photographic lenses
III. The photographic process
  Black-and-white films
  Picture-taking technique
  Processing and printing
  Colour photography
IV. Special techniques
  Polaroid photography
  Special photosensitive systems
  Special techniques and applied photography
V. The photography industry

I. History and basic technology

HISTORY

Early discoveries. The forerunner of the camera was the camera obscura, a dark chamber or room with a hole (later a lens) in one wall through which images of objects outside the room were projected on the opposite wall. The principle was probably known 1,000 years ago and was described in the manuscripts of Leonardo da Vinci. The Italian scientist and writer Giovanni Battista della Porta, late in the 16th century, demonstrated a camera obscura with a lens. Various types of camera obscura were in use by artists in the 17th century. Although the blackening of silver salts was known in the 16th century, and possibly earlier, it was not until the early 18th century that it was recognized that the darkening was caused by light and not heat. The first photograph made in a camera was achieved in 1826 by Joseph-Nicéphore Niépce, a French lithographer–inventor. He used the action of light on an asphalt solution to make his picture.

Daguerreotype and calotype

Louis-Jacques-Mandé Daguerre, a French scenic painter, in 1839 published a detailed description of the procedures of his daguerreotype process. He treated a silver-plated sheet of copper with iodine vapour to form light-sensitive silver iodide and exposed it in the camera. The image was developed by mercury vapour at 60° C (140° F) and fixed with sodium thiosulfate solution (originally sodium chloride) to remove unexposed silver iodide. The image highlights consisted of milky white mercury amalgam against a plain silver background. The daguerreotype process yielded single laterally reversed pictures and became the first commercially utilized photographic process. The original plates were 6½ × 8½ inches in size; the still current (in England) "whole plate" negative and paper formats are directly derived from Daguerre. The first daguerreotypes required exposures of about 15 to 20 minutes in bright sunlight. The daguerreotype with its unique positive images proved to be a dead end in the development of photographic technology.

The negative–positive photographic system now in use was the contribution of William Henry Fox Talbot, a British amateur scientist. In the final form (1840) of Talbot's calotype (later talbotype) process, silver iodide paper was exposed in the camera, developed in a silver nitrate and gallic acid solution, and fixed with sodium thiosulphate. Talbot printed the resulting paper negatives on a silver chloride material to yield positive pictures.

The wet-collodion process. Introduced in 1851 by the English sculptor F. Scott Archer, wet collodion was the first sensitized material of sufficient speed for instantaneous exposures in good light. A glass plate coated with a nitrocellulose (collodion) solution containing potassium iodide was immersed in the dark in a silver nitrate solution, exposed in the camera while still wet, and immediately developed (originally in ferrous sulfate or oxalate, later in pyrogallic acid) and fixed. Photographers working away from their studio had to take portable darkrooms with them for this chemical treatment. The wet-collodion plate became the principal photographic process during the following 20 years and was used by the first news and feature photographers (*e.g.*, the American Civil War records by Mathew Brady and the Crimean War coverage by Roger Fenton). Its high-quality, fine-grain negatives kept the wet-collodion process in use in the graphic-arts field until the middle of the 20th century.

Dry plates and further improvements. The invention of the dry plate carrying a silver halide emulsion in gelatin (1871) initiated sensitized-materials manufacture. With such plates the photographer was no longer obliged to coat and sensitize his own material. Dry plates could be made and stored for months and, eventually, years before use. Ready prepared emulsions were on sale by 1873 and coated plates by 1876. The first orthochromatic (green-sensitive) plates were marketed in 1884, following the discovery of spectral sensitization with dyes. Panchromatic sensitizers followed in 1904 and infrared sensitizers in 1919.

Introduction of flexible supports

With the introduction of the wet-plate process, glass became the main support for negative materials. Hannibal Goodwin, a U.S. clergyman, in 1887 proposed the use of celluloid film carrying a silver bromide emulsion; this was marketed both as roll film and as sheet film from 1889 onward by the Eastman Kodak Company and was later replaced by the less flammable acetate film (safety film). By the 1950s and 1960s, sheet films coated on high-stability supports, such as polyester, had virtually displaced glass even in special fields.

Talbot first used silver chloride printing-out papers (*i.e.*, papers that gave an image without development), while albumen printing-out paper was suggested in 1850 and development papers in 1851. Developable silver bromide gelatin papers were manufactured from 1873 onward. Various other photographic printing processes appeared after 1842—among the most recent, transfer-diffusion process (1939), xerography (1938–48), and the Polaroid one-step process (1947).

Colour photography. Alexandre Becquerel, a French physicist, and Niépce de Saint-Victor (cousin of Joseph

Niépce) between 1848 and 1860 obtained colour images in specially treated silver chloride coatings; these images could not, however, be fixed. Gabriel Lippmann, a French physicist, in 1891 produced the first permanent colour photographs based on this principle, but the procedure was too complex for general application.

James Clerk Maxwell, in Britain, demonstrated in 1859 the reproduction of colours by combining red, green, and blue light in various proportions. (Colour screen plates based on this additive principle of colour reproduction were first marketed in 1907.) In 1868 the Frenchmen L. Ducos du Hauron and C. Clos suggested subtractive colour synthesis (the superimposition of dye images) as a means of forming colour pictures; this method provided the basis of assembly colour-print processes from the 1890s onward. In 1912 chromogenic development (converting a latent silver image directly into a colour image) was proposed; the first commercial colour film with three emulsion layers, Kodachrome, based on this principle, appeared in 1935.

Camera evolution. Daguerre's camera consisted of two boxes sliding inside each other; it weighed about 50 pounds, but smaller models soon appeared. For transportability some large cameras had bellows to connect the front and rear sections; these eventually evolved into the present-day view or technical camera of the studio photographer.

Studio cameras were scaled down to hand-camera size and were fitted with improved lenses (particularly the lens designed by an Hungarian mathematician, József Miksa Petzval, in 1840) and with more versatile shutters (the first leaf-type dating from about 1887). The large-size press cameras and single-lens reflexes in use from the 1890s onward were derived from this type of hand camera.

Development of roll-film cameras

In 1875 the first roll-film camera was proposed; it used a roll of paper coated with emulsion that was stripped off after exposure. The Eastman Kodak box camera of 1888—with its handy roll of negative paper (later negative film) and with Eastman's widely available processing and printing service—gave impetus to the development of amateur photography as a hobby on a mass scale. Roll films in more or less their present form followed in 1891. Roll-film cameras, at first converted plate-type hand cameras, evolved into various folding and rigid designs.

Numerous small camera designs existed from the very beginning of photography. In 1924 the Leitz optical works in Germany introduced the Leica, the first commercially successful miniature camera taking perforated 35-mm (millimetre) film. Originally designed in 1913 by Oskar Barnack, head of the Leitz experimental department, this camera became the prototype for a large number of similar cameras and the starting point of 35-mm photographic technique, including comprehensive camera and accessory systems for general and scientific photography.

Lighting and light sources. Daylight was the photographer's only source of light until his materials became sensitive enough to permit reasonable exposures by such artificial sources as gaslight and tungsten-filament lamps. Limelight, in which a dish of lime was heated by a flaming jet of oxygen and hydrogen gases, was already in use for photography in 1840. In the 1850s it was discovered that burning magnesium wire produced an intense light. Flash powder, a mixture of magnesium powder and potassium chlorate, appeared in the 1880s, and the first prototype flash bulb in 1925.

The first high-speed flash photograph was made in 1851 by Talbot, who used the spark discharge of a Leyden jar battery. Ernst Mach, the Austrian physicist, in 1885 used sparks to take high-speed photographs of bullets in flight. Experiments with gas discharge flash tubes at the Massachusetts Institute of Technology in the early 1930s led to the present-day electronic, or speed, flash (first marketed in 1940).

BASIC TECHNOLOGY

Black-and-white photography. The most widely used photographic process is the black-and-white negative–positive system (Figure 1). In the camera, the lens projects an image of the scene being photographed onto a film coated with light-sensitive silver salts, such as silver bromide. A shutter built into the lens admits light reflected from the scene for a given time to produce an invisible but developable image in the sensitized layer, thus exposing the film.

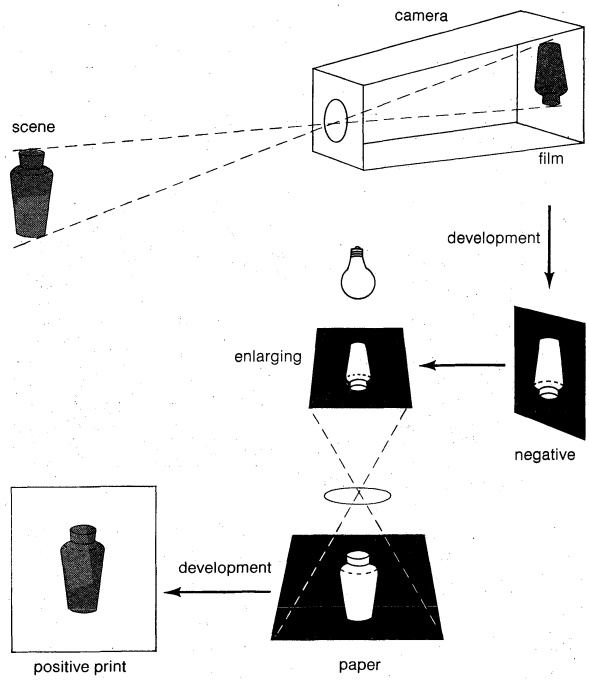


Figure 1: Sequence of negative–positive process, from photography of the original scene to enlarged print (see text).

During development (in a darkroom) the silver salt crystals that have been struck by the light are converted into metallic silver, forming a visible deposit or density. The more light that reaches a given area of the film, the more silver salt is rendered developable and the denser the silver deposit that is formed there. An image of various brightness levels thus yields a picture in which these brightnesses are tonally reversed—a negative. Bright subject details record as dark or dense areas in the developed film; dark parts of the subject record as areas of low density; *i.e.*, they have little silver. After development, the film is treated with a fixing bath that dissolves away all undeveloped silver salt and so prevents subsequent darkening of such unexposed areas. Finally, a wash removes all soluble salts from the film emulsion, leaving a permanent negative silver image within the gelatin layer.

Creation of a negative image

A positive picture is obtained by repeating this process. The usual procedure is enlargement: the negative is projected onto a sensitive paper carrying a silver halide emulsion similar to that used for the film. Exposure by the enlarger light source again yields a latent image of the negative. After a development and processing sequence the paper then bears a positive silver image. In contact printing, the negative film and the paper are placed face to face in intimate contact and exposed by diffused light shining through the negative. The dense (black) portions of the negative image result in little exposure of the paper and, hence, yield light image areas; thin portions of the negative let through more light and yield dark areas in the print, thus recording the light values of the original scene.

Colour photography. Present-day colour photographic processes rely on the fact that the eye sees colours in terms of primary stimuli of red, green, and blue to which corresponding receptors in the eye are believed to be selectively sensitive. Any colour sensation can thus be produced by a mixture of red, green and blue light in suitable proportions. Colour films record the proportion of the red, green, and blue components of colours and colour mixtures in nature by means of three sensitive layers that respond respectively to the three primary col-

ours. The images formed in the three layers are converted into coloured dye images, which in turn control the amount of red, green, and blue light left in a white light beam reflected from a colour print or shining through a colour transparency (see below *Colour photography*).

Basic camera functions. In its simplest form (Figure 2), the camera is a light-tight container carrying a lens, a

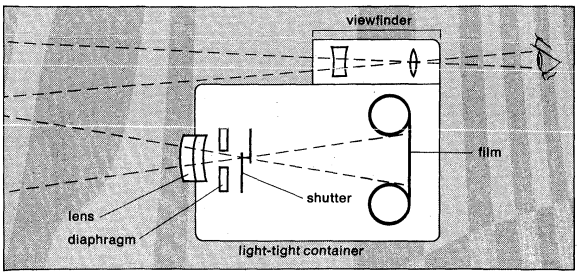


Figure 2: Basic camera elements.

shutter, a diaphragm, a device for holding (and changing) the film in the correct image plane, and a viewfinder to allow the camera to be aimed at the desired scene.

The lens projects an inverted image of the scene in front of the camera onto the film in the image plane. The image is sharp only if the film is located at a specific distance behind the lens. This distance depends on the focal length of the lens (see below *Photographic lenses*) and the distance in front of the lens of the object. As a camera is expected to be able to photograph near and far subjects, a focussing adjustment is embodied in all but the simplest cameras. Usually this adjustment functions by altering the distance between the lens and the film plane (typically, by moving the lens forward or back in the camera) to make objects at the selected distance produce a sharp image on the film. In some cameras focussing adjustment is achieved by moving only the front element of the lens, image sharpness resulting from modifying the focal length of the lens.

The two basic elements of exposure

The shutter consists of a set of metallic leaves mounted in or behind the lens or a system of blinds positioned in front of the film. It can be made to open for a predetermined time to expose the film to the image formed by the lens. The time of this exposure is one of the two factors controlling the amount of light reaching the film. The other factor is the lens diaphragm or aperture, an opening with an adjustable diameter. The combination of the diaphragm opening and exposure time is the photographic exposure, which can be expressed in terms of shutter speed and lens opening. To obtain a film image that faithfully records all the tone gradation of the object, this exposure must be matched to the brightness (luminance) of the subject and to the sensitivity or speed of the film. Many modern cameras incorporate light meters that measure the subject's luminance and indicate the required combination of shutter and diaphragm to yield a correctly exposed image. In automatic cameras the light meter may directly set the lens diaphragm or shutter or both in response to the subject luminance.

II. Cameras and lenses

CAMERA TECHNOLOGY

Principal camera types. The simplest camera type, much used by casual amateurs, has most of the features listed in the previous section—lens, shutter, viewfinder, and film-holding system. The light-tight container traditionally had a box shape. Such box cameras were usually intended for roll films; present-day compact versions frequently take instant-loading cartridges. Typically, a fixed shutter setting gives about $\frac{1}{40}$-second exposure; the aperture is fixed; and the lens is permanently set to record sharply all objects more than about six feet (two metres) from the camera. Box cameras are simple to handle, but their use is restricted to pictures of stationary or slow-moving subjects in bright daylight. More advanced versions may incorporate a fitting for flash bulbs or flash cubes for indoor picture taking or a light meter or both.

The 35-mm miniature camera. Perforated 35-mm film (originally standard motion-picture film) in cartridges holding 20 or 36 exposures with a nominal picture format of 24 × 36 mm is employed in miniature cameras. Smaller image formats—such as 24 × 24 mm, 24 × 32 mm, and 18× 24 mm (half frame)—also are available. The 35-mm camera has a comparatively fast lens (*i.e.*, with a large maximum aperture), a shutter with exposure times typically from 1 second to $\frac{1}{500}$ or $\frac{1}{1,000}$ second, and can focus on subject distances from infinity down to about three or four feet. A winding mechanism (usually a lever) advances the film from one frame to the next and, at the same time, cocks (tensions) the shutter for each exposure. An interlock prevents the shutter from being released until the film has been advanced and the film from being advanced until after an exposure. At the end of the film load, the film is rewound into the cartridge for removal from the camera in daylight.

Usually a 35-mm camera has a direct vision viewfinder, often combined with a rangefinder for accurate distance measurement. Advanced 35-mm cameras may have interchangeable lenses and a variety of other accessories. Many 35-mm cameras are single-lens reflex types (see below). Its instant readiness for action and fast shooting make the 35-mm camera popular for unobtrusive candid and news photography. In the early 1970s the 35-mm camera generally incorporated a light meter coupled with the exposure settings of the camera.

Advantages of the 35-mm camera

The ultraminiature or subminiature. This type takes roll film in widths narrower than 35 mm, either 16 mm or 9.5 mm, in special cartridges or chargers. The picture size may range from 8 × 11 mm to 12 × 17 mm. The ultraminiature is usually between matchbox and cigarette-pack size, permitting almost unobtrusive use. It has many of the features, suitably scaled down, of a 35-mm camera.

The view or technical camera. For studio and commercial photography the view or technical camera takes single exposures on sheet films (formerly plates) usually between 4 × 5 inches and 8 × 10 inches. A front standard carries interchangeable lenses and shutters; a rear standard takes a ground-glass screen (for viewing and focussing) and sheet-film holders. The standards move independently on a rail or set of rails and are connected by bellows. Both standards can also be displaced laterally and vertically relative to each other's centre and swung or tilted about horizontal and vertical axes. These features provide versatility in image control (sharpness distribution, subject distance, and perspective), though not speed in use. The view camera is nearly always mounted on a tripod or other support.

The press camera. Press cameras are loaded with sheet film (typical formats from 2½ × 3½ inches to 4 × 5 inches) for fast, hand-held shooting; they are traditionally of folding-bellows design with a lens standard on an extendable baseboard. In addition to sheet-film holders, press cameras take various types of roll film and 70-mm film magazines and often have coupled rangefinders. Modern modular designs have a rigid body with interchangeable front and rear units.

The folding roll-film camera. Obsolescence has overtaken the folding roll-film camera. Its design resembles the 35-mm miniature camera in shutter and viewfinder equipment, but such a camera has bellows and folds up to pocketable size when not in use. Generally it takes roll films holding eight to 16 exposures; typical picture sizes are 2¼ × 2¼, 2¼ × 3¼ or 1¾ × 2¼ inches. Some 35-mm cameras were also produced with bellows.

The single-lens reflex. The ground-glass screen at the back of the studio, or view, camera slows down picture taking because the screen must be replaced by the film for an exposure. The single-lens reflex camera (Figure 3) has a screen, but the film remains constantly in position. A 45° mirror reflects the image-forming rays from the lens onto a screen in the camera top. The mirror moves out of the way during the exposure and back again afterwards for viewing and focussing the next picture. The image on the screen therefore temporarily disappears from view during the exposure. Present-day single-lens reflexes are either 35-mm cameras or advanced roll-film models.

Reflecting the image through the lens

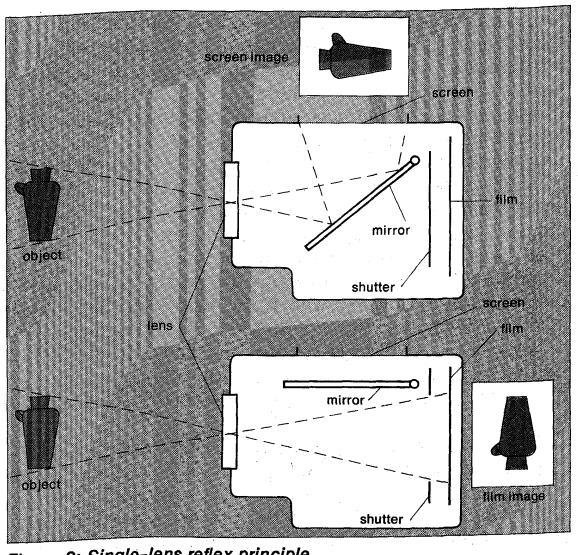


Figure 3: *Single-lens reflex principle.* (Top) Before exposure, upright but laterally reversed image is visible on top screen. (Bottom) During exposure, with mirror up and shutter open, an upside down but laterally correct image is projected on the film.

Most 35-mm reflexes have optical prism systems for eye-level viewing of the screen, built-in light meters, and numerous other refinements. Often the camera is part of an extensive accessory system.

The twin-lens reflex. The twin-lens reflex is a comparatively bulky dual camera (Figure 4) with a fixed-mirror

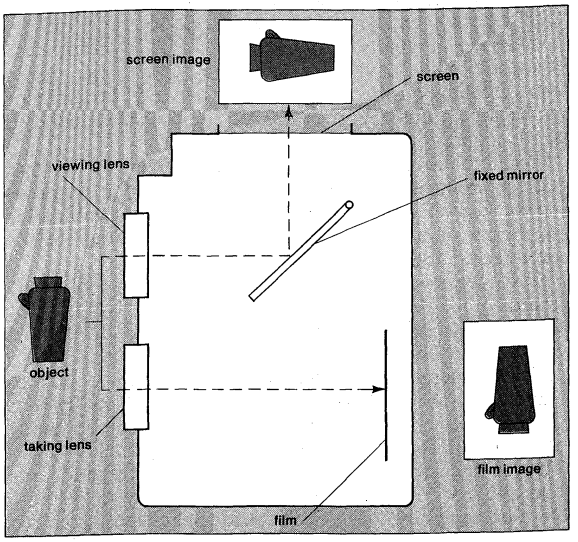


Figure 4: *Principles of the twin-lens reflex camera.* The fixed mirror deflects light rays coming through the viewing lens to a top screen, which shows the image upright but laterally reversed. Light from the object also goes through the taking lens, which is mounted on a common panel with the viewing lens, and is projected on the film.

reflex housing and top screen mounted above a roll-film box camera. Its two lenses focus in unison so that the top screen shows the image sharpness and framing as recorded on the film in the lower section. The viewing image remains visible all the time, but the viewpoint difference (parallax) of the two lenses means the framing on the top screen is not exactly identical with that on the film.

Shutter and diaphragm systems. Principal present-day shutters are the leaf shutter and the focal-plane shutter.

The leaf shutter. The leaf, or diaphragm, shutter consists of a series of blades or leaves fitted inside or just behind the lens. The shutter opens by swinging the leaves simultaneously outward to uncover the lens opening. The leaves stay open for a fixed time—the exposure time—and then close again. A spring drives the mechanism while an adjustable escapement controls the open time—typically between one and $\frac{1}{500}$ second. For longer exposures (time exposures) a special setting holds the shutter open as long as the exposure button (release) is held down; the shutter closes on letting go of the release. The mechanical escapement mechanism can be replaced by an electromagnet and an electronic circuit to control the length of time the shutter remains open.

The control of exposure time

Focal-plane shutter. The focal-plane shutter consists of two light-tight fabric blinds or a combination of metal blinds (Figure 5) moving in succession across the film

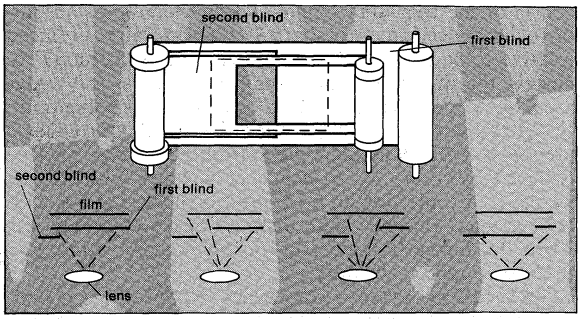


Figure 5: Focal-plane shutter consisting essentially of (top) two blinds that, moving in sequence (bottom), form a travelling slit (see text).

immediately in front of the image plane. The first blind uncovers the film and the second blind covers it up again, the two blinds forming a travelling slit the width of which determines the exposure time: the narrower the slit, the shorter the time. The actual travel time is fairly constant for all exposure times. A mechanism triggers the release of the second blind. Focal-plane shutters are usually adjustable for exposure times between one and $\frac{1}{1.000}$ second.

Diaphragm and shutter settings. In the lens diaphragm, a series of leaves increases or decreases the opening to control the light passing through the lens to the film. The diaphragm control ring carries a scale of so-called *f*-numbers, or stop numbers, in a series: such as 1.4, 2, 2.8, 4, 5.6, 8, 11, 16, 22, and 32. The *f*-numbers are inversely proportional to the square of the amount of light admitted. In the above series, each setting admits twice as much light as the next higher *f*-number, or stop (giving twice as much exposure), and half as much as the next lower value.

The series of *f*-numbers

Shutter settings on present-day cameras also follow a standard double-or-half sequence—*e.g.*, 1, ½, ¼, ⅛, $\frac{1}{15}$, $\frac{1}{30}$, $\frac{1}{60}$, $\frac{1}{125}$, $\frac{1}{250}$, $\frac{1}{500}$, and $\frac{1}{1.000}$ second. The shorter the exposure time, the "faster" the shutter speed.

Exposure values. The diaphragm and shutter control the amount of light reaching the film by adjusting the light intensity and the time during which it acts. Different aperture–shutter speed combinations may thus yield the same exposure—*e.g.*, diaphragm *f*/2.8 with $\frac{1}{250}$ second, *f*/4 with $\frac{1}{125}$ second, and *f*/5.6 with $\frac{1}{60}$ second. All such equivalent combinations can be assigned a single exposure value. Some cameras carry a scale of exposure values and have coupled aperture and shutter-speed settings. At a given exposure-value setting in such cameras, selecting various shutter speeds automatically adjusts the aperture to compensate and vice versa.

Exposure values run in a simple whole-numbered series, each step (EV interval) doubling or halving the effective exposure. The lower the EV number, the greater the exposure. Thus, EV 10 gives twice as much exposure as EV 11 or half as much as EV 9.

Automatic-diaphragm systems. On a camera with a ground-glass screen, such as a single-lens camera, a view camera, or reflex camera, viewing and focussing are carried out with the lens diaphragm fully open, but the exposure is often made at a smaller aperture. With a view camera the aperture must be reduced manually; but with many reflex cameras the lens aperture is semi-automatically or automatically reduced (stopped down) to the working aperture immediately before exposure. Typically, this is accomplished by an internal link between the release button and the lens.

Methods of focussing and framing. The ground-glass (sometimes grained plastic) screen is the most direct way of viewing the image for framing and for sharpness control. The screen localizes the image plane for observation. The image is also visible without a screen, but then the eye can locate the image plane of maximum sharpness only with a precisely focussed high-power magnifier. This aerial focussing method avoids interference of the ground-glass structure with sharpness assessment.

Focussing aids. The split-image wedge or wedge rangefinder consists of a pair of prism wedges fitted in the centre of the focussing screen. This system splits an out-of-focus image into two sharp halves laterally displaced relative to each other. When the camera is correctly focussed, the image becomes continuous across the wedge area. Since the eye can recognize slight image discontinuity better than it does slight unsharpness, this provides a precise way of focussing. The split-image wedge may be replaced by a ¼- to ½-inch spot covered with minute wedges. This microprism grid shows a blurred image with very ragged outlines and distinct microprism texture; when the image becomes sharp, the ragged outline and texture disappear abruptly.

Microprism grid of multiple wedges

The focussing screen is sometimes overlaid by a pattern of fine concentric lens sections. Called a Fresnel screen, it redirects the light from the screen corners toward the observer's eye and makes the image evenly bright.

Cameras without a screen generally are equipped with a distance scale, the lens being set to the estimated object distance. More advanced cameras have an optical rangefinder (Figure 6) as a distance-measuring aid; it consists

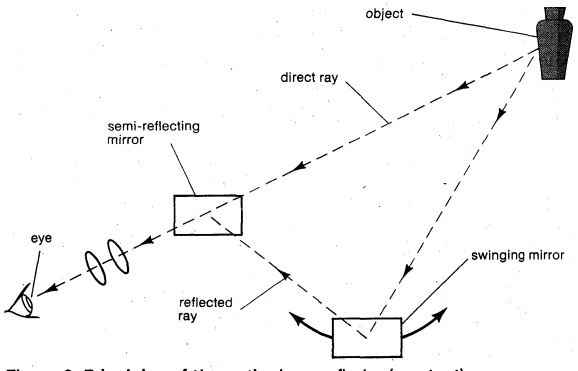


Figure 6: Principles of the optical rangefinder (see text).

of a viewfinder (see below) and a swinging mirror a few inches to one side of the viewfinder axis. As the eye views an image of the object, the mirror superimposes a second image from a second viewpoint. On turning the mirror through the correct angle, which depends on the object distance, the two images are made to coincide. The mirror movement can be linked with a distance scale, or coupled with the lens focussing adjustment. When the lens is incorrectly focussed, the rangefinder shows a double or split image. In place of a rotating mirror, the rangefinder may use swinging or rotating optical wedges (prisms).

Fully automatic focussing systems rely on photoelectric-cell circuits, which detect the contrast difference between a sharp and an unsharp image, thus eliminating visual judgment of image sharpness or continuity. The signal output from such a photocell system can power a small motor that automatically adjusts the lens setting to maximum sharpness. So far, such systems have been used only in professional motion-picture cameras. Another approach to automatic rangefinding employs extremely short laser pulses as a kind of optical radar—a system that requires considerable auxiliary equipment.

Viewfinders. The sighting devices in cameras lacking screens are called viewfinders; they show how much of the scene will appear on the film. The simplest viewfinder is a wire frame above the camera front, with a second frame near the back to aid the eye in correct centring. Most present-day finders are built into the camera and are compact lens systems. Bright-frame finders show a white frame reflected into the view to outline the field recorded on the film. An alternative form is the reflecting viewfinder in which the photographer looks down into a field lens on top of the camera. The upper section of a twin-lens reflex camera is such a reflecting finder, with provision for sharpness control (see above).

As the viewfinder axis in a camera other than a single-lens reflex does not usually coincide with the lens axis, the finder's and the lens's views do not exactly match. This parallax error is insignificant with distant subjects; with near ones it is responsible for the familiar fault of a portrait shot of a head that appears partly cut off in the picture even though it was fully visible in the finder. Camera viewfinders come with various parallax-compensating devices, such as correction marks to show the limits of the view at near range, or embody a system that moves the finder frame in unison with the distance settings.

Parallax problems with close-at-hand subjects

The optical finder gives a direct upright and right-reading view of the subject with the camera held at eye level. The traditional reflex camera, held at waist level, showed a laterally reversed view. Modern reflexes have a pentaprism arrangement that permits upright, right-reading, eye-level viewing by redirecting the image from the horizontal screen on top of the camera.

Exposure-metering systems. Exposure meters or light meters measure the light in a scene to establish optimum camera settings for correct exposures. A light-sensitive cell generates or controls an electric current according to the amount of light reaching the cell. A microammeter measures this current to indicate exposure settings. The two main types, the selenium cell and the cadmium sulfide cell, differ in that light falling on a selenium cell generates a current, while light falling on a cadmium sulfide cell changes its electrical resistance.

On separate exposure meters the values read off are transferred manually to the camera. Many modern cameras have built-in, coupled exposure meters; the operation of lining up the needle of the meter with an index at the same time sets aperture or shutter speed controls of the camera, or both.

In automatic exposure systems, the meter on its own sets the lens aperture or exposure time, or both. Or, pressing the shutter release may set the exposure combination by means of a selector mechanism immediately before the exposure itself.

On cameras with electronic shutter control, often a cadmium sulfide cell measures the light cumulatively during the exposure and closes the shutter as soon as the cell has registered a sufficient amount of light.

Measurements of the screen image brightness in a reflex camera can also provide exposure readings. Usually such through-the-lens meter systems cover the screen image with one or two photocells in the pentaprism housing. The readings depend on the light coming through the lens and so allow for the lens's angle of view, any light filters, close-up exposure corrections, stray light, and other factors.

Through-the-lens meter systems

In other systems, some of the light from the lens is diverted to a photocell before it reaches the screen (*e.g.*, by beam-splitting arrangements or through the use of a photocell behind an only partly reflecting mirror). Alternatively, with a focal-plane shutter camera a photocell may be mounted on an arm in front of the image plane and be swung out of the way while the picture is being taken.

Flash systems. Flash is a widely used artificial light source for photography, providing a reproducible light of high intensity and short duration. It can be synchronized with an instantaneous exposure. Being battery powered, small flash units are self-contained. The two main systems are flash bulbs and electronic flash.

Flash bulbs. The flash bulb consists of a small, oxygen-filled glass envelope containing two sealed-in wires with a filament, a firing paste, and a specific amount of aluminum or zirconium wire. A pulse of current through the lead-in wires fires the flash, making the filling wire burn away virtually instantaneously. The firing system is built into a flash gun, which consists of a small casing

holding a battery circuit, a reflector, and the fitting to take the bulb. Flash bulbs come in various sizes, from about ½ × 1¼ inches to the size of a 60-watt tungsten bulb.

So-called flash cubes consist of four miniature flash bulbs mounted in small reflectors on four sides of a one-inch cube. They plug into a rotating fitting in the camera, which has the necessary firing circuit and battery built-in. After every shot, the cube rotates to point a fresh bulb at the subject. Some flash cubes are triggered mechanically and need no batteries.

Rotating flash cubes

Electronic flash. Electronic flash is a high-voltage discharge through a gas-filled tube; but, unlike a flash bulb, it can flash any number of times. A capacitor charged up to several hundred volts (by a step-up circuit from low-voltage batteries or from the line voltage supply) provides the discharge energy. The flash is triggered by a low-voltage circuit generating a high-voltage pulse. The light output of an electronic flash is comparable with that of a flash bulb, but its duration is much shorter (typically, $\frac{1}{1,000}$ second compared with $\frac{1}{50}$ second for a bulb), while the maximum flash intensity is correspondingly higher. The smallest electronic flash units clip onto the camera and are about cigarette-pack size. Large units with flood and spotlight fittings are used in professional studio photography and often have adjustable power levels to provide different flash intensities.

Special electronic flash units with ring-shaped tubes mounted around the camera lens are used for shadowless lighting on small objects in close-up photography, particularly for medical and forensic purposes.

Firing and synchronization. Flash guns and units have a firing button for the triggering circuit but are usually fired by a switch in the camera shutter to synchronize the flash with the shutter action. A flash lead or a contact in the camera's accessory shoe connects the gun with this shutter switch.

Usually the shutter contact closes the instant the shutter is fully open. The shutter must then stay open while the flash reaches its peak brightness—$\frac{1}{60}$ to $\frac{1}{50}$ second with flash bulbs or 15 to 20 microseconds (millionths of a second) with electronic flash. Those limits determine the fastest usable shutter speed. On some shutters the flash is fired before the shutter opens, to allow for the firing delay of flash bulbs. With focal-plane shutters the transport time of the blinds affects the synchronizable shutter speeds.

System cameras. From the development of the 35-mm miniature camera in the 1930s evolved the concept of the system camera that could be adapted to numerous jobs with a range of interchangeable components and specialized accessories. Today, most moderately advanced 35-mm miniatures take interchangeable lenses, close-up and photomicrographic attachments, filters, flash units and other accessories. The most elaborate camera systems also include such accessories as alternative finder systems; interchangeable reflex screens, film backs, and magazines; and remote-control and motor-drive systems. Modular professional roll-film and view cameras are built up from a selection of alternative camera bodies, film backs, bellows units, lenses, and shutters. This is the nearest approach to the universal camera, assembled as required to deal with practically every type of photography.

Modular systems

PHOTOGRAPHIC LENSES

Characteristics and parameters. The lens forming an image in the camera is a converging lens, the simplest form of which is a single biconvex (lentil-shaped) element. (The word lens is derived from the Latin name of the lentil.) In theory, such a lens makes a light beam of parallel rays converge to a point (the focus) behind the lens. The distance of this focus from the lens itself is the focal length, which depends on the curvature of the lens surfaces and the optical properties of the lens glass. An object at a very long distance (optically regarded as at "infinity") in front of the lens forms an inverted image in a plane (the focal plane) going through the focus. Light rays from nearer objects form an image in a plane behind the focal plane. The nearer the object, the farther behind the lens the corresponding image plane is located—which is why a lens has to be focussed to get sharp images of objects at different distances.

Focal length and image scale. The image scale or scale of reproduction is the ratio of the image size to the object size; it is often quoted as a magnification. When the image is smaller than the object, the magnification of the object is less than one. If the image is $\frac{1}{20}$ the size of the object, for example, the magnification may be expressed either as 0.05 or as 1:20.

For an object at a given distance, the scale of the image depends on the focal length of the lens (Figure 7). A

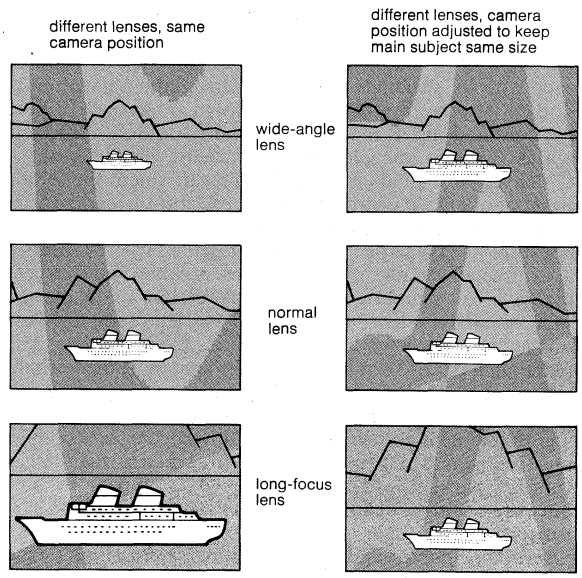


Figure 7: Effects of using lenses of different focal lengths.

normal camera lens usually has a focal length approximately equal to the diagonal of the picture format covered. A lens of longer focal length will give a larger scale image, but necessarily covers less of the scene in front of the camera. Conversely, a lens of shorter focal length yields an image on a smaller scale but—provided the angle of coverage is sufficient (see below)—takes in more of the scene. Many cameras, therefore, can be fitted with interchangeable lenses of different focal lengths to allow varying the image scale and field covered: long-focus lenses for magnified views of distant detail and short-focus, wide-angle lenses to cover wider views from nearer distances. The focal length of a lens in millimetres (sometimes in inches) is generally engraved on the lens mount.

Aperture. The aperture, or *f*-number, is the ratio of the focal length to the diameter of an incident light beam as it reaches the lens. For instance, if the focal length is 50 mm and the diameter of the incident light beam is 25 mm, the *f*-number is 2. This incident beam diameter is often roughly the lens-diaphragm diameter, but with some lens systems it may be appreciably larger or smaller. The maximum aperture (*f*-number at the largest diaphragm opening) is also marked on the lens, usually in the form *f*:2, *f*/2, or 1:2.

Angle of coverage. A lens must cover the area of a camera's film format to yield an image adequately sharp and with reasonably even brightness from the centre to the corners of the film. A normal lens should cover an angle of at least 60°. A wide-angle lens covers a greater angle—about 70° to 90° or more for an ultrawide-angle lens. A long-focus lens covers a smaller angle; for instance, the diagonal of a 24 × 36 mm picture format only subtends a 28° angle with a 135-mm lens.

The angle of coverage depends on the lens design. Designations like "wide angle" or "narrow angle" are not necessarily synonymous with "short focus" and "long focus," as the latter terms refer to the focal length of the lens relative to the picture format.

Optical performance. A simple lens produces a very imperfect image, which is usually blurred away from the centre. The image may have colour fringes around object outlines, and straight lines may be distorted. Such defects, called aberrations, can be eliminated—and even then not completely—only by replacing the single lens element by a group of elements of appropriate shape and separation. Aberrations arising from some of the lens elements then counteract opposite aberrations produced by other elements. The larger the maximum aperture, the greater the angle of coverage and the higher the degree of correction aimed at, the more complex camera lenses become. Lens design for relative freedom from aberrations nowadays involves advanced computer programming to calculate the necessary geometrical parameters of every lens element.

Groups of lens elements for correcting aberrations

Aberrations. There are a number of lens aberrations, each with its own characteristics. Chromatic aberration is said to be present when the lens forms images by different-coloured light in different planes and at different scales. Colour-corrected lenses largely eliminate these faults. Spherical aberration is present when the outer parts of a lens do not bring light rays into the same focus as the central part. Images formed by the lens at large aperture are therefore unsharp but get sharper at smaller apertures. Curvature of field is present when the sharpest image is formed not on a flat plane but on a curved surface. Astigmatism occurs when the lens fails to focus image lines running in different directions in the same plane; in a picture of a rail fence, for instance, the vertical posts are sharp at a focus setting different from the horizontal rails. Another aberration, called coma, makes image points near the edges of the film appear as irregular unsharp shapes. Distortion is present when straight lines running parallel with the picture edges appear to bow outwards (barrel distortion) or inwards (pincushion distortion).

Resolving power and contrast transfer function. One way of testing lens performance is to observe the image it forms of patterns of increasingly closely spaced black lines separated by white spaces of line width. The closest spacing still recognizable in the image gives a resolving power value, expressed in lines per millimetre. Photographs of such line patterns or test targets show resolving power of the lens and film combination. For example, a resolution of 80–100 lines per millimetre on a fine-grain film represents very good performance for a normal miniature camera lens.

The visual sharpness of an image depends also on its contrast. Opticians, therefore, often plot the contrast with which the image is reproduced against the line spacing of that image. The resulting contrast transfer curve, or function, gives a more reliable indication of the lens performance under practical picture-taking conditions.

Special lens types. Apart from general purpose camera lenses of various focal lengths, there are lenses of special characteristics or design.

Telephoto lenses. Long-focus lenses are bulky, because they comprise not only the lens itself but also a mount or tube to hold it at the appropriate focal distance from the film. Telephoto lenses are more compact; their combinations of lens groups make the back focus (the distance from the rear lens element to the film) as well as the length of the whole lens appreciably shorter than the focal length. Telephoto lenses are therefore popular for interchangeable-lens cameras. Strictly, the term telephoto applies only to a lens of this optically reduced length; in practice, long-focus lenses of all types tend to be called indiscriminately telephoto or "tele" lenses.

Tele-converter lenses

If a camera lens is interchangeable, an accessory teleconverter lens group can be positioned between the prime lens and the camera. This turns a normal lens into an even more compact telephoto system, which is less costly than a telephoto lens but which greatly reduces the speed of the prime lens and usually impairs sharpness performance.

Wide-angle and retrofocus lenses. Short-focus, wide-angle lenses are usually mounted close to the film. Single-lens reflex cameras need a certain minimum lens-to-film distance to accommodate the swinging mirror. Wide-angle (and sometimes normal-focus) lenses for such cameras therefore use so-called retrofocus designs—in effect, inverted telephoto systems. In these, the back focus is appreciably longer than the focal length. In practice, both a telephoto and a retrofocus lens must be specially designed for its particular use to ensure optimum image performance.

Fish-eye lenses. For image angles greater than 110°, it becomes difficult to bring the lens close enough to the film to allow the rays between lens and film to diverge sufficiently. The fish-eye lens overcomes this difficulty by making the rays diverge less behind the lens than they do in front. The resulting image shows appreciable distortion with image details near the edges and corners progressively compressed. Fish-eye lenses usually cover angles between 140° and 210° and are used for unusual wide-angle effects where the distortion becomes a deliberate pictorial element. They also have certain scientific applications, for instance, to cover a horizon-to-horizon view of the sky in recording cloud formations.

Mirror lenses. Images can also be formed by light reflected from curved mirrors. This method, long used in astronomical telescopes, is applied to long-focus lens systems of short overall length by folding the light path back onto itself. A mirror lens or catadioptric system has no chromatic aberrations. Other aberrations are corrected by incorporating one or more appropriate lens elements. The arrangement of the system, with a central opening in the primary mirror, makes stopping down with a customary diaphragm difficult. Catadioptric systems therefore generally use neutral-density filters to control light transmission.

Variable-focus lenses. In variable-focus lenses, the focal length can be varied by movement of some of the elements or groups within the lens system. One lens can thus replace a range of interchangeable lenses. Such a lens also allows instant assessment of the pictorial effect of using different focal lengths.

The variable-focus, or zoom, lens was originally developed for motion-picture photography, in which adjustment of the focal length during a shot produced a zooming-in or zooming-out effect (hence the name). It is increasingly used in still cameras, especially 35-mm single-lens reflexes since the reflex finder permits accurate continuous assessment of image coverage. During zooming, the image changes in scale but not in sharpness. To ensure reasonably constant performance over their focal-length ranges, zoom lenses are complex systems containing from a dozen to 20 elements. Zoom lenses for still cameras usually have focal-length ratios of 2:1 to 3:1 (*e.g.*, 40–120 mm for a 35-mm reflex), occasionally up to 6:1.

Lens-changing systems. Miniature and roll-film cameras hold interchangeable lenses in screw or quick-change bayonet mounts. In a focal-plane shutter camera the usable range of focal lengths is practically unlimited. In cameras with leaf shutters, the lens is either mounted in front of the shutter or the lens is changed complete with shutter. Some designs use convertible lenses with the rear components built into the camera together with the shutter; interchangeable front groups then provide different focal lengths in combination with the fixed rear group. View-camera lenses—usually with their own shutters—are mounted on lens boards that clip into and out of the front camera standard.

Convertible lens with fixed rear component

Afocal attachments provide the effect of alternative focal lengths with a fixed camera lens. They are magnifying or reducing telescopes without a focal length (hence afocal), yielding a virtual image that the camera lens projects onto the film. Their designated magnification factor indicates the effect on the image scale; *e.g.*, a 1.5× tele attachment magnifies the image on the film 1½ times, while a 0.7× wide-angle attachment reduces the image scale to 0.7 times that of the prime camera lens. There are also afocal zoom and afocal fish-eye attachments.

Lens coating. When light passes from one optical medium to another (especially from air to glass and vice versa in a lens), about 4 to 8 percent of it is lost by

reflection at the interface. This light loss builds up appreciably in complex multi-element lenses. Some of the reflected light still reaches the film as ghost images or light spots or as general contrast-reducing scattered light.

To reduce such losses, the air to glass surfaces of modern lenses typically carry a microscopically thin coating of metallic fluorides. The coating eliminates most reflected rays. Complete elimination can occur only for light of one wavelength if the coating thickness and refractive index are exactly right. In practice, a coated lens surface reflects about 0.5 percent of incident white light—1/10 of the light lost by an uncoated lens. Multiple coatings can reduce reflections over a wider wavelength range.

Lens coating has made possible modern optical designs of a dozen or so lens groups (20 or more air–glass surfaces); with uncoated lenses the limit was around ten surfaces, which cost a light loss of nearly 60 percent.

III. The photographic process

BLACK-AND-WHITE FILMS

The latent image. The sensitive surface of ordinary film is a layer of gelatin carrying minute suspended silver halide crystals or grains (the emulsion)—typically silver bromide with some silver iodide. Exposure to light in a camera produces an invisible change yielding a latent image, distinguishable from unexposed silver halide only by its ability to be reduced to metallic silver by certain developing agents.

Formation of the latent image

Current theories postulate that silver halide crystals carry minute specks of metallic silver—so-called sensitivity specks—which amount in mass to about 1/100,000,000 part of the silver halide crystal. A silver halide is a compound of silver with fluorine, chlorine, bromine, or iodine, but only the last three are light sensitive. When light action releases electrons from the silver halide crystal, they migrate to the sensitivity specks. The resulting electric charge on the specks attracts silver ions from the neighbouring silver halide; and as the silver ions accumulate, they become metallic silver, causing the speck to grow. Halide (*e.g.*, bromide) ions at the same time migrate to the surface of the silver halide crystal and are there absorbed by the gelatin of the emulsion. When the sensitivity speck is large enough, it provides a point of attack for the developer, which can then reduce the whole silver halide crystal to silver. Developers are selective organic reducing agents that attack only silver halide crystals that have sufficiently large sensitivity specks. The halide grains carrying a developable sensitivity speck make up the latent image.

Sensitometry and speed. The sensitivity or speed of a film determines how much light it needs to produce a given amount of silver on development. Sensitometry is the science of measuring this sensitivity, which is determined by giving the material a series of graduated exposures in an appropriate instrument (the sensitometer). After development under specified conditions, the density of the silver deposit produced by each exposure is measured and the densities are plotted on a graph against the logarithm of the exposure. The resulting characteristic curve, or $D/\log E$ curve (see below, *Contrast*), shows how the film reacts to exposure changes. A specified point on the curve also serves as a criterion for calculating film speed by methods laid down in various national and international standards.

The most widely used speed scale is ASA speed. It is arithmetic, with the speed number directly proportional to the sensitivity. A film of 200 ASA is twice as fast (and for a given subject requires half as much exposure) as a film of 100 ASA. Some speed scales, such as the DIN system used on the European continent, are logarithmic. Every increase of three in the DIN speed number doubles the actual speed. A 50 ASA film is equivalent to 18 DIN, 100 ASA to 21 DIN, and so on.

All-around films for outdoor and some indoor photography have speeds between 80 and 160 ASA; fine-grain films for maximum image definition between 20 and 64 ASA; high-speed films for poor light and indoor photography between 200 and 500 ASA; and ultraspeed films for very low light conditions from 650 ASA up.

Colour sensitivity. Initially, the silver halide emulsion is sensitive to ultraviolet radiation and to violet and blue light. Most films contain sensitizing dyes to extend their colour sensitivity through the whole visible spectrum. Such films (called panchromatic films) record subject colour values as gray tones largely corresponding to the visual brightness of the colours. Some high-speed panchromatic films are extra red-sensitive for photography by red-rich tungsten lighting.

Non-colour-sensitized or blue-sensitive emulsions (without sensitizing dyes) are used for copying monochrome originals and similar applications needing no extended colour sensitivity. At one time, orthochromatic films—sensitive to violet, blue, green, and yellow but not to red—were also used for general photography; nowadays they are employed mainly for photographing of phosphor screens, such as cathode-ray tubes, and for other purposes requiring green but not red sensitivity. Both non-colour-sensitized and orthochromatic films can be handled in the light of appropriately coloured darkroom safelights.

Use of orthochromatic films

Infrared films are sensitized to invisible infrared wavelengths. They are used in aerial photography to cut through atmospheric haze (which scatters blue light but not infrared rays) and for special purposes in scientific and forensic photography.

Filters. Filters can modify the way in which a film records colours as monochrome tone values. They are disks of coloured glass or gelatin with controlled transmission characteristics. Placed in front of the camera lens, they preferentially transmit light of their own colour and hold back light in other colours. A yellow or yellow-green filter is often used in landscape photography to prevent overexposure of the blue sky and to bring out detail in cloud formations. Orange and red filters make the sky still darker and cut through haze by absorbing scattered blue light.

Contrast filters differentiate between the gray values of objects of different colour but of similar brightness. For instance, a red flower and green foliage record in similar shades of midgray. A red filter holds back green light to darken the green foliage, making the flower lighter; a green filter absorbs red light, thus darkening the flower. Such deliberate tone distortion is widely used in photomicrography and other fields.

Other filter types used in photography include ultraviolet, infrared, and polarizing filters. Ultraviolet-absorbing filters screen out ultraviolet rays at high altitudes (*e.g.*, in mountain photography). Since camera lenses are not normally corrected for such rays, they can reduce image sharpness, even though the lenses allow only a small amount of ultraviolet to be transmitted. Infrared filters are used with infrared film to hold back visible light. Polarizing filters polarize light and can absorb polarized light if suitably oriented. Light reflected at certain angles from shiny surfaces of nonmetallic media (glass, water, varnish) is polarized; a properly oriented polarizing filter subdues such reflections in a picture. Such filters are also employed in scientific photography.

Since a filter screens out part of the light, its use calls for extra exposure, the amount of which is indicated by a filter factor—*e.g.*, 2×, which means the exposure time must be multiplied by 2. For cameras with an exposure-value scale, a filter may specify an exposure value reduction (−1, −1½, etc.; *i.e.*, the indicated exposure value must be reduced by this amount). The correct filter factor depends—apart from the filter colour—on the spectral sensitivity of the film, the colour quality of the lighting, the type of subject and effect aimed at, and other exposure conditions. Factors published by filter manufacturers usually apply to average daylight conditions and panchromatic films.

Filter factors for proper exposure

Other film characteristics. Of practical interest to the photographer are the graininess, resolving power, and contrast of a film. Although they are characteristics of the film itself, they are influenced by the conditions of development (see below *Processing and printing*).

Grain. The image derived from minute silver halide crystals is discontinuous in structure. This gives an ap-

pearance of graininess in big enlargements. The effect is most prominent with fast films, which have comparatively large silver halide crystals; slower, fine-grain emulsions are therefore preferable for miniature and ultraminiature negatives intended for considerable enlargement.

Resolving power and acutance. The fineness of detail that a film can resolve depends not only on its graininess but also on the light scatter or irradiation within the emulsion (which tends to spread image details) and on the contrast with which the film reproduces fine detail. These effects can be measured physically to give an acutance value—today preferred to resolving power as a criterion of a film's sharpness performance. Fine-grain films with thin emulsions yield the highest acutance.

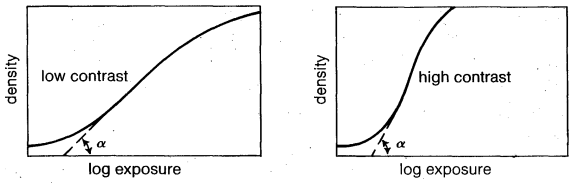


Figure 8: Characteristic curves of low-contrast and high-contrast film (see text).

Contrast. High-contrast films reproduce tone differences in the subject as great density differences in the image; low-contrast films translate tone differences into small density differences. The characteristic curve of a film obtained by plotting the density against the logarithm of the exposure (mentioned earlier under *Sensitometry and speed*) can be used to express a film's contrast (see Figure 8). The slope of the straight-line section of the curve (sometimes called the gamma, actually the tangent of the angle α) indicates contrast: the steeper the slope, the higher the contrast rendering. General-purpose films yield medium contrast (gamma 0.7 to 1). High-contrast films (gamma 1.5 to 10) are used for copying line originals and other specialized purposes; low-contrast films for continuous-tone reproduction. Gamma is also used to indicate degree of development, since increased development, with a given emulsion, results in a higher gamma.

Film structure and forms. Viewed in magnified section (Figure 9), a film consists of a number of layers and

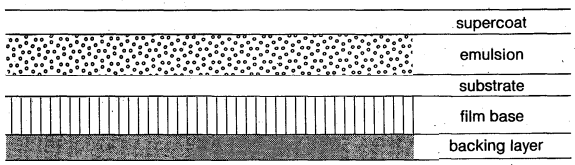


Figure 9: Structure of a modern black-and-white film (see text; thicknesses not to scale).

components: (1) A supercoat of gelatin, a few microns (one micron is 0.001 millimetre) thick, protects the emulsion from scratches and abrasion marks. (Pressure and rubbing can produce developable silver densities.) (2) The emulsion layer (silver halide suspended in gelatin) is usually nine to 12 microns (up to $\frac{1}{2,000}$ inch) thick but may sometimes reach 25 microns. (3) A substrate or subbing layer promotes adhesion of the emulsion to the film base. (4) The film base or support is usually cellulose triacetate or a related polymer. The thickness may range from 0.08 to 0.2 millimetre (0.003 to 0.008 inch). Films for graphic arts and scientific purposes are often coated on a polyethylene terephthalate or other polyester support of high dimensional stability. Glass plates—once the commonest support for negative materials—are now used only for applications requiring extreme emulsion flatness. (5) A backing layer on the rear of the film base counteracts curling. Usually it contains also a nearly opaque dye to suppress light reflection on the rear support surface. Such reflection (halation) reduces definition by causing halolike effects around very bright image points. Some film bases (especially in 35-mm films) are tinted gray to absorb light that has passed through the emulsion layer.

Use of anti-halation dye

Sheet film. Press and studio cameras generally take sheet film—single sheets (typical sizes range between 2½ × 3½ and 8 × 10 inches) loaded in the darkroom into light-tight film holders for subsequent insertion in the camera.

Perforated film. Some film is perforated along its edges and rolled up on its own inside a light-tight cartridge, which can be loaded into the camera in daylight. Once the camera is closed, a transport sprocket engaging the edge perforations draws the film from the cartridge onto a spool and advances it from picture to picture. The commonest film width is 35 mm (for 35-mm miniature cameras), and its cartridge typically holds enough film for 36 or 20 exposures. A 70-mm-wide film for larger cameras and 16-mm-wide strips for ultraminiatures are packed and used in a similar way.

Roll film. The term roll film is usually reserved for film wound up on a spool with an interleaving light-tight backing paper to protect the wound-up film. The spool is loaded into the camera in daylight, the backing paper leader threaded to a second spool, and the film wound from picture to picture once the camera is closed. This is the classical roll film of roll-film cameras. Common current film widths are 62 mm ($2\frac{7}{16}$ inches) and 45 mm ($1\frac{13}{16}$ inches). The rear of the backing paper carries sets of consecutive numbers spaced at frame intervals for different image formats. In some roll-film cameras these numbers are visible through a window in the camera back and show how far the film must be wound from one picture to the next. Instant-loading cartridges also use paper-backed roll film.

PICTURE-TAKING TECHNIQUE

The main areas of practical camera handling in photography concern sharpness control, exposure, and lighting.

Sharpness control. The image on the film is sharpest when the lens is focussed to the exact object distance. Usually, however, a scene includes objects at varying distances from the camera. Various factors affect the sharpness distribution in a picture of such a scene.

Depth of field. The sharpness in the image of objects in front of and behind the focussed distance falls off gradually. Within a certain range of object distances this sharpness loss is still comparatively unnoticeable. This range is the depth of field and depends on: (1) the amount of sharpness loss regarded as acceptable: miniature negatives requiring big enlargement must be sharper than larger-format negatives, which are enlarged less; (2) the lens aperture used: stopping down the lens (higher *f*-numbers) increases the depth of field; (3) the object distance: the depth of field is smaller for near objects than for more distant ones; and (4) the focal length of the lens: depth of field is reduced with longer focus lenses (and with larger picture formats requiring lenses of longer focal length); the depth increases with shorter focus lenses. A depth of field indicator, often included on the focussing mounts of lenses, shows on the distance scale how far in front of and behind the focussed distance objects will be in focus at different diaphragm openings.

Factors influencing depth of field

Subject and camera movement. Movement of the subject while the camera shutter is open for the exposure leads to a blurred image. The exposure time must therefore be short enough to keep the blur within acceptable limits. The shutter speed required depends on the movement speed of the object, the scale of the image (movement blur gets greater the nearer the subject or the longer the focal length of the lens used) and the movement direction; movement across the direction of view produces more blurring than movement receding from or approaching the camera.

Movement blur can be reduced, even with comparatively slow shutter speeds, by moving the camera (panning) to follow the subject during the exposure. This records the moving object comparatively sharply against a blurred background and emphasizes the impression of speed. Deliberate movement blur is sometimes aimed at for pictorial effect.

Camera shake through unsteady support during the ex-

posure also creates image blur—over the whole picture in such cases. Hand-held shots generally demand shutter speeds of $^1/_{30}$ second or shorter. For longer times a firm camera support—such as a tripod—is essential.

Exposure technique. The correct exposure (aperture and shutter settings) can be derived from tables or calculators or by direct measurement of the subject luminance with a light meter.

Meter readings. Straightforward readings with the meter cell pointed at the scene measure average reflected-light intensity, and they are reliable for subjects of average contrast and brightness distribution. Abnormal subjects of extreme contrast, or large highlight or shadow areas, need overriding exposure corrections. Spot readings taken with a meter covering a narrow angle measure selected subject details; *e.g.*, a medium tone for correct exposure irrespective of brightness distribution. This is practical with certain through-the-lens meter systems. Readings of the light reflected from the brightest scene highlight or of the light falling on the scene (incident light readings) measure the lighting level. This method is suitable for reversal processes (including colour) in which the exposure must be accurately matched to the lightest subject detail.

The selection of the appropriate aperture and shutter speed among combinations of equivalent camera exposures depends on the requirements of depth of field and of subject movement.

Lens aperture as the key to flash exposure

Flash exposures. The lens aperture controls flash exposure, since the flash duration is generally shorter than the open time of the camera shutter. The light intensity on a scene is, by basic illumination laws, inversely proportional to the square of the distance between light source and subject. The light intensity reaching the film is also inversely proportional to the square of the diaphragm *f*-number. For a given flash source and film speed, increased flash distance can therefore be compensated by corresponding *f*-number adjustment; the exposure is constant as long as the product of distance and *f*-number is constant. Flash and film manufacturers quote this product as a guide number for various flash–film combinations. For rapid exposure calculation, dividing the guide number by the flash to subject distance gives the required *f*-number; dividing the guide number by the *f*-number gives the distance at which the flash must be arranged. Different guide numbers are needed for shutter times shorter than the flash duration.

Special meters can measure flash light quantity on a scene during a test firing of the flashes. This is a practical procedure with studio setups using electronic flash units, but it is wasteful with flash bulbs. Cadmium sulfide cells of automatic cameras control an electronic shutter to allow for the light from flash bulbs (but not electronic flash) in their exposure-control functions.

There are two types of flash exposure automation. In one type, coupled focus settings and lens diaphragms stop down the lens when focussing on nearer distances to keep the flash exposure constant. The coupling is usually adjustable for different flash guide numbers. In the other type, used with computer electronic flash units, a photocell monitors the light intensity reflected from the subject and controls the flash duration and hence the exposure accordingly. The control range may typically cover $^1/_{50,000}$ to $^1/_{1,000}$ second in flash duration and 1½ to 20 feet (½ to six metres) in subject distances.

Exposure latitude. The ideal negative exposure records the darkest subject shadows as a just visible density. More exposure yields a denser negative, which, however, can still give an acceptable print by appropriate print-exposure adjustment. This range of usable negative exposures, the exposure latitude, depends on the film and the subject. This latitude is greater, the lower the subject contrast and the greater the film's exposure range (and, generally, the lower the film contrast). Thanks to exposure latitude, simple cameras with limited exposure adjustability can still yield acceptable pictures under differing light conditions.

Lighting technique. The kind of lighting on the scene governs the way in which the picture reproduces the subject. Orientation of the subject—as in taking a portrait—with respect to the light direction can often control the effect. Lighting from behind the camera gives flat effects, light from one side yields depth and modelling, while the principal light from behind the subject produces dramatic against-the-light effects of high contrast. Artificial light setups in the studio, with tungsten lamps or electronic flash, offer the greatest flexibility. Under such conditions, the photographer can arrange two or more lamps for various lighting effects.

Flexibility with studio lighting

Directional lighting improves detail contrast and brilliance. Excessive subject contrast, however, makes accurate exposure settings difficult and may lead to loss of picture detail in the highlights or shadows. Fill-in lighting, by a flash or other light source on or near the camera, can illuminate heavy shadows facing the camera.

PROCESSING AND PRINTING

Negative development. Amateurs usually process films in developing tanks. In this type of development, a roll or miniature film is wound around a reel with a spiral groove, which keeps adjacent turns separated and allows access by the processing solutions. Once the tank is loaded (in the dark), processing takes place in normal light, the processing baths (developer, intermediate rinse, fixer) being poured into the tank at the appropriate intervals. Sheet films are similarly treated in small tanks or held in hangers and immersed sequentially in the different processing solutions. Large-scale commercial processing laboratories use machines that automatically feed the films through the solutions in proper sequence.

Developers and their characteristics. The developer consists typically of one or more developing agents, a preservative (such as sodium sulfite) to prevent oxidation by the air, an alkali (such as sodium carbonate) to activate the developer, and a restrainer or antifoggant to ensure that the developer acts only on exposed silver halide crystals. A developer's main characteristics are activity, development speed, and effect on film gradation, graininess, and sharpness. Developers may be prepared on the basis of published formulas or bought as ready-mixed powders or concentrates for dilution with water.

The developer is allowed to act for a specific time to build up the image to the required density and contrast. This time depends on the developer, the temperature, the degree of agitation, and the film—as indicated by film and developer manufacturers' recommendations.

Fixing. The fixing bath contains a chemical (sodium thiosulfate) that converts the silver halide into soluble, complex silver salts that dissolve in the fixer. During this process the film loses its original silver halide milkiness overlaying the image and becomes clear. The fixer also contains a weak acid (to halt the development process) and a hardening agent to reduce gelatin swelling.

Washing and drying. Washing removes all residual soluble chemicals from the emulsion and must be thorough for image performance. Films are hung up to dry after removal from the tank.

High-speed processing. Greatly reduced processing times are possible with high-activity developers at elevated temperatures and with fast-acting fixing agents, such as ammonium thiosulfate. Such processes can cut access time to the negative down to less than a minute. One-bath (monobath) processing in a solution containing both fast-acting developing agent and fixing chemicals also reduces processing time. In special rapid-access processing equipment, films pass through chambers spraying the processing solutions onto the film surface or run in contact with monobath-soaked webs.

Monobath processing with developer and fixer

Printing. The simplest printing equipment is the contact printing frame in which the negative and printing paper are held together behind a glass plate during exposure to a suitable lamp. A printing box is essentially a printing frame with built-in light source. Contact printing gives a positive of the same size as the negative.

Enlargers. Usually negatives are enlarged to prints of the desired final size. The enlarger is a projection system on a vertical column mounted on a horizontal baseboard (Figure 10). It has a lens, a film holder (negative car-

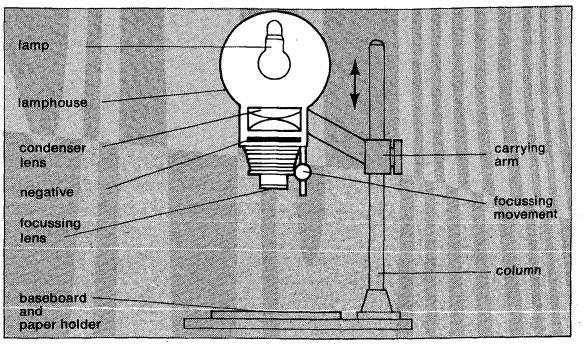


Figure 10: Basic enlarger features.

rier), and a lighting system (typically a lamp and condenser lens) for illuminating the negative. Raising or lowering the enlarger head on the column controls the image magnification; adjustment of the lens to negative distance focusses the image on the enlarging paper on the baseboard. In automatically focussing enlargers, these two adjustments are linked mechanically to keep the image sharp all the time. Enlargers are made in various sizes to take different maximum negative formats.

Printing papers. Papers for enlarging and contact printing are produced in grades of differing exposure range—*i.e.*, ratios of shortest to longest exposure to produce the lightest tone and a full black, respectively. The various grades yield prints of a normal tone range from negatives of different contrasts: a soft paper grade for a high-contrast negative, a normal paper for a normal negative, a hard paper for soft negatives, and so on. Paper grades are also numbered—typically from 0 to 5—in ascending order of contrast. Variable-contrast papers use a mixture of two emulsions of a different contrast and colour sensitivity; the contribution of each is controlled by filters in the path of the exposing light.

Paper grades in order of contrast

Other paper characteristics are the speed (slower for contact papers, faster for enlarging papers), image colour (blue black to warm brown), surface texture (glossy, velvet, mat), and base thickness (single or double weight). Some printing materials are coated on a white opaque plastic base or on plastic-laminated paper; these papers absorb no water during processing, allowing prints to dry more rapidly.

Printing exposures. Correct printing exposures are determined by trial and error or by test strips given a series of progressively increasing exposures. More sophisticated exposure control systems measure either the brightness of selected image portions projected on the enlarger baseboard or the average light intensity reaching the paper during the exposure. Printing papers are exposed and processed in a darkroom lit by an olive green or orange safelight. Printing papers are sensitive to violet, blue, and sometimes green light.

Print processing. The processing of prints consists of development, an intermediate rinse or stop bath, fixing, and washing. The developer and fixer are similar in principle to those used for negative films. In the normal method, dish or tray processing, prints are immersed successively in the solutions in dishes laid out side by side. Development is checked visually, the print remaining in the developer until the image has reached its full density. For drying, the prints may be clipped to a line, placed in a heated print dryer, or squeegeed onto a mirror-finished plate for a high-gloss surface.

Stabilization processing. Certain rapid processing papers incorporate developing agents in their emulsions and are processed on a roller processor. This processor runs the paper through an activating bath for instant development and then through a stabilizing bath, followed by a pair of squeegeeing rollers from which the print emerges merely damp. This process takes about ten to 15 seconds; the prints, however, do not keep quite as well as conventional prints, since unexposed silver salts are not removed from the emulsion but only converted into moderately light-stable compounds. Such prints can be made more permanent by subsequent fixing and washing.

Disadvantage of prints finished by stabilization processing

Dry processing. Processing baths can be completely eliminated by incorporating in the emulsion of the paper development and stabilization chemicals that become active on heating. One method is to disperse the processing chemicals in the emulsion in microscopic capsules containing the solution and a blowing agent. On passing the exposed paper over a heated roller, the blowing agent bursts the capsules, and the liberated processing solutions act on the silver halide immediately around each capsule. The liquid solvent instantly evaporates, leaving a dry print. Such encapsulation materials are used for such purposes as making proof prints of negatives and re-enlarging microfilm images. Certain nonsilver processes in photocopying systems also offer dry processing.

COLOUR PHOTOGRAPHY

Colour reproduction. Present-day colour photographic processes are tricolour systems, reproducing different colours that occur in nature by suitable combinations of three primary coloured stimuli. Each of these primary colours—blue-violet, green, and red—covers roughly one-third of the visible spectrum. Tricolour impressions can be produced by (1) combining coloured lights (additive synthesis) or (2) by passing white light through combinations of complementary filters each of which holds back one of the primary colours (subtractive synthesis).

In additive synthesis, a combination of red and blue-violet light (*e.g.*, light beams of the two colours directed on the same spot of a white screen) gives a purplish pink (magenta); equal parts of red and green produce yellow, and equal parts of green and blue violet produce bluish green (cyan). Superimposition of all three light beams on a screen yields white; combinations of varying proportions of two or three of the colours produce virtually all the other hues.

In subtractive synthesis, yellow, magenta, and cyan filters or dye layers subtract varying proportions of the primary colours from white light. The yellow filter absorbs the blue component of white light and so controls the amount of blue present in a white light beam that has passed through the filter. Similarly, the magenta filter controls the amount of green light left, and the cyan controls the amount of the red component. A cyan and magenta filter superimposed in a white-light beam holds back both the red and the green component, making the emerging beam blue. Similarly a cyan and yellow filter together yield green, and a yellow and magenta filter together yield red. Superimposing such filters or dye images of different densities in a white-light beam can therefore re-create any colour impression in the same way as superimposing light beams of the primary colours.

Subtraction of primary colours by filters or dye layers

The difference between additive and subtractive synthesis is the approach: in additive synthesis colours are built up by combining different intensities of primary-coloured light and in subtractive synthesis, by removing different proportions of primary-coloured light from white light. All modern colour films are based on subtractive synthesis. In theory either method of colour synthesis should be capable of reproducing every existing colour in nature. In practice, the reproduction is imperfect. No filter dyes meet the required ideal specifications. Nevertheless, for most purposes, reproduction is adequate.

Colour films. *Reversal colour films.* To reproduce colour by subtractive three-colour synthesis (Figure 11),

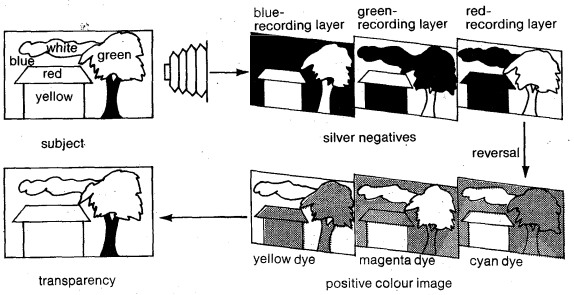


Figure 11: Colour reproduction sequence with subtractive reversal film (see text).

colour films first break down the colours of an image into their primary components by means of three separate sensitized layers, each of which responds exclusively to blue, green, or red light. The image in each layer is reversal-processed to yield a positive dye image in a colour complementary to the layer's spectral sensitivity. Thus, the blue-sensitive layer first yields a negative image of everything blue in the original scene (*e.g.*, the blue sky) and then a positive image of everything that is not blue. This positive image is coloured yellow. Similarly, the green-recording layer yields a magenta positive image of everything that is not green, and the red-recording layer a positive cyan image of everything that is not red. Blue sky, for instance, does not figure in the yellow positive image, but does figure in the magenta positive image (not being green) and in the cyan positive image (not being red). The magenta and cyan dyes in the areas that were blue sky are superimposed; white light passing through the resulting transparency loses its green and red, but not its blue, component; thus, the sky appears blue. Similarly, green subject components end up as positive yellow image density in the blue-recording and positive cyan density in the red-recording layer, combining to green in the transparency. Yellow records as a negative image in the green-recording and red-recording layers, hence leaving a positive yellow image only in the blue-recording layer. All other colours are formed by similar combinations of different densities of the dye images.

Negative colour films. Negative colour materials work in a similar way but yield a negative dye image by direct development. Blue subject tones record in the blue-sensitive film layer to produce a yellow negative image. Green colour components yield a magenta dye image in the green-responding layer, and red components yield a cyan dye image in the red-recording layer. With respect to the subject, the colour negative therefore reverses the tones in brightness as well as in colour. Printing the colour negative on a colour paper with three differentially responding layers reverses the process once more, reconstituting the original subject colours in a positive print.

Colour-film structure. In a magnified section (Figure 12) a reversal colour film has these components:

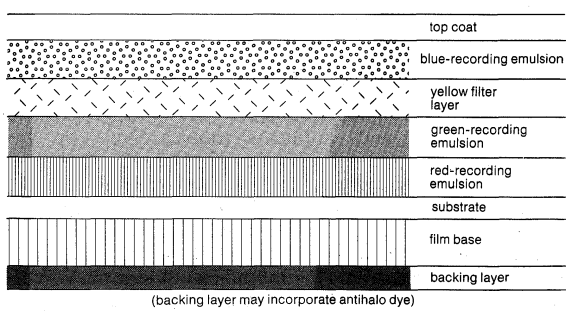


Figure 12: Structure of reversal colour film (see text; thicknesses not to scale).

1. A top layer of plain gelatin, which protects the underlying layers against abrasion and damage.
2. The first emulsion layer, which contains blue-sensitive silver halide plus a yellow-forming colour coupler. This is a colourless substance that reacts with the decomposition products of the developing agent to generate dye in all areas where a silver image is produced and in proportion to the density of that silver image.
3. A yellow filter layer of finely divided (colloidal) silver, which holds back blue light from the subsequent emulsion layers. It disappears during the bleaching stage of processing.
4. The second emulsion layer, which contains blue-and-green sensitive silver halide plus a magenta-forming colour coupler. The blue sensitivity is suppressed by the yellow filter layer.
5. The next emulsion, which is blue-and-red sensitive (blue again being suppressed) and contains a cyan-forming colour coupler.
6. A substrate, which ensures optimum adhesion of the emulsion layers to the film base and may also contain light-absorbing silver to prevent the scattering of light by reflection from the support surface (halation).
7. The film base or support of clear cellulose acetate derivative (or sometimes polyester), typically about 0.005 inch thick.
8. The back of the support, which carries a light-absorbing layer (an alternative to the antihalation layer in the substrate); on roll films this also acts as an anticurl layer.

Variations in reversal colour films

The above scheme is subject to variations. Some reversal colour films do not incorporate colour couplers in the emulsions but introduce them in the processing solutions. Processing such nonsubstantive colour films is more complex than processing substantive-coupler films containing couplers in the emulsions.

Negative colour films are similar, but the couplers are themselves often already coloured yellow or red or both. The unreacted couplers remain as positive images, which compensate for some deviations of the image dyes from the colour characteristics of ideal dyes. Such negatives are characterized by an apparently overall reddish or orange tint.

Some colour films come with different combinations of colour sensitivity and dye colour formed in the individual layers for deliberately falsified colour rendering (false-colour films) in special applications.

Colour positive (print) materials have a paper or white, opaque film base instead of transparent film and have no antihalation layer. The emulsion sequence may be different from the above scheme, as may be their spectral sensitivities. The latter may be keyed to the transmission characteristics of the negative dyes for better colour reproduction.

The dye images are reasonably lightfast, but fade on prolonged exposure to ultraviolet-rich radiation. Colour transparencies and prints intended for continuous display may be protected by ultraviolet-absorbing coatings or filter layers.

Colour balance. Colour film reacts to all hue and tone differences including the prevailing light colour. A film recording approximately natural colours in daylight reproduces scenes photographed by tungsten light with a reddish overall tint—because this lighting is richer in red rays than is daylight. This spectral balance of different "white" light sources may be rated numerically by the colour temperature—a concept of theoretical physics that, with tungsten lighting, corresponds roughly to the absolute lamp-filament temperature. Such absolute temperatures are expressed in degrees Kelvin (°K). The higher the colour temperature the richer the light is in bluish and the poorer it is in reddish rays and vice versa. Average daylight is rated at about 5,500° K, the light from an overcast sky from 6,500° K up; the colour temperature of tungsten lamps ranges between 2,600° K and 3,400° K.

To ensure correct "white light" colour reproduction with different types of lighting, the sensitivities of the three film layers must be matched to the colour temperature of the light. Reversal colour films are therefore made in different versions. Daylight films are balanced for faithful colour rendering with 5,500° to 6,000° K light (average daylight, blue-coated flash bulbs, electronic flash). Type A films are balanced for 3,400° K light (high-temperature tungsten filament and photoflood lamps). Type B films and tungsten films are balanced for around 3,200° K to match tungsten studio lamps.

"Universal" colour films

Such accurate film balance matching is less vital with negative colour films since the colour rendering of the print can be modified during colour printing. "Universal" amateur negative colour films are usable with any light, from tungsten to daylight. For high quality, professional negative colour films are still preferentially balanced to either daylight or tungsten sources.

Strongly coloured filters are suitable only for special effects; they overlay the colour image with the filter colour. Pale correction filters can match a film to a light source other than that for which it is balanced—*e.g.*, pale blue, with a daylight-type film used in tungsten lighting,

to raise in effect the colour temperature. Pale pink or amber filters similarly reduce the colour temperature for using artificial-light-balanced films in daylight. Colour-film manufacturers publish detailed recommendations of actual filters required for such conversion.

In outdoor photography, especially involving distant views, an ultraviolet-absorbing filter is often required, as ultraviolet radiation records in the blue-sensitive layer of the film, producing an overall blue cast in the transparency. A pale pink skylight filter for outdoor subjects lit only by skylight counteracts the cold bluish colour rendering resulting from such illumination.

Colour-film processing. The processing sequence for colour materials is longer than for black-and-white films and requires more solutions. Development needs very precise timing and temperature control. Colour films can be processed in amateur developing tanks; professionals use sets of tanks in temperature-controlled water jackets with provision for standardized solution agitation.

Reversal colour processing. Each make and type of colour film requires its own individual chemicals (often available in kits) and sequence for processing. A typical sequence for a substantive (couplers in the film) colour film is (1) first development, to form a negative silver image in each emulsion layer; (2) stop bath or hardening or both, to arrest negative development and harden the emulsion layers; (3) fogging, to render the remaining silver halide in each emulsion layer developable (in this step, the film is either exposed uniformly to white light or treated with fogging chemicals); (4) colour development, to produce a positive silver image in the remaining silver halide plus a coincident dye image by reaction with the colour couplers; (5) bleaching, to reconvert the negative and positive silver images into silver halide; (6) fixing, to dissolve all silver halide out of the emulsion, leaving only the three dye images; (7) the final wash, to remove all soluble chemicals (rinsing and washing are also done after each of the above processing stages); and (8) stabilization and drying (the stabilization bath in some colour processing sequences improves the lightfastness of the dyes).

Processing of colour films lacking couplers

Processing of nonsubstantive colour films, in which the couplers are in the colour developer, is more complex because each emulsion layer is re-exposed by appropriately coloured light and colour-developed separately. This operation requires automated processing machinery.

Negative colour processing. Negative colour films are practically all of the substantive-coupler type. In a typical processing sequence, colour development is done first; it forms a negative silver image in each emulsion layer, together with a corresponding dye image. Next, a stop bath stops the development process and may or may not include hardening. Bleaching then converts the silver image into silver halide. In the fourth step, fixing dissolves the silver halide out of the emulsions, leaving the negative dye images. Finally comes washing, stabilizing, and drying. Washing or rinsing also follow each of the first three stages.

Substantially the same procedure is followed for processing positive colour papers. The bleaching and fixing stages may be combined in one bleach-fix stage.

Colour printing. Colour print processing can be done in dishes or trays, but professional colour laboratories use either tank sets of suitable size to take larger print formats or drum processors, with the print attached to the circumference of a rotating drum that dips into a trough of processing solution. The liquid in the trough is changed at time intervals appropriate to the processing sequence; in automatic systems the solution change is program controlled. Various processes are in current use.

Methods of obtaining colour control

Positive prints from colour negatives. Positive prints are obtainable from colour negatives by enlarging them onto a positive colour paper. Colour control consists of modifying the colour of the printing light by yellow, magenta, and cyan filters of various densities—singly or in combination—to obtain a print of correct or desired colour balance. In simple colour enlargers such filters are placed in a filter drawer between the negative and light source. Advanced colour enlargers use variable-density cyan, magenta, and yellow filters in front of the light source; the light is then thoroughly mixed in a diffusing globe or box before reaching the negative. Alternatively, high-density filters may be inserted to varying degrees into the light path.

The above method of subtractive or white-light printing depends on subtracting or holding back colour components of white light. In additive printing the prints are given successive exposures through high-density red, green, and blue filters. Each of these exposures forms the image in one of the emulsion layers of the paper; colour balance depends on the proportions of the individual exposure times. The red, green, and blue filters are mounted in a turret in front of the enlarger lens or interposed by a relay mechanism between the lamp and the negative. In automatic colour printing systems these relays are controlled by photocells that evaluate the red, green, and blue components of light transmitted by the negative.

An alternative method of additive printing starts with a white light exposure (equivalent to simultaneous blue, green, and red light) and then interposes high-density yellow, magenta, and cyan filters in the light path to terminate the exposure to blue, green, and red light, respectively.

Reversal colour printing. Colour transparencies can be printed on a reversal colour paper similar to a reversal film and processed in an analogous way. The same kind of colour control with filters is again possible, but the colour effect of the subtractive filters or of the additive filter exposures is reversed.

Dye-destruction processes. Dye-destruction processes differ from chromogenic colour materials (where colour images are produced during development) in starting off with emulsion layers containing the final dyes. During processing these are bleached in proportion to the silver image formed. Straightforward processing of a dye-destruction or dye-bleach material yields a positive image from a positive original and consists of: development to form a silver image; stop-fixing to arrest development and remove unexposed silver halide; dye bleaching to bleach the dye in the areas containing a silver image; silver bleaching to convert the silver image into silver halide; and fixing, to remove residual silver halide. Washing is done between all the processing stages.

Obtaining a positive image from a negative requires a more elaborate processing sequence, analogous to reversal processing in a chromogenic system. Dye-bleach materials use far more light-stable dyes than those produced by colour-coupling development. The positive–positive procedure also yields duplicate transparencies on dye bleach materials with a transparent film base.

Assembly colour prints. The original method of producing colour prints was based on separation negatives obtained by photographing the original scene on separate black-and-white plates or films through a blue, green, and red filter, respectively. This method analyzes the subject in terms of its tricolour components in the same way as the initial negative images in a three-layer colour film. Positive prints from the separation negatives, converted into colour images (*e.g.*, by toning) and superimposed on top of each other, yield a subtractive tricolour print.

Prints by the dye transfer process

The main surviving assembly print process, dye transfer, uses a set of separation positives on a panchromatic matrix film made either from separation negatives of a colour transparency or by separation (three filtered exposures) from a colour negative. Appropriate processing converts the matrix film into a gelatin relief image whose depth is proportional to the positive silver density. Each matrix is soaked in dye solution (yellow for the matrix derived from the blue-filter negative and so on), and the dyes from the matrices are transferred in succession to a single sheet of gelatin-coated paper. Elaborate care is required to ensure accurate superimposition (registration) of the dye images; the result is a positive colour print.

Transparency projection. Many amateur colour pictures are in the form of transparencies, particularly on 35-mm film. These are usually mounted in plastic or card frames or bound between glass for projection on a screen

in a darkened room. The projector consists of a lens, a holder for the slide, and a lighting system (lamp, reflector, and condenser lenses to concentrate the light onto the slide). Modern slide projectors take the slides in magazines or trays holding 30 to 50 or more slides. An automatic slide transport feeds each slide from the tray into the light path of the projector and may be operated from a remote control unit or by pulses from a tape recorder, which can also record a commentary to the complete slide series. Some projectors feature remote-controlled or automatic focussing to keep each successive slide image sharp on the screen.

The standard miniature slide size is 2 × 2 inches for transparencies up to 1⅝ × 1⅝ inches; the most usual transparency format in such slides is 24 × 36 millimetres. Projectors for larger slides (*e.g.*, 2¾ × 2¾ inches for transparencies up to 2¼ × 2¼ inches) and ultraminiature projectors (*e.g.*, 10 × 14 millimetre transparencies in 3 × 3 centimetre slide frames) are suitably scaled up or scaled down versions of the standard models.

The size of the projected image depends on the distance of the projector from the screen and the focal length of the projection lens. Projectors may also project from the back onto a translucent screen; such rear projection setups are more compact, and the image is often bright enough for viewing in daylight. The rear-projection system is used in schools and for commercial displays. Elaborate slide shows are produced by linking two or more projectors aimed at the same or adjacent screen areas. With a suitably assembled slide set, the pictures can be made to change, overlap, and assemble, according to a predetermined program.

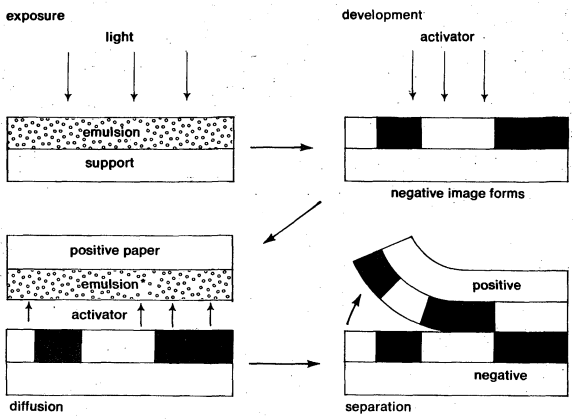
IV. Special techniques

POLAROID PHOTOGRAPHY

History and evolution. The diffusion-transfer process, the basis of the Polaroid system, was invented almost simultaneously in 1939 in Belgium and Germany. It relies on a silver halide negative material with a special processing procedure to produce positives in one stage. The process was initially used in office copying systems, and it is still in use in such systems.

Land's adaption of process for camera use

In 1947 Edwin Land, a U.S. inventor, introduced a version of the process adapted for camera use in which the processing takes place in the camera and yields a positive print within a minute or less of exposure. The original Polaroid material was a dual roll of negative and positive paper. Later, film packs and sheet films appeared, for use in special cameras or in camera backs to fit professional cameras taking interchangeable film holders or magazines. Present-day Polaroid films include high-speed emulsions, high-contrast process, transparency, and scientific materials based on the same process. Early in 1963 a diffusion-transfer colour material appeared for positive colour prints; similar materials for positive transparencies and for colour negatives are in preparation.



*emulsion with development nuclei

Figure 13: Steps in the production of a positive print by the diffusion-transfer process.

Black-and-white diffusion transfer. The negative material—paper or film—carries a silver halide emulsion incorporating developing agents. After exposure (Figure 13), soaking in an alkaline solution containing a silver halide solvent activates the developer to yield an almost instant negative image. The negative is brought into contact with an unsensitized print material containing development nuclei. The silver salts from the unexposed image areas diffuse into the positive layer and deposit metallic silver on the development nuclei. The whole process takes one-half to one minute, after which the negative and positive sheets can be peeled apart. The positive is the finished print; the negative is usually discarded or, if on film, it may be washed and treated to give a conventional negative for normal enlarging.

In the Polaroid instant-picture process the negative emulsion is fast enough for camera exposure. The positive material carries pods of jellied processing solution. After exposure, the negative and positive sheets pass between pressure rollers built into the camera or camera back; the pressure ruptures the reagent pod and the processing jelly is spread between the two surfaces. After the requisite processing time the finished print is again peeled off the negative material.

The Polacolor process. The colour version of the Polaroid material uses six active layers: a blue-sensitive silver halide emulsion backed by a layer consisting of a yellow dye and developer compound, a green-sensitive layer backed by a layer of magenta dye and developer, and a red-sensitive layer backed by a cyan dye and developer. The dye-developer in each case consists of dye molecules (not colour couplers) chemically linked to developing agent molecules.

After exposure and activation by the alkaline jelly, the dye-developer molecules in each layer migrate into the adjacent silver halide layer and there develop any exposed silver halide to a negative image. This at the same time anchors the dye component of the dye-developer molecule in position. Dye-developer molecules in unexposed image areas are not used up by development but migrate into the receiving layer of the positive material. Here they are immobilized, remaining as dye images corresponding to a positive of each silver halide layer in the negative film. The dyes thus re-create a full-colour positive image. The process depends on the controlled diffusion of the dye-developer molecules, achieved by spacing layers and balanced exposure and development time. Developing takes about one minute. Like the black-and-white Polaroid material, Polacolor film is available in rolls, film packs, and sheets.

Applications. The Polaroid process has an advantage in applications that need quick access to a finished print, provided that the limited print size (3¼ × 4¼ inches in the film pack and roll film, or 4 × 5 inches in sheet film) is not too serious a drawback. The initial field for the process was amateur snapshooting and instant portraits, from which evolved the taking of identification pictures for work and security passes. Such passes are made with special cameras that record a portrait together with personal details on a composite print, which is then laminated to form a tamperproof identity card of a standard size. In studio photography, Polaroid prints provide a quick method of making exposure tests, and checking the effects of lighting arrays. Polaroid prints have found some use in press photography. Commercial applications are numerous and include progress pictures on a building site and shots of properties for real-estate advertisements. In the laboratory, Polaroid prints yield instant records of experimental setups and are used for photomicrography, infrared photography (with a special infrared material), and for recording oscilloscope traces and other data from instruments. Polaroid prints of close-ups are also used in forensic and medical photography.

Limitations of the Polaroid process

Instant diffusion-transfer prints are not suitable for rapid-sequence photography, for applications requiring high image definition, or for applications in which a number of duplicate prints are required. Nor is the Polaroid process applicable under conditions precluding immediate

processing of the picture (*e.g.*, in underwater or space photography).

Key patents have largely reserved the use of this type of instant photography to materials and equipment of the Polaroid Corporation. As these patents are expiring, other major manufacturers are patenting their own versions of instant-picture cameras and processors. Proposed modifications of the process include negative films loaded in standard spools or cartridges in normal cameras, which are then fed—in a daylight processor—through an activating bath together with a separate positive material. The pictures are not available singly after each exposure, but the processing procedure at the end of a roll is virtually as rapid and convenient as Polaroid-camera processing. Polaroid has demonstrated a movie-film system based on this idea.

SPECIAL PHOTOSENSITIVE SYSTEMS

The high working speed (efficiency of converting light into permanent images) of silver halides makes them almost the only materials suitable for camera use. Numerous light-sensitive systems not using silver have been known since the beginning of photography. In view of an increasing silver shortage and its consequent high price, a number of substitute systems have grown in importance, and new ones have appeared during recent years. Most of them are limited to office copying, microfilming, the graphic arts, and other applications in which flat copy is reproduced.

Electrophotography. Electrophotography covers a number of processes that rely on photoconductive substances whose electrical resistance decreases when light falls on them. A layer of such a substance with a grounded backing plate is given a uniform electrostatic charge in the dark. When a light image is projected onto the surface, the photoconductor allows the electrostatic charge to leak away in proportion to the exposure. This leaves an "image" charge that can be converted, in various ways, into a visible image.

Principles of xerography

In xerography, the photoconductive layer is selenium, and the image is made visible by dusting the plate with an electrostatically charged powder (toner) having a charge that is the opposite of that of the electrostatic image. The powder adheres to the image portions only and is then transferred to a sheet of plain paper also under the influence of electrostatic fields. A final heat treatment fuses the powder into the paper for a permanent picture. The process usually makes a positive from a positive original. In office copying machines (the main application of xerography) the whole operating sequence is programmed and automated. A zinc oxide coated paper may replace the selenium plate; if so, the pigment powder deposit is fused directly into the paper surface.

The process is used mainly for line images without intermediate tones between black and white. Modified procedures permit continuous-tone reproduction and—with coloured pigments—also colour printing.

In the electroplastic process, a transparent thermoplastic serves as the photoconductive layer. After charging and exposure, the residual electrostatic charge forms stresses in the thermoplastic. Controlled heating deforms the surface in the image areas into a grain pattern, which is frozen into the plastic on cooling. The resulting image is light scattering and is viewed by reflection or in special projection systems. The electroplastic image can be erased by reheating the plastic layer, which then becomes reusable.

Colloid and photopolymer processes. A comparatively early nonsilver process depended on organic colloid (gum or gelatin) treated with a bichromate. Exposure to light hardened the gelatin, rendering it insoluble, while unexposed portions could be washed away with warm water, leaving a relief image.

Photopolymer systems substitute a plastic precursor in place of the gelatin. The plastic precursor polymerizes to an insoluble plastic when exposed to light, and the unexposed soluble material is washed out by a suitable solvent. Photopolymer processes have been adapted for forming resists (protective coatings) for etching, as, for instance, in the manufacture of printed circuits. In indirect photopolymer systems a light-sensitive substance is mixed with a plastic precursor and on exposure decomposes into compounds that initiate polymerization of the plastic. The polymerizable layer may include a pigment for a final coloured image. Superimposing colour images derived from separation negatives can yield positives; systems of this type are used for quick colour proofing in photomechanical reproduction.

Diazonium processes. A diazo or dyeline process depends on the decomposition by light of organic diazonium salts. These salts can also couple with certain other compounds to form dyes. After exposure, only the exposed (and decomposed) diazonium salt forms dye, producing a positive image from a positive original.

The materials are usually papers or transparent supports impregnated with the required chemicals. They are mainly sensitive to ultraviolet rays and can therefore be handled by normal tungsten lighting.

The light-decomposition of diazonium compounds also produces gaseous nitrogen. This phenomenon is utilized in vesicular processes that incorporate the diazonium compound in a thermoplastic layer. The nitrogen slowly diffuses out of this layer, but if heat is added immediately after exposure, the expanding nitrogen gas forms minute light-scattering bubbles visible as an image. The scattering power corresponds to the exposure. Further general exposure, after the plastic has cooled, decomposes the residual diazonium compound with gradual diffusion of the nitrogen out of the layer, destroying the latter's light sensitivity. This process and thermal dyeline systems are dry-processing instant-access systems and are used for making microfilm duplicates.

Photochromic systems. Certain dyelike substances can exist in a colourless and a coloured state. They are called photochromic compounds. The coloured state is formed by exposure to radiations of a certain wavelength. The compound reverts to its colourless state either in the dark or on treatment with radiation of a different wavelength. This reversibility is a primary characteristic of photochromism, and it is an instant-image system involving no processing.

Reversibility of the photochromic systems

Photochromic systems are used in microrecording (see below *Microfilming and micro-reproduction*). As the change of state takes place on a molecular level, the images are practically grain free and resolution is limited only by the resolving power of the optical system. Photochromic materials can be negative or positive working. With some photochromic compounds the dye image can be rendered permanent by optical, gas, or solution treatment.

Glasses containing certain metal compounds also act as photochromic materials. Exposure to light breaks down the compounds into metal that forms a visible (and permanent) image in the glass. Another type of photochromic glass contains silver halide crystals dispersed in the glass melt. The action of light decomposes the silver halide, forming a visible silver deposit. The halogen cannot escape from the glass, so it recombines with the silver in the dark and the image fades. Such photochromic glasses are incorporated in automatic light-control devices; light transmission decreases as the intensity of the light reaching the glass rises. Such glass has found use in certain types of sun glasses.

SPECIAL TECHNIQUES AND APPLIED PHOTOGRAPHY

High-speed and stroboscopic photography. High-speed photography is generally concerned with exposure times shorter than about $\frac{1}{1,000}$ second (one millisecond) and often exposures shorter than $\frac{1}{1,000,000}$ second (one microsecond). This field partly overlaps that of high-speed cinematography—sequences of very short exposures. Exposure times can be reduced by high-speed shutter systems or by short duration flash sources.

High-speed photography, together with high-speed cinematography, aids in the study of missiles, explosions, nuclear reactions, and other phenomena of military and scientific interest. In industry, high-speed pictures show up movement phases of machinery, relays, and switches;

dynamic fractures of materials or insulation breakdown; and, in natural science studies, flight movement of birds and insects.

High-speed shutters. The shortest exposure with mechanical shutters is about $\frac{1}{3,000}$ second. Special high-speed shutter systems are magneto-optical, electro-optical, or electronic.

A magneto-optical shutter (Faraday shutter) consists of a glass cylinder placed inside a magnetic coil between two crossed polarizing filters; so long as the filters remain crossed, virtually no light can pass through. A brief current pulse through the coil generates a magnetic field that rotates the light's plane of polarization in the cylinder so that during the pulse some light passes through the second polarizing filter.

Principle of the Kerr cell

The electro-optical shutter (Kerr cell) is made up of a liquid cell of nitrobenzene fitted with electrodes and again placed between two crossed polarizers. An electric pulse applied to the electrodes changes the polarization properties of the nitrobenzene so that this arrangement again transmits light. Minimum exposure time is around five nanoseconds (5×10^{-9} second).

Image converter tubes electronically transmit and amplify an optical image focussed on one end of a tube onto a phosphorescent screen at the other end. Electrons flow in the tube only in the presence of an electric field, which can be controlled by short-time pulses down to a few nanoseconds.

High-speed light sources. The shortest electronic-flash duration is around one microsecond. Spark discharges in air between electrodes yield still shorter exposures; discharge voltage may go up to tens or hundreds of thousands of volts. Short-duration pulses applied to X-ray tubes produce X-ray flashes for high-speed radiography. The shortest exposures are between 20 and 50 nanoseconds. Special switching modes turn lasers into high-speed sources with durations down to a fraction of a nanosecond.

Synchronization. Generally the event photographed is made to trigger the exposure (the current pulse to operate the shutter or flash or spark source) to ensure correct synchronization. Examples are bullets interrupting a light beam to a photocell, or self-luminous phenomena (explosions) triggering the system via a photocell circuit. The event and the exposure may be also triggered together by a signal from a common source.

Stroboscopic photography. Electronic-flash units designed to flash in rapid succession (up to several hundred times a second) can photograph a moving subject in front of a stationary camera with its shutter open to yield multiple images of successive movement phases. The technique has been used in pictorial and sports photography (*e.g.*, recording the movement of dancers or golfers) and for analyzing movement cycles without a motion-picture camera. With very rapid cyclic movement (*i.e.*, rotating machine components) illumination by the stroboscopic flash synchronized with the movement cycle makes the moving component appear stationary.

Aerial photography. Photographs from airborne or spaceborne vehicles either provide information on ground features for military and other purposes (reconnaissance) or record the dimensional disposition of such features (surveying).

Reconnaissance photography for capturing detail

Reconnaissance photographs call for maximum sharpness and detail rendering. Infrared films are often used to bring out details not discernible visually. In nonmilitary applications such photographs may reveal ecological factors (tree diseases, crop variations) and traces of archaeological sites not visible from the ground. Such shots are generally taken with cameras using five- or 9½-inch roll film in large magazines, built into the aircraft and operated electrically by the pilot or other crew member, or automatically at set intervals. Some systems incorporate a shutterless technique; the film runs continuously past a slit at a rate matched exactly to the image movement in the camera's focal plane as the aircraft flies over the ground (image motion compensation).

Aerial survey is a systematic procedure of photographing the ground for map production; exposures are made at intervals to partly overlap the view of successive pictures. The aircraft must fly on a steady course at a constant height, which, together with the focal length of the lens, determines the image scale. Larger areas are covered by successive adjacent tracks. The individual photographs are enlarged to the same degree and then assembled in a precise mosaic.

For vertical aerial surveys the camera points directly down. Oblique surveys, with the camera pointed at an angle, reveal more perspective detail but need optical rectification when subsequently used in map making. Aerial photographs taken under precisely specified conditions can serve for accurate measurements of ground details by stereoscopic evaluation (see below *Photogrammetry*).

Underwater photography. Underwater photography requires either special watertight cameras or pressure-resistant housings for normal cameras. In both cases, camera functions are controlled through pressure-tight glands. A flat glass or plastic window is usually in front of the camera lens. Light sources are battery-powered tungsten or tungsten-halogen lamps or electronic flash units (again in self-contained pressure-proof housings) or flash bulbs in suitable guns. For comfortable handling the weight of the housing with camera is adjusted to slight negative buoyancy. Complete camera and lighting outfits may be built into sledgelike or torpedolike units with an electric or compressed-air motor for self-propulsion through the water.

Since the refractive index ratio of glass to water is lower than for glass to air, the light-bending power of a glass lens is less in water than in air. This factor reduces the lens's angle of view and makes objects appear at about three-fourths of their actual distance. This difference must be allowed for in focussing—possibly by a suitably calibrated distance scale or by fitting the housing with a compensating porthole, which acts as a diverging lens.

Underwater cameras

Underwater cameras with lenses designed for direct contact with the water eliminate the air space between the lens and the porthole. Such lenses can cover wider angles of view without distortion, but they do not give sharp images outside the water.

As light absorption by the water—depending as it does not only on depth but also on water clarity—is difficult to estimate visually, exposure meters (in watertight cases) are needed. The red and yellow absorption of the water more than a few feet below the surface turns colour photographs taken by daylight into virtually monochrome shots; hence artificial light is essential to show up the full colour range of fish and other underwater subjects.

Close-range and large-scale photography. Near photography to reveal fine texture and detail covers several ranges: (1) close-up photography at image scales between 0.1 and 1 (one-tenth to full natural size); (2) macrophotography between natural size and 10–20× magnification, using the camera lens on its own; (3) photomicrography at magnifications above about 20×, combining the camera with a microscope; and (4) electron micrography with an electron microscope at magnifications of 10,000 to 1,000,000×, which involves photography of the electron microscope's phosphor screen or placing a photographic emulsion inside the vacuum chamber of the electron microscope to record directly the image formed by the electron beams.

Close-up and macrophotography. Supplementary close-up lenses or extension tubes (placed between the lens and camera body) allow the camera to focus on near distances for large scales of reproduction. Special close-up rangefinders or distance gauges establish exactly the correct camera-to-subject distance and precise framing of the subject field. Special simple close-up cameras, as in fingerprint recording and certain fields of medical photography, are permanently set to a fixed near distance and have a distance gauge or similar device built in. Screen-focussing cameras (view and single-lens reflex) need no such aids, as the finder screen shows the precise focus and framing.

Extension tubes or extension bellows or both or "macro" lenses of extended focussing range are used for the macro

range of distances. For optimum image quality, macrophotographic lenses specially corrected for large image scales may be used or the camera lens reversed back to front.

Photomicrography. There are two principal methods of photographing through a microscope. In the first, the camera, with its lens focussed at infinity, is lined up in the optical axis of the microscope, which is also focussed visually on infinity. In the other method, the camera without lens is positioned behind the microscope eyepiece, which is focussed to project the microscope image directly onto the film.

Special photomicrographic cameras generally employ the second method. Microscope adapters to provide a light-tight and rigid connection between the camera and microscope are available for both systems. Such microadaptors may incorporate their own shutter and a beam splitter system for viewing and focussing of the microscope image through a focussing telescope. Photomicrographs are the essential adjunct to all microscopy to record biological, bacteriological, physical, and other observations in black-and-white or colour.

Basis of depth perception

Stereoscopic and three-dimensional photography. Visual three-dimensional depth is perceived partly because of the fact that the human eyes see a scene from two viewpoints separated laterally by about 2½ inches (six centimetres). The two views show slightly different spatial relationships between near and distant objects (parallax); the visual process fuses these stereoscopic views into a three-dimensional impression. A similar impression is obtained by viewing a pair of stereoscopic photographs taken with two cameras or a twin camera with lenses 2½ inches apart, so that the left eye sees only the picture taken by the left-hand lens and the right eye only that of the right-hand lens. Binocular viewers or stereo-selective projection systems permit such viewing.

Stereo photographs can also be combined in a single picture by splitting up the images into narrow vertical strips and interlacing them. On superimposing a carefully aligned lenticular grid on the composite picture, an observer directly sees all the strips belonging to the left-eye picture with his left eye, and all the strips belonging to the right-eye picture with his right eye. Such parallax stereograms are seen in display advertising in shop windows or showcases. They also can be reproduced in print, overlayed by a lenticular pattern embossed in a plastic covering layer.

Photogrammetry makes use of stereo photography in measuring dimensions and shapes of ground objects in depth, as from successive exposure pairs made during an aerial survey flight. If all exposure parameters, including flying height, ground separation between exposures, and focal length of the aerial camera lens are known, the height of each ground feature can be measured. Photogrammetric plotting instruments do this and draw height contour curves of all features for aerial maps.

Similar photogrammetric evaluation of stereo photographs of nearby subjects can also be made. For instance, it is possible to reconstruct accurately the scene of a highway accident. In industry, a photogrammetric plot of an automobile model can be fed into a computer to program the machine tools that will shape the full-scale motor body components.

Infrared photography. Images formed by infrared and heat radiations can be recorded either directly, on films sensitive to them, or indirectly, by photographing the image produced by some other system registering infrared radiation.

Silver halide emulsions can be sensitized to infrared rays with wavelengths up to around 1,200 nanometres (one nanometre is $\frac{1}{1,000,000}$ of a millimetre). The usual sensitivity range is 800 to 1,000 nanometres. Direct infrared-recording aerial photography shows up ground features of differential infrared reflection but similar light reflection (*e.g.*, different types of foliage) and cuts through haze and mist. Special colour films with an infrared-sensitive layer and processed to colours different from the natural rendering (false-colour films) show up such differences still more clearly. In forensic photography, infrared pictures reveal ink alterations in forgeries, differentiate stains, and help to identify specific textiles and other materials. In medicine, infrared photographs show subcutaneous blood vessels, as the skin is transparent to infrared.

Applications of direct infrared photography

With suitable equipment, it is possible to convert an infrared image into one visible on a fluorescent screen, where it can be photographed. In infrared scanner systems a moving mirror scans the object or scene and focusses the radiation onto an infrared-sensitive cell. The cell generates electric signals to modulate a light source, which, in turn, scans a photographic film or paper synchronously with the mirror. The resulting image records hotter and colder parts of the object as lighter and darker areas and can establish actual temperatures of subject details—sometimes to within a fraction of a degree. This system has been used to record temperature variations in the skin for the diagnosis of cancer.

Ultraviolet photography. Invisible shortwave ultraviolet radiations can be recorded directly or used in fluorescence photography (see Figure 14).

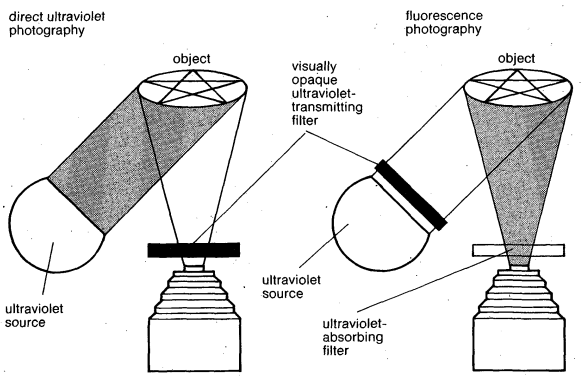


Figure 14: Arrangements for ultraviolet photography.

For direct ultraviolet recording, the photographically useful wavelength range lies between 400 nanometres (visible violet) and about 200 nanometres and needs special optical systems transparent to ultraviolet rays (quartz, silica, or fluoride elements or combinations thereof). Light sources rich in ultraviolet such as mercury vapour lamps—with an ultraviolet-transmitting, but visually opaque, filter in front of the camera lens—ensure that the photograph records only the ultraviolet-reflecting characteristics of the subject.

Fluorescence photography records the glow or visible light given off by certain substances when they are irradiated by ultraviolet rays. The object is illuminated only by ultraviolet rays by screening out the visible light with a filter that transmits only ultraviolet radiation, and another filter that absorbs the ultraviolet rays is placed over the camera lens, permitting only the visible light (fluorescence) to be recorded on the film. Normal lenses and panchromatic or colour materials are used.

Applications of ultraviolet photography

Ultraviolet photography can identify or separate pigments and fabrics and can detect forgeries of documents. Fluorescence photography can identify dyes, stains, and markings, specific chemical substances, and fluorescent components in microscope specimens. Ultraviolet microscopy offers increased resolution through the shorter wavelength radiations employed. Aerial and satellite photography by ultraviolet can show up ultraviolet-reflective ground features.

Radiography and other radiation recording techniques. Silver halide emulsions are sensitive to X-rays, gamma rays, and charged particles emitted by radioactive substances. Some of these rays penetrate visually opaque materials to varying degrees to show up internal structures. Radiography covers techniques of recording the subsurface features of objects.

X-ray radiography. X-rays (wavelengths between $\frac{1}{100}$ and $\frac{1}{100,000}$ that of visible light) are produced by high-voltage electron streams bombarding an electrode in a vacuum tube. For radiography the object to be

recorded is placed between an X-ray tube and the film; the film registers the differential absorption of the X-rays by the object's internal structure as a projection shadowgraph.

The most familiar application is in medicine for diagnosis and recording, including dental radiography. Industrial radiography permits nondestructive inspection of castings, welds, and engineering structures.

Gamma radiography. This technique of gamma-ray radiography is similar to that of X-ray radiography but relies on rays emitted by radioactive substances (*e.g.*, artificial isotopes like cobalt-60). Gamma rays have wavelengths from 100 to 1,000 times shorter than X-rays and correspondingly greater penetrating power. Small gamma-ray sources are placed in areas inaccessible to X-ray tubes such as inside pipelines. In all radiographic applications, the exposure occurs under conditions of normal light, from which the radiographic film is protected by a light-tight (but radiation-transparent) wrapping.

Autoradiography. Autoradiography records the distribution of radioactive materials in botanical and histological specimens placed in contact with a photographic emulsion. This technique has been applied to the study of metabolism of plants and animals; it records the activity of organic compounds of radioactive isotopes introduced into the system of the plant or animal. In engineering studies, autoradiography can be used to follow the transfer of radioactive substances from one surface to another in lubrication. The technique also has applications in machining and other metal-treatment processes.

Nuclear-track recording. Tracks of elementary particles, such as protons, electrons, and mesons, produced by nuclear reactions, can be recorded by photographic means. The most common technique is to photograph the visible traces of such tracks in bubble or spark chambers with special camera and lens arrangements. Different arrangements can provide for coverage of large fields or the recording of tracks simultaneously from several directions for three-dimensional reconstruction.

Particle tracks can be recorded directly in thick (up to one millimetre) emulsion layers or in emulsion stacks (up to 20 inches) carried in high-altitude balloons and in spacecraft and satellites. Special processing procedures are required to deal with these emulsion thicknesses.

Astronomical photography. By the cumulative effect of light received over a long period, a photographic emulsion can record celestial objects too faint to be visible. Before radio telescopes (see TELESCOPE), photography was the only way of detecting many such objects.

Astronomical cameras as units of telescopes

Astronomical cameras are film- or plate-holding units built onto high-power telescopes, typically reflecting systems. The telescopes run on precision, clock-driven mounts to keep the optical axis stationary with respect to the sky area as the Earth rotates during an exposure time, which can run into several hours. For increased recording sensitivity, the telescope image may be intensified electronically.

Astronomical photographs taken through narrow-band colour filters—including infrared or ultraviolet transmitting filters—show selective emission characteristics of stars. In the case of the Sun and of planets, such photographs can reveal some surface details not observable by white light. Colour photographs reveal colours not directly visible because the intensity of starlight is too low to stimulate the eye's colour-vision mechanism.

Spectrography records the composition of light emitted by stars and other objects, the star image of the telescope being photographed through a diffraction grating, a device that disperses white light into constituent wavelengths. Elements present in the star or the gas mantle surrounding it can be identified from their characteristic spectral lines. Displacement of such lines from their known wavelength position can indicate the velocity with which the distant stellar systems recede from or approach the Earth.

Microfilming and micro-reproduction. Microfilming is the copying of documents, drawings, and other such matter at a reduced scale—typically 1:15 to 1:25—for compact storage. Complete micro-reproduction systems include methods of filing the film copies for easy retrieval and re-enlargement.

Documents, periodicals, and other printed matter are usually microfilmed on 16-mm film with an image size between 10 × 14 and 14 × 20 mm in a copying camera taking 100-foot lengths of film. Engineering drawings of high information content are microfilmed on 35-mm unperforated film with a standard image size of 32 × 45 mm. Films of up to 105 mm in width are also used. Automated microfilm cameras run continuously, documents being fed onto a moving band carried past the camera at a steady speed while the film runs past a slit at a matched rate.

Readers and reader printers are desk-top projectors that display the frames re-enlarged to about natural size on a back projection screen. In a reader printer the image may also be projected on sensitized paper for full-size enlargements. Advanced readers have elaborate retrieval systems based on frame coding and run the microfilm rolls through at high speed until a specific searched image is reached.

Aperture cards or standard-size transparent jackets store microfilm images as single frames or groups of frames. Such unitized microfilms permit easier indexing and retrieval by classical or punched card indexing systems. Certain 35-mm microfilm cameras photograph the original document directly on film premounted in an aperture card and processed on the spot.

Microfiche system

A unitized system called microfiche carries up to 70 frames each about 11 × 16 mm on a 4 × 6-inch sheet of film—the equivalent of a 70-page document. The microfiche camera repositions the film frame by frame after every exposure. Microfiche can also be produced by mounting strips of 16-mm microfilm on self-adhesive 4 × 6-inch film sheets or jacketing such strips in multichannel plastic jackets of the same format.

Various duplication methods allow microfilm records to be extensively distributed.

For still greater space saving, microfilm images may be reduced beyond 1:100 on high-resolution photochromic image materials. Extreme fine-grain silver copies then hold 3,000 to 4,000 individual frames on a single 4 × 6-inch film. This method, proposed for complex catalogues and like purposes, offers easy retrieval of individual frames but requires a high-magnification reader.

V. The photography industry

Equipment manufacture

Present-day manufacture of cameras and other photographic equipment is concentrated in mass-production plants that make most of the components (camera bodies, lenses, shutters, and other parts) on more or less automated machines; the components are then assembled by semiskilled or skilled labour. Smaller manufacturers of low- and medium-priced cameras obtain components for assembly from such specialist suppliers as shutter manufacturers (about half a dozen leading firms in the world) and lens producers. High-quality precision cameras are produced on a smaller scale with automated fabrication of the engineering components but much more extensive manual assembly by highly skilled technicians. Components and functions of every camera are tested at every production stage; less expensive cameras are usually batch tested by a sampling procedure.

The raw material for lens manufacture covers a range of optical glasses of different optical characteristics. About ten major worldwide glass producers supply the several dozen optical firms offering lenses of well-known brands. The glass is cast into blanks for specific lens elements, ground and polished to the required exact specifications, and the elements assembled in metal (sometimes plastic) mounts. Extensive production tests and optical performance checks safeguard quality standards.

Silver halide emulsions are made by mixing silver nitrate with a solution of alkali halide—typically potassium bromide and iodide—in gelatin. The silver halide then precipitates out as fine crystals. After cooling to a jelly, shredding, and washing, the emulsion is remelted and treated to increase speed and contrast. Colour sensitizers and additives are introduced, and the gelatin emulsion is machine coated on wide continuous webs of paper or film. Gen-

erally several coatings are applied. Operations from emulsion mixing onward are carried on in total darkness. After cooling and drying, the material is batch tested for consistent characteristics then cut and packed.

Film base is produced either by coating a solution of the base material on large drums where it solidifies (film casting) or by extrusion of plastics, such as polyester, in film extruders. For print materials, paper of suitable purity is coated with a barium sulfate emulsion in gelatin, to provide a smooth white surface, and then with the silver halide emulsion.

Photofinishing laboratories process most amateur and some professional photographers' films and prints. In 1969, the total volume of business of the 2,000 to 3,000 laboratories in the United States was estimated at $850,000,000, of which $750,000,000 was colour processing.

Continuous processing of films

Photofinishing laboratories use machines that carry the films in spliced-together lengths or on racks through successive tanks of the processing solutions. Prints are usually made to standard formats on automatic enlargers, taking both the negatives and the paper in continuous rolls. The paper rolls of 250 or 500 feet are processed in continuous-strip processors, which deliver prints dry and with a glossy surface ready for automatic cutting. Many printers have automatic exposure measurement based on overall negative density, with automatic controlled colour correction for colour negatives. High-capacity colour printers of this type can produce 2,000 to 3,000 prints per hour. Coding systems identify individual films and corresponding prints by customer or order number for final re-sorting. More exacting processing services grade colour negatives before printing by light transmission measurements through different colour filters; the resulting exposure data may be punched as edge codes in the film itself or programmed on perforated paper tape. When the tape is run through the printer together with the film, the perforations directly control the colour exposures and corrections.

Enlargements to special sizes and colour printing for professional photographers require individual enlarging by skilled personnel on conventional enlargers with advanced automation features of focussing, exposure measurement, and colour control.

Other processing services include duplication of transparencies, various types of photocopying (partly on coin-operated copiers set up in public places), microfilming, and microfilm processing.

Industry economics

The world's largest producer of photographic goods is the United States, followed by Japan, West Germany, Great Britain, and France. The photo industry in the United States employs some 112,000 persons; the German and Japanese industries about as many again between them.

The world's major photo exporters are the United States, followed by Japan and West Germany. Principal foreign markets of United States photo products are Canada, Japan, West Germany, and Great Britain. The United States is also the largest photo importer, its main suppliers being Japan, West Germany, Belgium, and Great Britain.

In the United States the ratio of photo- (including movie) equipment production to photo-chemical production (films, papers, chemicals, etc.) is about 3:4; the corresponding ratio in Germany is 3:2, in Japan nearly 2:1 and in Great Britain about 1:2. Approximately 35 to 40 percent of sensitized materials production is bought by amateurs, the remainder for professional purposes, including applied photography and industrial recording. This ratio is changing in favour of professional uses.

BIBLIOGRAPHY

General Works: L.P. CLERC, *La Technique photographique* (Eng. trans., *Photography: Theory and Practice,* 6 pt., ed. by D.A. SPENCER and rev. by L.A. MANNHEIM *et al.*, 1970–71), classic treatise on the technology of photography; L.A. MANNHEIM *et al.* (eds.), *The Focal Encyclopedia of Photography,* rev. ed., 2 vol. (1965); C.E.K. MEES (ed.), *The Theory of the Photographic Process,* 3rd ed. by T.H. JAMES (1966), classic work on basic photography theory; H.J. WALLS, *How Photography Works* (1959), scientific background information on photographic processes.

History: W. BAIER, *A Source Book of Photographic History* (1964), a review with detailed bibliographies and literature references to photographic developments; J.S. FRIEDMAN, *The History of Color Photography,* 2nd ed. (1968); H. GERNSHEIM, *The History of Photography from the Earliest Use of the Camera Obscura in the Eleventh Century up to 1914* (1955), on the evolution of photographic equipment, techniques, and practices as well as of pictorial photography; E. STENGER, *Siegeszug der Photographie in Kultur, Wissenschaft, Technik* (1950; Eng. trans., *The March of Photography,* 1958).

Camera technology: LIFE LIBRARY OF PHOTOGRAPHY, *The Camera* (1970), elementary pictorial description of camera functions and use; A. MATHESON and W.D. EMANUEL (comps.), *Cameras: The Facts* (annual, 1957–64), detailed descriptions and diagrammatic illustrations of principal camera models on the market; J. STUPER, *Die Photographische Kamera* (1962), detailed treatise of camera components, and their design and operation; H.J. WALLS, *Camera Techniques: Fundamentals and Equipment* (1960).

Lenses and optics: H.M. BRANDT, *Das Photo-Objectiv* (1956; Eng. trans., *The Photographic Lens,* 1968), principles of geometric optics, photographic lens types, and the evolution of modern photographic lenses, with market-wide lens tabulation; A. COX, *Photographic Optics,* 14th ed. (1971), classic manual of lens principles and use with reference to image quality, measurement, and modern techniques, and *A System of Optical Design* (1964), basics of image assessment and of design techniques with a survey of current lens types (includes extensive mathematical treatment); C.B. NEBLETTE and A.E. MURRAY, *Photographic Lenses* (1965).

Films and taking technique: W.F. BERG, *Exposure: The Fundamentals of Camera Technique,* 4th ed. (1971), basic theory and practice; M. LANGFORD, *Basic Photography* (1965) and *Advanced Photography* (1970), manuals of practical technique for professional photographers; LIFE LIBRARY OF PHOTOGRAPHY, *Light and Film* (1970), elementary and pictorial presentation of taking techniques; W. NURNBERG, *Lighting for Photography,* 16th rev. ed. (1968), analysis of illumination techniques.

Processing and printing: C.I. JACOBSON, *Developing,* 17th ed. (1970), manual of all aspects of negative technique; and with L.A. MANNHEIM, *Enlarging,* 21st ed. (1972), manual of positive technique in black and white and colour; D.H.O. JOHN and G.T.J. FIELD, *A Textbook of Photographic Chemistry* (1963), basics of chemical reactions in black-and-white processing; L.F.A. MASON, *Photographic Processing Chemistry* (1966), detailed treatment of processing mechanisms and reactions.

Colour photography: R.M. EVANS, W.T. HANSON, and W.L. BREWER, *Principles of Color Photography* (1953), principles of colour rendering, response, and reproduction; R.W.G. HUNT, *The Reproduction of Colour,* 2nd ed. (1967), standard handbook with moderately advanced mathematical treatment; D.A. SPENCER, *Colour Photography in Practice,* rev. by L.A. MANNHEIM and VISCOUNT HANWORTH (1969), theory and practical techniques.

Special techniques and applied photography: R.F. SAXE, *High-Speed Photography* (1966), a condensed but comprehensive survey; J.B. HERSEY (ed.), *Deep-Sea Photography* (1967); O.R. CROY, *Alles über Nahaufnahmen* (1960; *Eng.* trans., *Camera Close-up,* 1961), practical how-to-do-it manual; J. BERGNER, E. GELBKE, and W. MEHLISS, *Einführung in die praktische Mikrofotographie* (1961; Eng. trans., *Practical Photomicrography* 1966), a comprehensive manual; N.A. VALYUS, *Stereoscopy* (1966; orig. pub. in Russian, 1962), comprehensive survey of technology and applications; W. CLARK, *Photography by Infrared,* 2nd ed. (1946); EASTMAN KODAK COMPANY, *Infrared and Ultraviolet Photography,* 6th ed. (1959); J.A. CROWTHER, *Handbook of Industrial Radiography* (1945); J. GIEBELHAUSEN, *Photography in Industry* (1967); C.E. ENGEL (ed.), *Photography for the Scientist* (1968); C.R. ARNOLD, J. ROLLS, and J.C.J. STEWART, *Applied Photography* (1971).

Photographic industry: Wolfman Report on the Photographic Industry in the United States (annual); statistics and reports of the German Photographic Industry Association and of the Japanese Camera Industry Association.

(L.A.Ma.)

Photometry, Astronomical

Astronomical photometry is a branch of astronomy dealing with the measurement and study of the brightness of celestial objects. In astronomy, the term photometry is most commonly applied to measurements of brightness that extend, in any one observation, over a limited band

of wavelengths only, rather than over all wavelengths of light. The relationship of the dominant, or effective, wavelength (represented by λ_0) to the bandwidth being observed (represented by $\Delta\lambda$) is a ratio ($\Delta\lambda/\lambda_0$) that is typically between one-third and one-three-hundredth. In contrast, the terms radiometry and bolometry, in astronomical usage, are applied to brightness measurements comprising all wavelengths, while the term spectrophotometry is generally used when only a very narrow band of wavelengths is measured at one time; *i.e.*, when the ratio $\Delta\lambda/\lambda_0$ is chosen close to zero. (For the related study of the electromagnetic spectra produced by celestial objects, see SPECTROSCOPY, PRINCIPLES OF.)

Originally, the role of photometry was to help identify stars in the compilation of star catalogs and charts. As photometric methods became more sophisticated, photometry became a developed branch of astronomy aimed toward the determination of the physical properties of stars; it is also used extensively in studies of most other classes of celestial objects, each of which requires special consideration and treatment or, in some instances, a special technique.

Historical survey. Photometry is one of the earliest branches of astronomy, the variety of brightnesses of the stars and planets having been obvious from time immemorial. Originally, it sufficed to arrange these celestial objects subjectively in sequence from the brightest to the faintest, as was done by Hipparchus around 130 BC. His work was expanded by Ptolemy about 300 years later to include about 1,000 stars. Both men used a system of "magnitudes," in which the brightest stars were ranked first magnitude and the faintest perceptible to the naked eye sixth. Further classification of the naked-eye stars by magnitudes was done in the late 16th and early 17th centuries by Tycho Brahe and Johann Bayer. After the telescope was invented, classification was rapidly extended to fainter stars by Galileo and others. All these early observers employed subjective estimates of stellar magnitude (sometimes using their own terminology and units, such as grades). As a result, there was little homogeneity to their systems.

The need for a comprehensive and consistent set of visually estimated stellar magnitudes was first met by Friedrich W.A. Argelander in several works, culminating with his great *Bonner Durchmusterung* ("Bonn Survey"), a catalog that gave magnitudes for about a third of a million stars in the northern half of the sky, down to about the tenth magnitude. The magnitude intervals were subdivided into ten subclasses, corresponding closely to the least perceptible differences in brightness as judged by eye; thus, Argelander's magnitude estimates appear in decimal form. For example, 4.3 designates a stellar brightness three subclasses fainter than fourth magnitude. Argelander fitted his magnitude scale to the traditional one of six magnitude classes for the naked-eye stars. This system of magnitudes, while convenient and useful, was entirely subjective in nature and lacked a physical or quantitative definition.

Early in the 19th century, astronomers began to intercompare quantitatively the brightness of stellar images with visual photometers, devices with which the observer matches two star images, one of which may be that of an artificial star, in the field of view of a telescope. Gustav Fechner, a German scientist, observed that the response of the eye was characterized by approximately equal sensation for equal ratios of brightness, while others found that the old magnitude intervals corresponded to brightness ratios of between 2.2 and 2.8—or, in other words, stars of the second magnitude, for example, were from 2.2 to 2.8 times brighter than stars of the third magnitude. The English astronomer Norman Robert Pogson proposed in 1856 the adoption of 2.512 (the fifth root of 100) as the formal basis of a physically defined magnitude scale. This, the Pogson scale, is now universally adopted by astronomers and may be stated conveniently: the difference in magnitude between two stars is given by 2.5 times the logarithm (to base 10) of the brightness ratio. If the ratio is obtained from measurements at a specific wavelength, the result is a monochromatic (one-colour) magnitude; if the ratio is obtained in terms of the response of a particular light receiver to a specific band of wavelengths, the result is a heterochromatic (more than one colour) magnitude in what is called a natural system characteristic of that photometric instrument and technique.

Pogson scale of magnitude

Adopting the Pogson scale, astronomers made use of visual photometers toward the end of the 19th century to re-establish the magnitude system on a physical basis, while they fixed the magnitudes in the range from fourth to sixth to agree with the ancient and the *Durchmusterung* values. Several new catalogs achieved improved accuracy because of the physically defined scale. The most important of these were the *Potsdamer Photometrische Durchmusterung,* which included 14,199 stars brighter than 7.5 magnitude in the Northern Hemisphere, and the *Revised Harvard Photometry,* which included 9,110 stars brighter than 6.5 magnitude over the entire sky as well as 36,682 fainter stars to about ninth magnitude.

In the last decade of the 19th century, bright stars were found to produce photographic images larger and denser than those of faint stars, so that measures of these image properties could be used to establish magnitudes. It is notable, however, that the dimensions of the optical images of stars in the focal plane of a telescope are substantially identical for all stars, bright and faint. The form and dimensions of the image are determined chiefly by the disturbing influences of atmospheric turbulence upon the diffraction pattern formed by the objective. The optical images from bright and faint stars differ only in energy flux. On the other hand, photographic images of bright stars are larger because of the turbidity and halation (a blurring caused by reflection) characteristics of the photographic emulsion; in particular, the size of a stellar photographic image is not a manifestation of the star's diameter.

Photographic photometry, despite calibration difficulties, soon largely displaced visual photometry. The photographic method not only determined photographic magnitudes in the blue spectral region, to which it was primarily sensitive, but, with suitable colour filters and orthochromatic emulsions, established photovisual magnitudes that closely matched the colour characteristics of the visual system of magnitudes. The photographic and photovisual magnitudes were adopted as the International System and served until about the middle of the 20th century, when photoelectric techniques were adopted because of their greater accuracy.

The International System of photometry

The earliest efforts to measure stellar brightness by means of the photoelectric effect, the conversion of light to electric current, were made in 1913. The technique was restricted to the brightest stars, however, because of inadequate sensitivity, although its inherent accuracy surpassed that of all other photometric methods. One of the chief difficulties was the accurate measurement of the feeble current produced by photocells. Until electronic amplifiers displaced the more primitive electrometers, photoelectric photometry remained severely handicapped. In 1932 in the United States Albert Edward Whitford made use of an electronic amplifier for the first time in astronomy. A second major advance took place when practical photomultipliers (see below *Principles of photometric methods: General techniques*) became available after World War II. Photoelectric methods now surpass all others in reliability, accuracy, and sensitivity.

The UBV system. What is called the UBV system of stellar magnitudes has been widely adopted as a successor to the International System. It is based on photoelectric photometry in three broad bands: the ultraviolet (U), blue (B), and visual (V) spectral regions; the latter two are similar to those of the old photographic and visual magnitudes. The additional U band includes the important spectral region that includes the Balmer discontinuity, in which light intensity drops suddenly owing to absorption by ionized hydrogen.

Apart from the Balmer discontinuity and the spectral lines, the distribution of stellar radiation with wavelength is known to approximate roughly the distribution of radiation from what the physicist calls blackbodies (*i.e.*,

bodies emitting radiation freely at all wavelengths) with temperatures ranging from about 2,000° to 50,000° K (0° K equals −273° C or −460° F). The relative strengths of the B and V magnitudes vary with stellar temperature and consequently with spectral type. For a given star, the quantity B–V, termed a colour index, expresses in magnitude form the ratio of brightnesses in two different wavelength bands. It might be expected, therefore, that a star's spectral type could be inferred from its B–V colour index; but this inference is valid only when the starlight has not passed through regions of interstellar dust. The dust scatters light more strongly at short wavelengths; and for every three magnitudes lost in the V band there are about four lost in B, resulting in an apparent reddening of the B–V colour index by about one magnitude. The apparent B–V of a star must therefore be corrected for this reddening, called colour excess, before any inference of the star's spectral type can be made. It is here that a third band in the ultraviolet plays a role, as it makes possible the determination of colour excess, total absorption, and spectral type for many stars. This kind of information about stars can be obtained more certainly when the UBV photometric data are supplemented by spectroscopic data, but UBV photometry can be carried out on stars far too faint for spectroscopic studies.

The colour index

Many more stars have been measured in the UBV than in any other photoelectric system. Such measurements are available to observers as standards of reference over the entire sky. Included are stars of most spectral types, luminosities, and degrees of reddening. The magnitudes have been made to coincide closely with the historical magnitudes for the bright stars; that is, the zero point for the V magnitudes has been fitted to selected stars in the North Polar Sequence, which in turn had been fitted to the old *Durchmusterung* values. The B–V and U–B colour indices are defined as zero for A0 stars of normal luminosity, thus conforming to the earlier zero-point convention. On the whole, the UBV bands are readily duplicated to a sufficient accuracy by appropriate combinations of photomultipliers and standard filters as well as standard photographic emulsions whenever photographic photometry is desired.

Principles of photometric methods. *General techniques.* In astronomical photometry, the process starts at the focal plane of a telescope, and from the start the principal methods differ greatly.

In photoelectric photometry, the image of a single star is passed through a small diaphragm in the focal plane and falls upon a photoelectric detector; all other images in the focal plane are blocked off and wasted. Photographic photometry, on the other hand, has the advantage of recording simultaneously all the images to which the plate is exposed. An additional advantage of the photographic method is that the plate becomes a permanent record, useful also for purposes other than photometry. Unfortunately, these advantages can be gained only at the cost of considerably reduced accuracy, for the complex relationships between the size and density of photographic images of stars and the brightness of the optical images are not subject to full control or accurate calibration. Nonuniform sensitivity or development over the surface of the emulsion, other photographic effects, the atmospheric dispersion and turbulence as they affect the chromatic form of the images, and field errors caused in the telescope optics all act to introduce large accidental and systematic errors.

Advantages and disadvantages of photography

Photoelectric photometry owes its accuracy primarily to two factors: the highly linear relation between the amount of incoming radiation and the electric current it produces and the use of precise electrical measurement techniques. In contrast, the photochemical processes and densitometric techniques (measuring the degree of darkening of photographic film) used in photographic calibration are relatively crude. For this reason, virtually all modern magnitude scales are established with photoelectric photometers, whereas the photographic methods are used only where their advantages may be exploited and deficiencies minimized. For example, after photoelectric measurements have been made, in a small area of the sky, of a select few individual stars that embrace an adequate range of magnitude and colour index, a photograph of the same area offers a means of determining the magnitudes of many more stars by interpolation. This simultaneously eliminates some of the chief shortcomings of the photographic method and provides much data with little telescope observing time.

The two chief stellar photometric instruments in current use are the photoelectric photometer, which attaches directly to the telescope, and the iris photometer, a device used to measure the degree of darkening of images on a photographic plate. As noted above, a diaphragm in the photoelectric photometer's focal plane serves to isolate the single image to be measured. Large diaphragms admit images of extended objects such as nebulae or planets; small ones isolate a star from a close companion or reduce the amount of background light from the sky when the star to be measured is very faint.

Two principal photometric instruments

The light comprising the stellar image passes through the diaphragm, an appropriate filter, a field lens, and into a photomultiplier, a device that produces a relatively strong electric current from a weak light input. The purpose of the field lens is to focus a small image of the telescope's entrance pupil onto the input of the photomultiplier. Typical output currents for stellar measurement lie in the range from 10^{-5} to 10^{-11} ampere, despite the large amplification factor achieved within the photomultiplier, typically around 10^6. The output current may be measured in a variety of ways.

A photoelectric photometer has some viewing eyepieces to allow selection of and alignment on the object to be measured or to permit guiding on a separate stellar image. There may also be provisions for simultaneous observing of several colours or multiple images or for scanning across an image.

In order to retain the highest accuracy in photoelectric photometry, all of the instrumentation must be stable, accurately calibrated, and free from effects that are dependent on such extraneous influences as temperature or orientation in the Earth's magnetic and gravitational fields. But even when these effects are minimized, there remain limitations to accuracy imposed by the atmosphere.

Limitation on accuracy

Because the transparency of the Earth's atmosphere is neither perfect nor constant and may not be uniform around the sky, it is usually necessary to compare photoelectric measurements of stars quickly and repeatedly, using nights of the greatest clarity. (Exceptions to this rule are sometimes permissible; for example, when the instrument is equipped to compare two nearby images simultaneously.) Reference stars of known magnitude as near the program stars as possible are customarily chosen to facilitate rapid manipulation of the telescope and to minimize the influence of differences in atmospheric conditions. The atmosphere's extinction of starlight is dependent on both the path length traversed by the light and the transparency of the air; it also depends on colour, being more pronounced at shorter wavelengths. Accordingly, all photoelectric measurements require slight corrections, since seldom will the stars be observed through identical lengths of air path. It is customary to apply these corrections to convert the magnitudes into what they would be in the absence of an atmosphere.

In the iris photometer, a beam of light is made to pass through the photographic plate, and the diameter of the beam is adjusted by means of an iris diaphragm to always pass a constant amount of light around and through a stellar image. The iris diameter thus adjusted will usually be found to be approximately proportional to the magnitudes of the imaged stars, over a range of several magnitudes. The stars of the calibrated sequence provide the link between arbitrary iris diameters and the magnitudes and permit the magnitudes of the program stars to be determined by interpolation. Another method of calibrating the image properties with magnitudes, believed by some to be superior in accuracy and range, consists of studying the image profiles with a microdensitometer; *i.e.*, by passing through them a very fine pencil beam of light.

Use of the iris photometer

Visual photometry continues in use in a few special areas of research. For extremely close visual binary stars, no other method has proved successful in measuring the magnitude of each component separately or in obtaining a difference in magnitude, Δm. Experienced observers of double stars attempt to estimate Δm directly, but the reliability varies with the seeing quality and proximity and Δm of the pair. Some observers of double stars make use of a double-image polarizing eyepiece that provides for equalizing the brightnesses in calibrated fashion; accuracies of about 0.05 magnitude can be obtained, although the measures are difficult.

Visual estimates of stellar brightness are also made by dedicated hobbyists, a particularly valuable pastime for those who observe the long-period variable stars whose slow variation few major observatories can follow thoroughly or economically. Although such visual estimates of magnitude seldom are more accurate than 0.1 magnitude, the amplitudes of variation of the long-period variables are so large that this accuracy is sufficient for most purposes.

From photographs of wider double stars, estimates of the difference in magnitude can also be made profitably by visual inspection, particularly when an objective grating has been used to limit the apparent difference in magnitude to 0.5 magnitude or less.

Selected areas; transfer photometry. At one time, the magnitudes of stars in one region of the sky were determined by making a photograph of the desired region and superimposing on it a second exposure of a region containing stars of known magnitude. In this process, termed transfer photometry, both the magnitude scale and the zero point of one region were effectively transferred to another. The reference regions were set up in several places in the sky, known as selected areas. Because one of the regions was centred about the north celestial pole and was therefore observable at all seasons in the Northern Hemisphere, it became the standard reference region. The photometrically calibrated stars in it make up the North Polar Sequence. The accuracy of the photographically determined magnitudes in this and other selected areas is insufficient to warrant their continued use; indeed, systematic errors of scale amount to as much as 30 percent at the 18th magnitude. All modern transfer work is therefore done with photoelectric photometers.

North Polar Sequence of stars

Broad-band photometry. If all stars were of identical colour—that is, of identical spectral energy distribution—their apparent magnitudes would be independent of the wavelength chosen for measurement and of the bandwidth used. In this idealized situation, any detector could be employed without spectral restrictions; and consequently, an observer would exploit the full spectral response of his most sensitive detectors in order to measure the magnitudes of the faintest objects in the most efficient way. In fact, however, the stars do have a wide variety of spectral energy distributions, and this imposes the necessity of determining magnitudes in restricted bands of colour.

From the definition given earlier of the Pogson magnitude scale, it would appear that monochromatic magnitudes pertaining to a specific wavelength are best suited to physical interpretation, because comparing the brightnesses of stars at wavelengths that might differ among stars of different energy distribution would have little physical meaning. But in order to study the most distant stars, whose apparent intensity is inherently so feeble, moderately broad bands are needed to provide sufficient energy to be detectable by even the most sensitive light receivers. Thus, the astronomer is compelled to make a compromise between a clear physical definition and the need to use sufficiently broad bands to examine stars in a consistent fashion, from the brightest to the faintest. In this discussion of the way in which broad-band magnitudes may be used and interpreted, it will simplify matters to exclude from consideration peculiar stars, novae, emission nebulae, and other sources of nonthermal radiation whose spectral energy distribution may be entirely dissimilar to that of most stars.

With any broad spectral band there can be associated a representative wavelength, λ_r, at which a monochromatic band of arbitrary narrowness would produce the same system of magnitudes for stars of a given spectral energy distribution when measured with the same telescope and detector system. Accordingly, the representative wavelength, λ_r, can be defined by an equation that includes variables for photometer response and for stars of differing spectral energy distribution. In comparing the values that the representative wavelength takes for the hottest and coolest stars, it is found that its range of variation depends on both the width and symmetry of the band in question. As a typical example, the V and B bands of the UBV system have values of representative wavelength that vary by about 2 or 3 percent over the normal range of stellar energy distributions. Experience has shown that this departure from monochromaticity is quite tolerable in the context of many applications, some of which will be mentioned presently, but that it is unacceptable for certain more critical applications, which will be taken up in the following section.

Representative wavelength of a broad spectral band

In practice, a broad-band-magnitude system is readily defined by (1) a set of magnitudes for specified stars as measured with a particular photometric apparatus and (2) a specification of the chief components of the apparatus, such as colour filters and type of detector, or, alternatively a specification of the instrument's response function. Although exact duplication of the apparatus response function cannot be expected in other photometers, sufficiently close approximation can usually be achieved, and small corrective transformations can be applied to produce results virtually consistent with those of the original equipment. It can be shown that broad-band magnitudes are in many instances adequate approximations of monochromatic magnitudes when the wavelength of the latter is defined by an equation that is somewhat different, though analogous, to that for λ_r. This is designated as the effective wavelength, λ_0, and it lies within the range through which λ_r varies over the range of stellar energy distribution encountered among stars and can be thought of simply as a kind of mean value of λ_r.

Some examples of applications to which broad bands are suited have been mentioned earlier: estimating the spectral types of stars as well as measuring the amount of reddening and absorption influencing their apparent colour indices and magnitudes. (In instances of extreme reddening, small corrections may be required if λ_r differs significantly from λ_0.) Other examples are stellar distances derived photometrically and absolute magnitudes.

Distances of remote stars can be deduced from their apparent brightnesses. In principle, it suffices to compare the apparent brightness of a star of unknown distance with that of one of the same spectral type and luminosity class at a known distance, since the ratio of apparent brightnesses (due allowance being made for any absorption) is simply the ratio of the square of their distances. A more customary way of treating this problem involves the concept of absolute magnitude, which is taken to be the apparent magnitude that a star would have if it were to be viewed from a standard distance of ten parsecs (about 32.6 light-years). The relation between the absolute and apparent magnitudes, M, m, and the distance, d, expressed in parsecs can be derived from the inverse-square law of diminishing brightness in magnitude form; and it states that the absolute magnitude is equal to the apparent magnitude minus a logarithmic value of the distance:

The concept of absolute magnitude

$$M = m - 5 \log \frac{d}{10}.$$

If there is any interstellar absorption, A, then it is also subtracted from the apparent magnitude:

$$M = m - A - 5 \log \frac{d}{10}.$$

These relationships serve to define a set of absolute magnitudes for stars of known distance, so that the relative intrinsic brightnesses and total energy outputs of different types of stars can be determined. A profoundly important relationship between the absolute magnitudes and spectral types of stars was established in this way by Ejnar Hertz-

sprung and Henry Norris Russell early in the 20th century and elaborated upon by numerous subsequent investigators. The relationship, commonly referred to as the Hertzsprung-Russell diagram from the form in which it is usually displayed, has played a fundamental role in studies of the constitution and evolution of the stars.

Conversely, the absolute magnitude of a star can be inferred from its spectral type and luminosity class, and this inference together with its apparent magnitude permits a determination of its distance.

Broad-band systems, such as the UBV system and its predecessors, have been of immense value in providing information about the stars in the Galaxy, extragalactic distances, and other important basic and statistical data. For this kind of application, monochromatic magnitudes would not have been significantly better.

Narrow-band photometry. Several types of investigation require narrow bands even at the cost of limiting sensitivity and thereby excluding application to the faintest stars. In general, the narrower and more numerous the bands are within a spectral range, the more may the photometric methods approach the spectroscopic in terms of the number of independent parameters that can be determined. In addition, the narrowing of a band toward monochromaticity aids in physical interpretation, particularly in relating the photometric data to theoretical stellar models and to spectrophotometry.

Two general types of narrow-band photometry have been profitably applied to stellar problems: one in which the strength of a stellar absorption line is measured, the other in which selected regions of the continuum are measured. In the former, the bandwidth is typically very narrow, ten to 30 Å (angstroms) wide, while in the second the bands are of intermediate width, typically 150 to 300 Å (one angstrom equals 10^{-10} metre).

The hydrogen-beta system

As an example of the former, the Hβ (hydrogen-beta) system introduced by Bengt Strömgren is based on measures of the strength of the Balmer Hβ line, a hydrogen absorption line at 4861 Å, which, for the stars of spectral type O9 to B9, is well correlated with luminosity. When UBV photometry is supplemented by Hβ photometry, the colour indices B–V and U–B (corrected for reddening) show spectral type, while the Hβ strength shows the luminosity. Despite the narrow band required for the Hβ measure, the efficiency of the photometric method, in terms of the time needed to reach a given faint star, is higher than that of the traditional spectroscopic method. Stars of other spectral types may be studied in an analogous way, making use of other appropriate spectral-absorption features.

In a different approach to obtaining two-dimensional classification of stars through photometry, several investigators have used four or more bands of intermediate width that, being less wasteful of light, allow the study of fainter stars. A number of general characteristics of stellar spectral energy distribution may be selected as criteria: a colour index based on the region at shorter wavelengths than the Balmer discontinuity (see above *The UBV system*), the depth of the Balmer discontinuity, a colour index in the region between 3800 and 4600 Å that is generally characterized by a different colour temperature than that at longer wavelengths than 4600 Å, or various combinations of these. In some instances it is possible to form compound colour indices that are insensitive to interstellar reddening but that remain sensitive to some inherent property of the star. In general, it has not been possible to accomplish two-dimensional classification over the entire range of spectral types; instead, particular systems are devised for use within restricted spectral ranges.

Spectrophotometry. In spectrophotometry, photometric methods are employed to study in great detail the spectral energy distribution of stars, generally with emphasis on the continuum rather than the spectral lines. Spectrophotometry has been done both photographically, by the scanning of a spectrogram and its calibration images with a microphotometer, and photoelectrically, by the direct scanning of a spectrum at the telescope. The scanning can be continuous or may consist of discrete measures made at selected places along a spectrum. If at selected places, the spectral regions may be isolated by filters; it is more customary, however, to use a spectrograph as a dispersing instrument with one or more slits located in the focal plane that admit restricted and well-defined bands of wavelengths into the light receivers. The bandwidth is generally so restricted that the measures are essentially monochromatic.

As in broad-band photometry, the photographic method is less wasteful of light, since the entire spectrum is photographed; photoelectric direct scanning wastes most of the light but is capable of very high precision. Spectrophotometry, at least with high resolution, is applicable only to the brightest stars and requires large telescopes for maximum light collection. Fortunately, among the bright stars there are examples of most types of normal stars, so that a large body of fundamental knowledge on stellar radiation can be obtained.

An important application of spectrophotometry is the study of interstellar absorption, particularly how the absorption varies with wavelength. For this purpose, two stars of the same spectral type, one reddened, the other unreddened, are compared spectrophotometrically. Another important result of spectrophotometry is the absolute calibration of stellar energy output through direct comparison with artificial standards of radiation. From such measurements, it is known, for example, that the radiation received at the Earth from an A0 star of 0 magnitude at 5560 Å (close to the effective wavelength of the V magnitude) is 3.8×10^{-9} erg per second per square centimetre per angstrom.

Surface photometry. Many celestial objects extend over an appreciable angle and present an apparent surface in the field of view. Examples are galaxies, nebulae, the Sun, the Moon, and most planets. Photometric studies of such objects can be made in two ways: in detail over the surface, as a function of position, or in whole, whereby the integrated light of the entire surface is taken.

The results of detailed surface photometry of an object are conveniently displayed in an isophotal map, a map in which contours of constant brightness are drawn much in the same fashion as altitudes in a terrestrial contour map. Customarily, isophotometric mapping is done by an instrument that follows contours of constant density on a photograph. This technique of isophotometry can be applied in the study of galaxies. If photographs have been made at different effective wavelengths, it is possible to distinguish regions containing the older and redder objects from those regions containing young stars and stellar associations.

Isophotometry of galaxies

A more recent technique consists of the superposition of a positive photograph taken in one colour onto a negative taken in another colour. By appropriate control of the balance of photographic densities, the images of selected types of stars will cancel, leaving the regions containing extreme blue and red objects showing up boldly as dark or light areas. This is an example of the use of photometric techniques in a qualitative rather than quantitative fashion.

While furnishing less information than detailed surface photometry, integrated magnitudes based on the entire surface of an extended object provide a convenient measure of the total radiation emitted. The ratio of the total luminosity to the total mass of a galaxy is an important parameter in the study of these objects. Moreover, for the most remote galaxies, integrated magnitudes provide a means of estimating their approximate distances.

In the study of emission nebulae, measures of the strength of various emission lines provide information about the constitution, density, and temperature of the nebulae. Recently developed techniques employing pressure-modulated Fabry–Pérot interferometers provide, in addition, details about the mass motions of the nebular gases.

Variation of light from astronomical objects. A considerable number of celestial objects vary in brightness, some being highly regular and repetitive in behaviour, others quite irregular. Many of these objects have provided important clues to astronomical problems.

Among the variable stars are some, such as the RR Lyrae variables and the Cepheids, whose light variation is caused by physical pulsation. Since early in the 20th century the Cepheids have played a vital role as astronomical distance indicators. Because those stars possess easily recognized light-variation characteristics in terms of the shapes, amplitudes, and periods of their light curves, they have been studied intensively. In particular, those Cepheid variables located in the neighbouring galaxies called the Magellanic Clouds show that their average brightnesses are highly correlated with their periods. Since all stars in the Magellanic Clouds are roughly at the same distance, the correlation must apply to their average absolute magnitudes also. Once their mean absolute magnitudes are known, it is sufficient to observe the period of any other Cepheid variable to infer its mean absolute magnitude and from this (and its mean apparent magnitude) its distance. This process, stated here in an oversimplified form, is used to determine the distances of the nearest galaxies. Although the RR Lyrae variables are not as intrinsically luminous and cannot be seen as individual stars at such distances, they serve a similar purpose in aiding the astronomer to deduce the distances of some star clusters within the Galaxy.

Pulsating stars, in some instances, also provide knowledge of stellar diameters. Their brightnesses are known to vary as a consequence of a pulsation process in which both diameter and surface temperature vary cyclically. If at two times during the cycle the colour indices and, thus, surface temperatures are the same, then the ratio of brightnesses must be proportional to the square of the radii at these times. If radial velocities are also known, the varying rate at which the stellar surface moves throughout the cycle is known; and, when the rate is integrated, the distance traversed between any two instants can be deduced. Thus, the difference between, as well as the ratio of, the two radii is available, and thus the radii can be determined. By carrying out such a process for many pairs of instants of common colour index, detailed knowledge of the star's pulsation in diameter can be obtained.

Eclipsing binary stars

An important group of stars is the eclipsing binaries, double stars too remote and too close together to be resolved, and whose orbital plane is oriented in space to pass through or very near the Earth. The combined light of an eclipsing binary, viewed of necessity in a single image, diminishes whenever one star eclipses the other; thus, there are two times per period when the apparent brightness reaches a minimum. From an analysis of the geometrical shape of the light curve, it is possible to infer important properties of the stars and their orbits. If radial-velocity curves are also available for the two components, it is possible to determine the stars' masses and diameters.

In the solar system, the rotational characteristics of many minor planets have become known from cyclic variation in their brightnesses, as has the rotational period of the planet Pluto. Observations of the occultation of stars by the Moon have provided important new information on the diameters of stars. Very fast photometry is required to follow the rapid extinction of light as the Moon's edge cuts the star from view; the shape of the extinction curve requires resolution in time on the order of milliseconds.

Two new types of objects discovered in recent years show characteristic variations in energy output, sometimes including the optical spectral region. They are pulsars and quasi-stellar sources, which are treated in detail in separate articles. In one pulsar, near the centre of the Crab Nebula, the optical brightness varies with the same periodicity (33 milliseconds) as its radio emission. Irregular brightness variations occurring in quasi-stellar sources over times of only months imply that their sizes must be no larger than the distance traversed by light in such times, and thus they must be many times smaller than galaxies.

Limits of observations. There are practical limitations to both the accuracy and dynamic range of astronomical photometry. The factors determining the limits are to be found in the statistical character of the detection of small numbers of photons, the irreducible units of light; in the nature of the photometric instrumentation; and in the behaviour of the terrestrial atmosphere.

The spectral transmission of the atmosphere is characterized by numerous wavelength regions called windows (regions of relative clarity) separated by regions of high opacity. Even within the windows, the atmosphere not only absorbs light and disturbs the images formed by telescopes; it also emits light, so that even a moonless sky is not totally dark. In addition, there is a background of scattered light from interplanetary dust and distant unresolved stars, nebulae, and galaxies. On a dark, clear night, the surface brightness of the sky is equivalent to about one star of 22nd visual magnitude in every square second of arc. Ultimately, this background light limits the extent to which faint stars can be detected on a fully exposed photograph.

23rd-magnitude stars

The faintest stars that have been photographed thus far by traditional methods are at about the 23rd magnitude, approximately one-third as bright as the sky background. This limit, 23rd magnitude, also represents the extent to which photoelectric photometry has been carried, because, in the absence of photographic star charts, the positions of any fainter stars remain unknown. If photoelectric photometry of a 23rd-magnitude star is carried out with a 5″-of-arc diaphragm, the light from the star would represent only about 2 percent of all the sky light admitted. At this faint level, only about two photons of visual light arrive per hour per square centimetre at the Earth from the star. Thus, quite apart from the brighter sky background, a very large telescope aperture and a very long observing time are necessary to collect a statistically significant quantity of light. Similar statistical considerations may limit the speed with which even bright objects may be observed, as in the case of rapid-scanning devices or observations of occultations of stars.

At the faintest levels, it appears that electronographic techniques, as developed by André Lallemand and Gerald E. Kron, will make possible the measurement of fainter stars beyond the reach of either conventional photoelectric or photographic methods. Kron has extended photometry in this fashion almost to the 24th magnitude. The essential principle employed is the replacement of the photographic plate by a photoelectric imaging device. In the latter, an image is produced on a photographic plate by highly accelerated electrons emitted by a photocathode placed at the optical focal plane of the telescope. The advantage over conventional photography lies in the much higher efficiency and sensitivity, while the advantage over conventional photoelectric photometry lies in the fact that a feeble star does not have to compete for detection with all the light from the surrounding sky background as admitted by a diaphragm, but only with that sky light arising from the small area occupied by the image itself.

There are also limitations on the working range of photometric processes. A photoelectric photometer, for example, typically is restricted to a range of less than 15 magnitudes because of factors such as saturation and fatigue of instrumental components at the brightest levels and electrical or statistical limitations at the faintest levels. Much wider ranges are possible by means of calibrated optical attenuation and overlapping intervals of measurement. The full range of photoelectrically measured astronomical objects is now about 50 magnitudes, extending from the Sun at an apparent magnitude of −26.8 to the faintest star at about 23rd magnitude.

In terms of accuracy, photometric measures in principle can be carried out over sufficiently long time intervals or can be repeated indefinitely to reduce the residual uncertainty inherent in the averaged results. Even this procedure, however, reaches practical limits imposed by systematic errors caused by imperfect calibration or observational procedures, the approximate character of transformations to standard systems, residual errors in the known magnitudes of standard stars, and unrecognized sources of error. In practice, the best photoelectric magnitudes of bright stars are typically accurate within about

1 percent, while, for certain forms of differential photometry, accuracies about 50 times greater can be achieved. On the other hand, photographically interpolated magnitudes are seldom more accurate than about 4 percent.

BIBLIOGRAPHY

Historical: H.F. WEAVER, "The Development of Astronomical Photometry," *Pop. Astr.*, vol. 54 (1946); G. EBERHARD, H. KOHLSCHUETTER, and H. LUDENDORFF (eds.), *Handbuch der Astrophysik*, vol. 2 (1929–31); W. WIEN and F. HARMS (eds.), *Handbuch der Experimentalphysik*, vol. 26 (1937); JOEL STEBBINS, "The Electrical Photometry of Stars and Nebulae," *Mon. Not. R. Astr. Soc.*, 110:416–428 (1950); *Publs. Washburn Obs.*, vol. 15, ch. 1 (1928).

Photographic photometry: H.F. WEAVER, "Photographic Photometry," in S. FLUGGE (ed.), *Handbuch der Physik*, vol. 54 (1962), in English; J. STOCK and A.D. WILLIAMS, "Photographic Photometry," with a section by B.M. MIDDLEHURST, in W.A. HILTNER (ed.), *Astronomical Techniques*, ch. 17 (1962); C.E.K. MEES, *The Theory of the Photographic Process*, 3rd ed. (1966).

Photoelectric photometry: A.E. WHITFORD, "Photoelectric Techniques," in S. FLUGGE (ed.), *Handbuch der Physik*, vol. 54 (1962), in English; F.B. WOOD, *Photoelectric Photometry for Amateurs* (1963).

Spectrophotometry: A.D. CODE, "Stellar Energy Distribution," in J.L. GREENSTEIN (ed.), *Stellar Atmospheres*, ch. 2 (1960); J.B. OKE, "Absolute Spectral Energy Distributions in Stars," *A. Rev. Astr. Astrophys.*, 3:23–46 (1965).

General: G.P. KUIPER and B.M. MIDDLEHURST (eds.), *Stars and Stellar Systems*, vol. 2 and 3 (1962–63), contain many articles dealing directly or indirectly with astronomical photometry. The UBV system was first described by H.L. JOHNSON and W.W. MORGAN in "Fundamental Stellar Photometry for Standards of Spectral Type on the Revised System of the Yerkes *Atlas*," *Astrophys. J.*, 117:313–352 (1953).

(R.H.H.)

Photoperiodism

Photoperiodism is the functional or behavioral response of an organism to changes of duration in daily, seasonal, or yearly periods of light and darkness. Flowering in plants, breeding in animals, and migration in birds are among the many well-known photoperiodic responses that have received intensive study. Though such reactions are genetically determined, temperature, nutrition, and other environmental factors modify the responses.

Modifying factors

Plants. Photoperiodism was first reported in 1920 when the flowering of certain plants was demonstrated to be controlled by the daily duration of light. Investigators established the existence of three types of plants on the basis of day lengths favourable to flowering: short-day plants, which fail to flower when days exceed a certain duration called the critical day length; long-day ones, which fail to flower unless the days exceed a critical day length; and indeterminate ones, which flower about equally well regardless of day length. Critical day lengths for both long- and short-day plants differ widely from species to species and even among varieties of a single species.

Other responses that are influenced by the photoperiod include the formation of bulbs, tubers, roots, and runners; the coloration of stems and leaves; the falling of leaves; and the development of certain anatomical features.

The photoperiodic response of plants actually depends on the duration of the darkness, and not on light, a fact that was not understood at the time the phenomenon was named. Thus, a short-day plant flowers when nights are long and a long-day one when nights are short or absent. A brief period of artificial light in the middle of a long dark period divides it into two short ones and thus promotes the flowering of long-day plants and inhibits that of short-day ones. The dark reactions are, therefore, the time-measuring reactions.

Control of flowering

Growers of chrysanthemums can produce these highly seasonal flowers throughout the year by proper control of night length and other cultural conditions for adapted varieties. Increasing the length of day by artificial light keeps the plants from blooming until they attain sufficient size or until the desired date is reached. The light is then discontinued, or in some times of the year the plants are covered with black cloth during part of the day, thereby increasing the night length adequately to bring them into bloom.

Artificial control of day length is also used to regulate the time of flowering of orchids, China asters, feverfew, and tuberous-rooted begonias. Experiments show that the growth of some kinds of woody nursery plants can be maintained in the greenhouse during the winter by the use of artificial light. The uses of artificial light to improve production of various kinds of plants are constantly increasing.

The photoperiodic reaction occurs in the leaf, but the response generally occurs elsewhere. Thus, a product generated in the leaves evidently moves to the growing points, where it controls flower-bud formation. The product is a flower-promoting substance produced during dark periods. It may be a special flower-forming hormone, florigen, but such a substance had not been identified by the mid-1970s. Other growth substances, especially a group called gibberellins, are known to play a key role in flower formation.

Knowledge of the initial photochemical step in photoperiodism was obtained by studies of the way in which light acts. Red light is more effective than any other kind in preventing flowering of short-day plants and promoting flowering of long-day ones. A second but far less effective region occurs in the blue area of the light spectrum. Flowering is controlled in both kinds of plants by light of the same wavelength ranges, indicating that the basic photochemical reactions are identical, although in one type they suppress, and in the other promote, flowering.

The inhibition of flowering of some short-day plants by exposure to red light can be prevented by exposing them immediately afterward to light of somewhat longer wavelength, namely, far-red light. A similar cancellation of the action of red by far-red radiation occurs in the germination of light-sensitive seeds and in the control of growth of the stem and other parts of dark-grown seedlings, indicating that the same initial photoreaction is present in all these phenomena. Many of these responses can be reversed repeatedly by successive alternating exposures to red and far-red light. The ultimate response to such a series of exposures is dependent on the kind of light last given.

Reversibility arises from the photochemical interconversion of two forms of a photoreceptive pigment called phytochrome. Thus, a red-absorbing form is converted by red light to the far-red-absorbing form and this in turn is reconverted by far-red light to the red-absorbing form. The far-red-absorbing form of phytochrome is believed to be biologically active, probably an enzyme; but the action it catalyzes is not yet known. The red-absorbing form is thought to be inactive.

Animals. The most conspicuous activities of animals closely correlated with certain seasons of the year, and hence with changes in day length, are bird migration, reproduction, and changes in coat and plumage. Each of these occurs with marked regularity at a particular time each year. By retaining animals in captivity and subjecting them to artificial increases and decreases in day length, bird migration, reproduction, and other activities have been induced out of season. It has been concluded, therefore, that in nature the changing daily periods of light and darkness determine to a great extent when many seasonal activities occur.

Before their northward migration to breeding grounds in the spring, birds undergo a marked change in their physiological state. The change is shown by increased activity of the reproductive organs and deposition of large amounts of fat. When these same changes were induced experimentally in winter by exposing captive birds to artificial increases in day length, the birds migrated northward in winter when released. By employing different schedules of day length at various seasons of the year, it has been demonstrated that the rate of change in physiological state depends on the daily amount of illumination. As little as nine hours of light per day is stimulating. The response is rapid under long days and slow

under short days. A gradual increase in the amount of light per day is not necessary. Constant day lengths are effective, as are decreasing day lengths, provided that the daily amount of light is stimulating. Therefore, migratory birds that winter on or near the Equator, where the day lengths are constant, or in the Southern Hemisphere, where the day lengths decrease after December 21, could have the onset of their spring migration regulated by day length. Prerequisite to the response to light that occurs in winter and spring is exposure to short days or long nights in autumn. This phase of the annual cycle is called the preparatory phase, since failure to complete it precludes the subsequent response. Under experimental conditions completion occurs uniformly and more rapidly with longer nights. Day lengths in autumn wind the "clock" within the bird, so to speak, while the day lengths of winter and spring make it run and determine how fast it runs.

Regulation of reproduction

In most birds, mammals, and other vertebrates, breeding is seasonal. By experimental manipulation of day length, reproductive activity has been induced out of season. For the species that normally breed beginning in autumn (*e.g.*, brook trout, deer, sheep, and goats), decreasing daylight or short days following long days will be effective. For species that breed in the spring (*e.g.*, starling, junco, ferret, and mink), increasing daylight or long days following short days will be effective. Changes in breeding season occur in some birds and mammals when they are transported from the Southern to the Northern Hemisphere or from the region of the Equator to temperate latitudes. Although day lengths are constant on the Equator and change but little in the tropics, there is some evidence that day length does regulate the breeding and plumage cycles of tropical birds. Experimental modification of the breeding cycle of invertebrate animals (*e.g.*, snails, crustaceans, and insects) by means of day length has also been demonstrated.

Attempts to discover how the length of daylight affects the breeding cycle have shown that the light stimulates the eye or part of the brain near the eye. The light stimulates the release of special hormones from the brain, which eventually reach the pituitary gland (the master endocrine gland located at the base of the brain). The pituitary, in turn, secretes the hormones that, among other functions, control the growth of the reproductive organs. There is good evidence that the length of the dark period and the ratio of light to darkness are important factors in determining the reproductive response. The effects of day length on the migratory response seem to apply as well to the reproductive response in some birds. Response to light can be modified by temperature, nutrition, and inherent internal factors. Important among the last is a daily rhythm of about 24 hours' duration (termed circadian), which makes it appear as though the organism were keeping track of each day. The response to light-dark cycles seems to depend in part on the time that the light and dark periods occur within these 24-hour periods. A few species of birds and mammals studied did not show any influence of light on the reproductive cycle.

The most conspicuous seasonal change in coat and plumage occurs in those mammals and birds that are white in autumn and winter and brown in spring and summer (*e.g.*, varying hare, ermine, and ptarmigan). By experimental manipulation of light, either of these colours can be produced out of season. Molt into white plumage or coat can be induced by short days. Temperature is not a factor. Day length also influences the time of annual molt in birds, growth of wool in sheep, and production of the winter coat in mink and ferret.

A number of less obvious seasonal activities also occur in response to day lengths. In some aphids, the production of winged individuals (as opposed to wingless individuals) normally occurs in autumn and is dependent on a period of darkness of 12 to 14 hours. In certain other insects the occurrence of sexually reproducing individuals, which occurs normally in autumn, can be produced out of season by shortening the days.

Practical applications of knowledge of photoperiodism are common in poultry management, for the length of daylight affects laying, sperm production, and body weight of the fowl.

BIBLIOGRAPHY. STANLEY D. BECK, *Animal Photoperiodism* (1963), a concise general presentation of the subject on mammals, man, birds, and insects; *Insect Photoperiodism* (1968), an advanced book on all aspects of photoperiodism in insects, with an extensive bibliography of the original research; BRIAN LOFTS, *Animal Photoperiodism* (1970), a popular work written for the general audience; ERWIN BUNNING, *The Physiological Clock*, rev. 2nd ed. (1967), a short monograph especially concerned with the biological clock used to measure the length of daily light and dark periods and involved in the photoperiodic responses of plants and animals; JURGEN ASCHOFF (ed.), *Circadian Clocks* (1965), proceedings of a summer school session held in September 1964 on the subject of circadian (approximately 24-hour) rhythms; ROBERT B. WITHROW (ed.), *Photoperiodism and Related Phenomena in Plants and Animals* (1959), proceedings of the first international conference on photoperiodism held in Gatlinburg, Tennessee, in 1957—an excellent classic source for reviews of original theories and observations.

The following are advanced review articles of current research; almost all have extensive bibliographies and attempt to bring each subject area up to date as of time of publication. ALBERT WOLFSON, "Animal Photoperiodism," in ARTHUR C. GIESE (ed.), *Photophysiology*, vol. 2, pp. 1–49 (1964); "Environmental and Neuroendocrine Regulation of Annual Gonadal Cycles and Migratory Behavior in Birds," *Recent Prog. Horm. Res.*, 22:177–244 (1966); STERLING B. HENDRICKS, "Photochemical Aspects of Plant Photoperiodicity," in *Photophysiology*, vol. 1, pp. 305–331 (1964); DONALD S. FARNER, "Comparative Physiology: Photoperiodicity," *A. Rev. Physiol.* 23:71–96 (1961); "The Photoperiodic Control of Reproductive Cycles in Birds," *Am. Scient.*, 52:137–156 (1964); H.A. BORTHWICK, S.B. HENDRICKS, and M.W. PARKER, "Photoperiodism" in ALEXANDER HOLLAENDER (ed.), *Radiation Biology*, vol. 3 pp. 479–517 (1956); MIKHAIL K. CHAILAKHYAN, "Internal Factors of Plant Flowering," *A. Rev. Pl. Physiol.*, 19:1–36 (1968); L.T. EVANS, "Flower Induction and the Florigen Concept," *ibid.*, 22:365–394 (1971).

(A.Wo.)

Photoreception

Photoreception is the activation of a biological process by means of illumination. Most organisms, including man, respond to visible light; some react to wavelengths of light not seen by man; and still others can react to properties of light not detectable by man, such as polarization (vibration of light waves in a definite pattern). This article is concerned with the sensory processes by which animals detect information carried by light.

Light energy is necessary for life on Earth. Green plants require light for photosynthesis, the process by which water and carbon dioxide are transformed into carbohydrates; plants also show adaptive responses (*e.g.*, germination and flowering) to annual changes in daily light periods. Animals depend on plants for food and thus are indirectly dependent upon photosynthesis. In some animals, response to variations in day length is of great importance in the regulation of annual reproductive cycles. (For additional information about the above responses to light, see PHOTOSYNTHESIS; PHOTOPERIODISM; STEREOTYPED RESPONSE.)

Light, the name given to the mediator of the sensation of sight in higher animals, including man, and the equivalent of this sensation in lower animals, is the part of the electromagnetic spectrum that is visible to animals; it includes the range of wavelengths from about 300 nanometres (1 nm = 10^{-6} millimetre) in the near ultraviolet to about 700 nanometres in the deep red (300 nanometres is beyond violet and does not evoke sensation in the human eye).

The entire cell of a unicellular animal such as *Amoeba* may be sensitive to light so that the cell moves toward or away from it. Some unicellular animals (*e.g.*, *Euglena*) have developed a light-sensitive receptor, or eyespot—a region with a lower threshold for light stimulation than occurs in the rest of the cell. Some multicellular animals have photoreceptive cells, or eyespots, scattered in various parts of or throughout the body; those in the outer covering of the earthworm (*Lumbricus*) serve in directional orientation, which involves comparison of light intensities at different directions. Most animals

have localized photoreceptors of varying complexity—*e.g.*, the ocellus of certain mollusks and arthropods; the compound eyes of arthropods; and the camera eyes of cephalopods and vertebrates.

Evidence indicates that the eyes of certain insects can make use of the information carried by near ultraviolet wavelengths of light as well as that carried by visible wavelengths; both carry information related to the sensation of colour. The eyes of many invertebrates, such as certain arthropods and mollusks, have evolved in such a way that they can detect polarized light; *i.e.*, it evokes a sensation and provides information used for navigation. Visual sensation in higher organisms is primarily a complex response to the intensity and the spatial and temporal distribution of light on the photosensitive retina (the innermost layer of nervous tissue within the eye). Different eyes, different specialized parts of the same eye, and even the same parts of the same eye vary in their responses to illumination. Both the properties of light and those of the eye are thus important determinants of visual sensation. The great differences in the light-analyzing capacities of animals are reflected in the great diversity of gross structural organization involved in photoreception. The fundamental mechanism of photoreception—photochemical activation of a light-receptive pigment and the primary excitation-initiating process—seems to be similar among most animals, however.

This article deals with the optical properties of eyes, including the basic arrangements for image formation and light detection and the morphology of photoreceptors; the photochemistry of light detection; and the physiological functioning of the receptor cells that initiate nervous activity. These initial processes of photoreception provide information for the neural centres of the retina and higher nervous centres. The neural events involving the higher centres lead to visual perception.

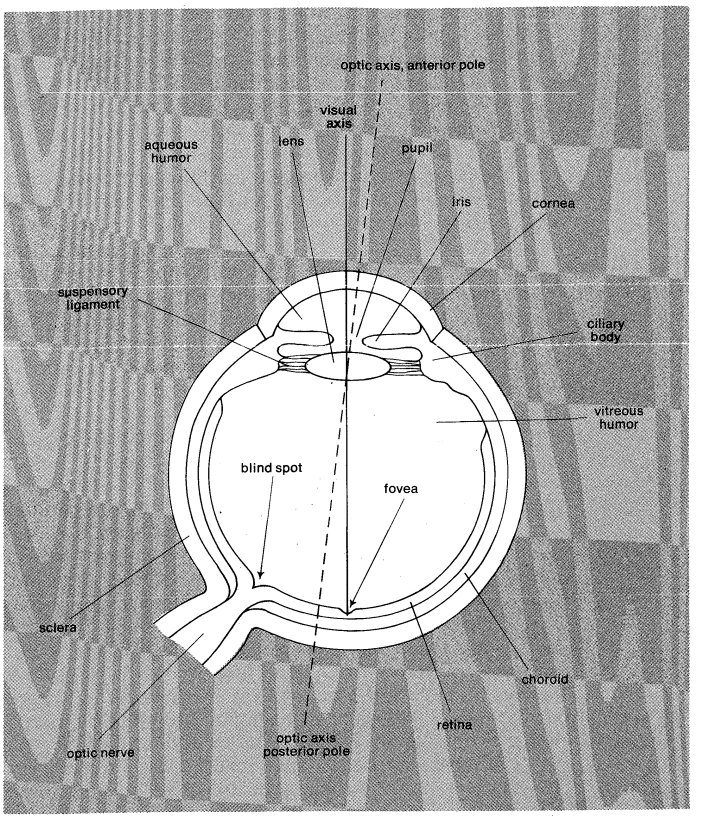


Figure 1: General structure of the mammalian eye.
From P.B. Weisz, *The Science of Zoology*, copyright 1966; used with permission of McGraw-Hill Book Co.

THE OPTICAL PROPERTIES OF EYES

The first active step in vision is the absorption of light by a photosensitive substance, a visual pigment. Various devices within the eye assist vision by directing incoming light to this pigment; *i.e.*, they act as light guides by refracting (bending), reflecting (turning back), or guiding light. The arrangement of the optical structures influences the resolving capability and other basic properties of visual sensation.

Camera eyes. *In vertebrates.* The mammalian eye (Figure 1), somewhat like a camera, has a cornea (a transparent, anterior portion) and a lens; it functions as a dioptric system—*i.e.*, a system in which light rays are refracted so as to focus on the retina; the image projected on the retina is inverted. In the retina the light first passes through several layers of nerve cells before impinging on the photoreceptor cells, called rods and cones. The dimensions and refractive powers of all the optical parts of the human eye are known, making it one of the best understood vertebrate eyes (see EYE AND VISION, HUMAN).

The large size of the camera-like vertebrate eye makes it potentially the most efficient of all eyes because it can project a large image on a large surface area containing a high density of receptors. Both vertebrate and invertebrate eyes reflect the influence of the animal's environment. The optical arrangement of the eyes of animals active during the night (*i.e.*, nocturnal) suggests that resolution is sacrificed for light-gathering power (see Figure 2). The opossum lens, for example, is so large that it almost touches the retina—*i.e.*, it has a short focal length, the distance from the centre of the dioptric system to the place at which the image of distant objects is focussed (retina). The short focal length combined with a wide aperture results in a low focal ratio, or f-number

The influence of environment on vertebrate eyes

$$\left(\frac{\text{focal length}}{\text{aperture diameter}}\right),$$

and ensures high light-gathering ability. In the eyes of animals active during the day (diurnal), the lens is smaller; as a result, the optical centre is closer to the front of the eye, and its front surface is flatter. Thus, the focal length of the system is longer, the f-number is higher, and the image on the retina is larger and dimmer than in the nocturnal eye. Assuming that the large image can be detected by the photoreceptors, resolution is improved at the expense of the speed of the lens system.

From G.L. Walls, *The Vertebrate Eye*, Cranbrook Institute of Science

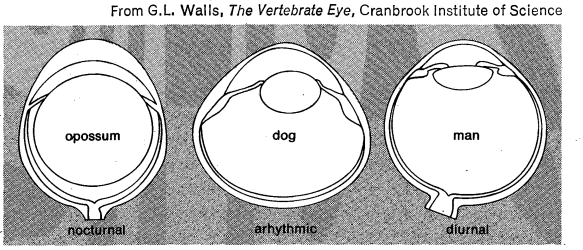


Figure 2: Influence of the environment on the optical arrangements of animal eyes (see text).

In order to utilize efficiently a large image, the retinas of diurnal animals have localized areas with many photoreceptors; *i.e.*, a higher receptor density. The receptors in this area, called the area centralis, are usually cones, the receptors of daylight and colour vision. These areas for sharp vision, often circular, are seldom located exactly in the optic axis (an imaginary line drawn through the centre of the cornea and the central point of the eye, see figure 1). The eyes of most birds have two such areas, the centre of each of which is specialized by a thinning of the retina to include only the receptors. This gives rise to a depression called the fovea, also found in teleost fishes, certain reptiles, and man. When the area centralis contains a yellow pigment, it is called the macula lutea. The macula lutea is found in higher primates (simians) and possibly chameleon lizards. This pigment filters out the shorter wavelengths of light and improves the sharpness of the image by reducing the chromatic aberration (variation of the focal length with different wavelengths of light) that would be caused by the inability of the lens to bring the long and short wavelengths to the same focus.

Distant objects are in focus on the retina of the normal human eye. In order for objects closer than about six

Accommodation in the eyes of vertebrates

metres (20 feet) to be in focus, however, an adjustment called accommodation is necessary; otherwise, the image would fall in back of the retina, and the object would appear fuzzy. In mammals, birds, and reptiles other than snakes, the accommodative adjustment consists of sharpening the curvature of the lens so as to shorten its focal length. In snakes, elasmobranchs (*e.g.*, sharks), and amphibians, accommodation is achieved by moving the lens —hence its focal plane—forward. In lampreys and teleost fishes the eye is adjusted for near objects, and accommodation for distant vision is carried out by a backward movement of the lens. Some species have evolved anatomical adaptations that make accommodation unnecessary. The retina of the fruit bat (*Pteropus medius*), is in folds, ensuring that some part of it will intercept an image at any location. The ray *Raja batis* and the horse have so-called ramp retinas, in which a continuous and gradual change occurs in the distance between the lens and retina in certain parts of the eye. Specific areas in these animals' eyes are presumably used to view objects at varying distances much as the human eye directs the image for detailed vision onto the fovea.

Because vertebrate species are adapted to almost every aquatic and terrestrial environment, they have evolved equally diverse eyes. In air, for instance, the front surface of the cornea can function effectively for image formation; in underwater eyes, however, the refractive index (the ratio of the speed of light in air to that in a given medium) of water and the cornea are almost identical, and the corneal front surface does not refract light. In these eyes the lens does much of the image formation.

Vertebrates have two types of photosensitive cells, rod cells and cone cells. The rod cells, which are long and fat, contain large amounts of visual pigment; they are the photosensitive cells for vision under conditions of dim illumination (scotopic vision). The cone cells, which are relatively small, mediate daylight vision (photopic vision) and colour sensation in many animals. The photosensitive photoreceptor outer segments of rods and cones are stacks of disks, or lamellae, with the planes of the disks at right angles to the long axis of the rod and cone cells. The retinas of animals active both day and night, as are those of humans, contain both rods (for night vision) and cones. In parts of the human retina the rods and cones are intermingled; elements of the nervous system provide the switching mechanism that permits adjustment for light conditions. The specialized fovea contains only cone cells; in the fovea the switching function is accomplished by eye muscles that change the direction of the field of vision in order to bring the image to the fovea.

The amount of light reaching the photoreceptor cells is controlled to some extent by the pupil, the opening of the eye through which light passes. The iris, the coloured portion of the eye surrounding the pupil, constitutes a diaphragm. Its muscles cause the pupil to change in diameter, decreasing the size of the pupil when light enters and increasing it when little or no light enters. The area of the pupil increases about 15 times in going from one millimetre to four millimetres in diameter, a relatively small increase in comparison with the range of light intensities under which the eye effectively operates. Since the amount of light entering the eye is proportional to the size of the pupil, it can be seen that changes in pupil size modify the amount of light only over a small range. Changes in pupil size are important in the human eye because they allow the lens to be used most effectively for visual acuity. When the whole lens is used, as in dim light when the pupil is large, the image formed by the lens is rather poor, chiefly because of chromatic aberration. The neural image on the retina is already poor, however, because the responses of thousands of rods must be pooled to obtain maximum sensitivity. Use of the whole lens is beneficial because it adds further light without reducing the image quality. When illumination is bright, the pupil is small, and only the aberration-free central part of the lens is used. This high-quality image is used effectively by the cone receptors of the fovea. There, no pooling of receptor responses occurs.

Pupils that form a circle when closed cannot greatly change in area; however, a pupil that forms a slit when closed can close almost completely. When nocturnally active animals find themselves in bright sunlight, they need additional protection for their sensitive rod-containing retinas; such animals have evolved pupils that close to form a slit. Many nocturnal vertebrates also show eyeshine (*e.g.*, the glow of a cat's eyes reflecting light at night). Eyeshine, which is caused by a mirror-like reflection from either the retina or choroid (a layer of blood vessels and connective tissue), enhances the sensitivity of the eye. The reflection of the light outward means that it passes the receptors a second time, giving them a chance to absorb light that was not absorbed during the inward passage through the receptors. Some animals thus have smaller rod receptors than they would otherwise need.

Adaptations of the eyes of nocturnally active animals

In cephalopods. The eyes of the invertebrate cephalopods—octopus, squid, and cuttlefish—are usually cited as examples of convergent evolution because they have independently evolved large camera-like eyes similar to those of vertebrates. The cephalopod eye lies within a cartilaginous cup. It consists of a cornea, lens, iris, and retina with the same basic relations to one another as are found in vertebrate eyes. Iris muscles can enlarge and narrow the pupil. Many details, of course, are different; for example, although the maximum density of photoreceptors in cephalopods is high—about 50,000 per square millimetre in *Loligo* and 100,000 per square millimetre in *Sepia*—the retinal structure otherwise bears little resemblance to that of the vertebrate. The photosensitive cells are of two types and are organized to detect polarized light (see below *Morphological features*). In addition, unlike vertebrate receptors, those of the cephalopod are the first element of the retina to be illuminated (rather than the last, as in vertebrates), and the optic ganglion (a mass of nerve tissue) is separate from the retina (rather than an intimate component, as in vertebrates). Both vertebrate and cephalopod retinas show pigment migration and movement of the receptors in response to adaptation for conditions of light and dark.

The cephalopod cornea does not have any focussing function. Image formation is accomplished entirely by the lens, which is forced forward for viewing nearby objects. The pupil is round in deep-sea cephalopods, such as *Loligo*, and slit-shaped in shallow-water dwellers, such as the octopus. Cephalopod photoreceptors are very long, an adaptation for nocturnal or deep-sea and low-light level conditions. The length allows for the presence of more visual pigment and hence greater absorption of light by each receptor cell.

Eyespots. Eyespots, the most primitive eyes, are found in the protozoan flagellates (unicellular animals with a flagellum, or whiplike structure, used for locomotion), flatworms (Platyhelminthes), and segmented worms (Annelida). An eyespot may be a specialized part of a cell as in protozoans, a single photoreceptor cell, or a small cluster of receptors with few or no accessory optical and neural structures. The entire epithelium (skin) of the annelid earthworm *Lumbricus* contains isolated light-sensitive cells that are eyespots. These photoreceptor cells are rather spherical in shape. A rhabdomere—a structure containing photosensitive pigment—lines a vacuole, or internal cavity, of the photoreceptor cell. The function of the vacuole may be to gather light.

Eyespots in planarians and amphioxus

The flatworm *Planaria* has a more highly developed eyespot. A number of photoreceptor cells are clustered under the epidermis. All of the rhabdomeres, which occur together within a cup-shaped collection of pigment cells, are located on slender filaments some distance from a cell that also gives rise to a nerve fibre.

A third type of eyespot is found in the nerve cord of the cephalochordate amphioxus. Each of a small cluster of photoreceptors, the Hesse cells, has a rhabdomere along the edge that faces a pigmented cell.

Ocelli. The ocellus, which is recognized as a true eye, is similar to a camera in that it usually projects an inverted image onto a light-sensitive layer. The ocellus is distinguished from the compound eye, which has many lenses, and from the more highly developed camera-like eye of the mollusks and vertebrates.

In mollusks. In a simple ocellus, that of *Nautilus*, the photoreceptor cells are bipolar—*i.e.*, the rhabdomere is at one end, the nerve fibre at the other—and arranged in a cup-shaped sheet. There is no lens or cornea, only a pinhole opening.

The more complex ocellus of the slug *Agriolimax reticulatus* is located at the tip of the tentacle (see Figure 3); there is a cornea under the epithelium, a vitreous body (a mass of clear jellylike material), and a lens, as well as a main retina and an accessory retina. The accessory retina is believed to function as an infrared receptor. As the tentacle is withdrawn, the accessory retina is rotated so that it is exposed to incoming radiation. The few photoreceptor cells are surrounded by pigment-containing cells. The ovoid eye is about 0.18 millimetre (0.01 inch) in its longest diameter. Distant objects appear in focus in the photoreceptor cells of *Agriolimax*, and the shape of the eye changes as the state of retraction of the tentacle varies. The change in shape may provide a mechanism for accommodation, although accommodation may not be particularly useful, because indications are that this ocellus does not clearly distinguish form.

The molluscan ocelli described above resemble a miniature simple camera, in which an inverted image is projected onto a photosensitive retina. The optics of other ocelli, however, are more sophisticated; for example, the scallop *Pecten* (Figure 4) and related genera have about 100 eyes located along the fringe of the mantle, the lining of the inner surface of the shell. The eyes are one millimetre in diameter and have a double retina; one is called a distal retina, the other a proximal retina. The nerve fibres of the distal (*i.e.*, farthest from the body axis) retina respond only to decreases in light intensity; the nerve fibres of the proximal (*i.e.*, closest to the body axis) retina, on the other hand, respond only to increases in light intensity. Because the photoreceptors in the proximal retina are inverted, light passes through all parts of them before reaching the light-sensitive pigment, and the photoreceptors just touch a reflecting structure, the tapetum lucidum. The scallop tapetum contains about 35 layers of thin crystals; the thickness of and the degree of separation between crystals are precise. The crystals and the intervening spaces function as an interference filter. The small reflection from each crystal interface adds to give a large net reflection from the surface of the tapetum. The fact that the photoreceptors just touch the tapetum ensures that they are illuminated on the second pass (after tapetal reflection) by the same light that passed through them the first time. The tapetum in the scallop eye acts as a concave (*i.e.*, depressed toward the centre) mirror, projecting light through the proximal retina and focussing an image of a distant object on the photoreceptors of the distal retina. The combination of tapetum lucidum and inverted photoreceptors in the proximal retina may thus enhance sensitivity without sacrificing resolution. The structural details of the tapetum lucidum have been described not only for the eye of the scallop *Pecten* but also for several arthropod and vertebrate eyes.

The tapetum lucidum

From P.F. Newell and G.E. Newell, *Invertebrate Receptors* (1968); The Zoological Society of London, No. 23

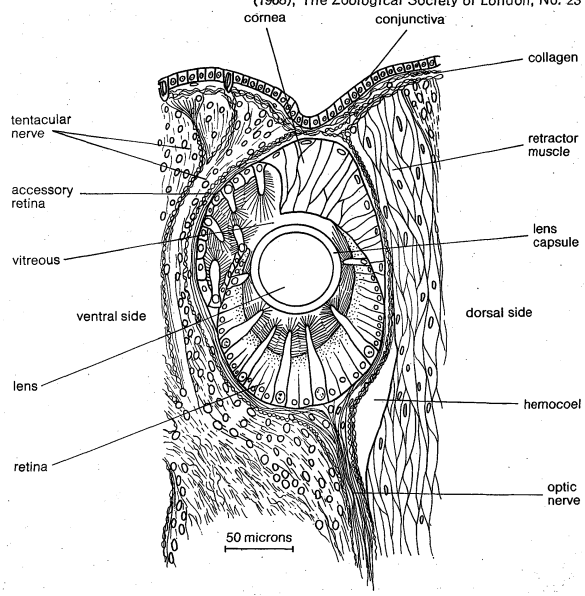


Figure 3: Longitudinal section of the eye of the slug *Agriolimax reticulatus.*

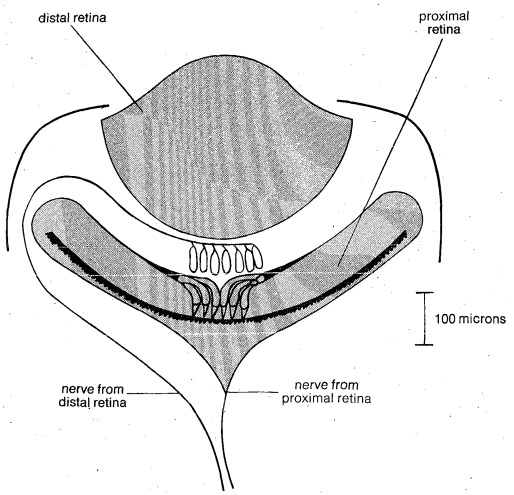


Figure 4: The central region of the eye of the scallop *Pecten.*

From M.F. Land, *Symp. Zool. Soc. London No. 23* (1968); Zoological Society of London

In arthropods. Among arthropods, ocelli are the main organs of sight in arachnids such as spiders and in insect larvae that undergo complete metamorphosis (*i.e.*, a radical physical change during development). Insects that undergo incomplete metamorphosis have three ocelli arranged in a triangle on the dorsal, or top, part of the head; these are subsidiary, however, to the main organs of sight, the compound eyes.

Spiders have two kinds of ocelli, the principal, or antereo-median, eyes and the lateral eyes. The principal eyes of jumping spiders are used when stalking prey and in courtship; the photoreceptors are arranged in four layers in the optic axis (see Figure 5). The two deepest layers cover the entire retina; the two most superficial layers are confined to the retina's central region. The rhabdomeres of the three deepest layers are rod shaped. Those of the most superficial layer are ovoid and oriented in the direction of propagation of light through the eye. Because the receptors are layered, the images of objects at different distances from the eye appear in different layers; it is possible that this type of retinal system is used for depth perception. It appears more likely, however, that the arrangement is used for colour vision. Images of distant red objects appear in the first retinal layer, blue-green images appear in the second layer, and ultraviolet images in the third layer, because the primitive lens system shows considerable chromatic aberration.

No image for visible light appears in the most superficial ovoid rhabdomeres of the principal eyes of spiders, and it has been postulated that they function to detect polarized light. The principal eyes also have a set of six muscles that produce eye movements when an image falls on the retina. The eye follows the image, which remains centrally fixed.

The lateral ocellus of the caterpillar, which has a group of five or six ocelli on each side of the head, resembles one ommatidium, or unit, of a compound eye. The optical apparatus includes a lens, or cornea; a crystalline cone, which is the principal focussing device; and two layers of photoreceptors (distal and proximal) in the optic axis, each of which is radially arranged like sections of an orange. The rhabdomeres of both cell types are in close apposition over much of their length. Objects at different distances project images at different vertical locations in the rhabdom—*i.e.*, a group of rhabdomeres in physical contact with one another. The proximal and distal

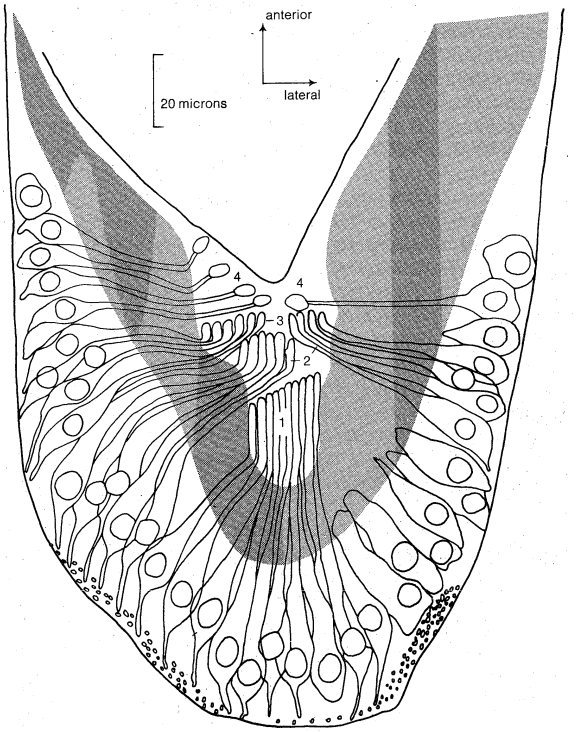


Figure 5: Frontal (horizontal) section close to the centre of the retina of the spider *Metaphiddipus aeneolus*, with the layers of receptor endings numbered 1 to 4. In layers 1, 2, and 3 the presumed receptive part of each receptor is the straight terminal portion; in layer 4 it is the terminal ovoid swelling.
From M.F. Land, *Journal of Experimental Biology* vol. 51 (1969)

rhabdoms form an optically continuous structure, a light guide. The optical information in this structure depends on the light accepted at the distal part of the rhabdom, rather than on the locations in the rhabdom at which images are projected, because the rhabdoms actually touch one another and effectively form one optical structure (see *Optical properties of photoreceptors* below).

Spatial resolution in ocelli and camera eyes

No optical system can form a point image of a distant luminous point, such as a star; because of diffraction (a modulation in which a redistribution of energy occurs), the image is instead a small spot, the Airy disk, with a bright centre enclosed in concentric, alternately bright and dim rings. The radius of the first dark ring indicates the size of the diffraction pattern. The smaller the aperture and the longer the wavelength, the bigger the Airy disk and the poorer the resolution.

The degree to which images can be detected is determined both by the size of the Airy disk and the size, separation, and density of packing of photoreceptors. Compared with the eyes of most invertebrates, a relatively large number of photoreceptors occurs in the principal eye of the jumping spider. This suggests that the image is good, that its quality is primarily determined by the size of the lens, and that the mosaic of receptor units makes maximal use of the image quality. By comparison, the theoretical resolution of the human eye is ten times better than that of the spider. This difference results solely from the difference in the size of the eye. In both the human eye and that of the spider, the focal length is adjusted to transmit the image onto a retina containing about the same density of photoreceptors. Actual resolution, however, is complex and depends on other factors, such as the nature of the target, properties of the illumination, the form of measurement, and neural processing in the eye and brain.

In the lateral ocellus of insect larvae, the size and density of the receptors are probably the limiting factors in resolution. In addition to poor spatial resolution, the light-gathering power of the lateral ocellus is also inferior to that of the spider eye.

Compound eyes. Aggregations of ocelli are found in several lower animals—in polychaete worms, for example. True compound eyes are found only in the arthropods, however. The compound eye of insects is composed of hexagonal or rectangular-shaped, closely packed optical units called ommatidia (small eyes); each ommatidium is virtually a single eye. In different species the size, number, and structure of ommatidia vary. An ommatidium (Figure 6) is composed of a corneal lens, or facet, which consists of a modified extension of the cuticle (the hard outer covering of arthropods) on the surface of the eye; four cells called Semper's cells or cone cells, which form the crystalline cone; and a sensory region called the retinula (small retina). In primitive insects (*e.g.*, the springtail *Lepisma*), in which the transparent cone cells are not specialized, the ommatidia are called acone ommatidia. In the more common eucone ommatidium, which occurs in moths and butterflies, the cone cells have a more complicated structure and contain granules of glycogen, or animal starch; because the granules are packed at various distances from each other, the refractive index varies in different positions in the cell. In certain beetles (*e.g.*, *Lampyris*) and in the horseshoe crab *Limulus*, the crystalline cone is an extension of the cornea.

Components of the ommatidium

The sensory part of the ommatidium, the retinula, consists of several radially arranged cells (retinular cells); each has a photoreceptor component, or rhabdomere (see Figure 6). The rhabdomeres of neighbouring retinular

From *Journal of Cellular and Comparative Physiology* (1965)

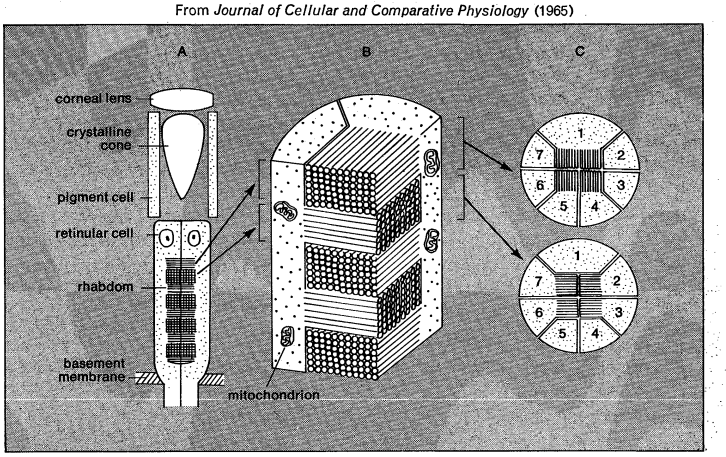


Figure 6: *An ommatidium from the compound eye of the crayfish Procambarus clarkii.*
(A) Longitudinal section through the optic axis with the cone stalk between the crystalline cone and the retinula omitted.
(B) Part of the rhabdom. The direction of the closely parallel microvilli in one layer is perpendicular to that in the next.
(C) Cross sections through the retinula at the levels of two neighbouring layers, rhabdomeres from retinular cells 1, 4, and 5 constituting one layer (upper figure); 2, 3, 6, and 7 constituting the next layer (lower figure); 1, 4, and 5 the third; and so on alternately.

cells may be either in contact (forming a rhabdom) or completely separate. The optical isolation of each ommatidium is enhanced by its being surrounded by light-screening, pigment-containing cells. During adaptation to light and dark conditions, migration of pigment in cells around the crystalline cone, the corneal process, the proximal part of the ommatidium, and within the retinular cells has been reported to occur in most types of compound eyes.

The apposition eye. In the compound eyes of diurnal arthropods, each ommatidium is separated from its neighbours by pigmented, or iris, cells under all conditions of illumination. As a result, the rhabdom of each ommatidium receives light only through its own corneal lens; light from the lenses of other ommatidia is blocked by the pigment. This is the basic structure of the apposition eye (Figure 7). There are, however, variations in structure.

In the honeybee *Apis mellifera*, the cornea is layered, each layer having a different refractive index. It has been shown that the corneal lens projects a small inverted image of a distant object in the crystalline cone some distance short of the rhabdom of an ommatidium. If the image were focussed exactly on the rhabdom, it might effectively transmit image-location information by means

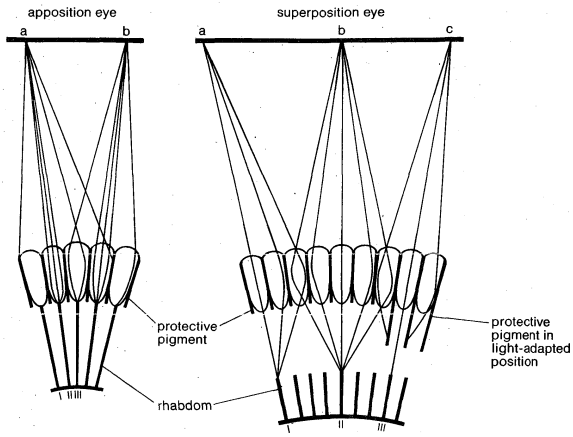


Figure 7: Image formation in apposition and superposition eyes: I, II, III represent different ommatidia; a, b, c represent point objects (see text).
From G.A. Mazohkin-Porshnyakov, *Insect Vision* (copyright 1969); Plenum Press

of differing responses of the individual retinular cells within an ommatidium. Because the image is focussed in front of the rhabdom, however, individual retinular cells within an ommatidium do not respond to image location. The defocussing also has the effect of broadening the field of view of the ommatidium. As the image moves, it can still be guided into the rhabdom by the crystalline cone, which is considerably wider than the rhabdom.

Specialization of the butterfly eye

It is not possible to generalize from the known optics of the compound eye of the bee to other apposition eyes as, for example, those of the butterfly, which superficially resemble the bee's in having a cylindrical rhabdom. Certain features of the butterfly eye are known, however (see Figure 8). Under the rhabdom of each ommatidium, the

From N. Yagi and N. Koyama, *The Compound Eye of Lepidoptera: Approach from Organic Evolution* (1963); Maruzen Co., Ltd.

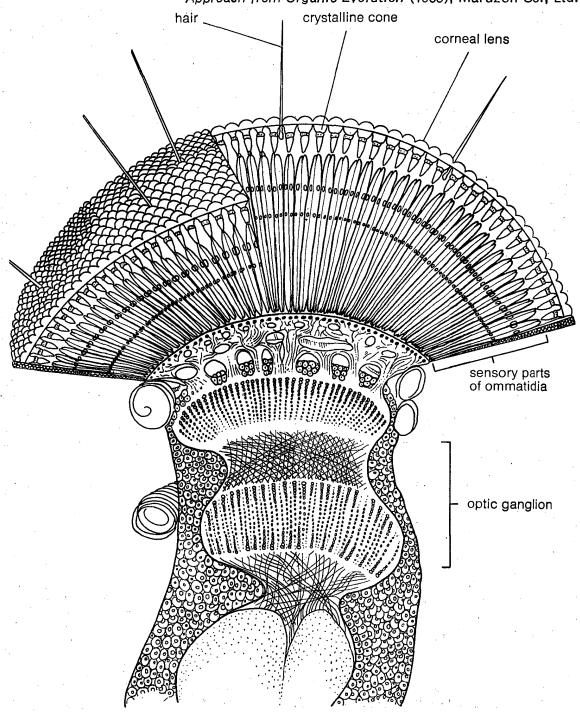


Figure 8: Section through the compound eye and optic lobe of a butterfly.

tracheole, or respiratory tubule, which supplies air to the ommatidium, has become specialized to form a tapetal filter comparable with that of the *Pecten* eye; it reflects highly saturated colours, usually blue colours in the dorsal ommatidia and more reddish colours in the ventral part of the eye. Each ommatidium thus has its own mirror. (The swallowtails, *Papilio*, are the only butterflies lacking this adaptation.) The light in the rhabdom of an ommatidium is tinted by the coloured light reflected from the mirror, and can be seen only with the aid of an ophthalmoscope—an instrument for viewing the interior of the eye. This filter system probably aids vision by enhancing contrast between objects of certain colours.

Butterflies with a tapetal filter, as well as nocturnal species (with a tapetum of different structure), have a corneal surface structure that acts as an anti-reflector. It consists of minute conical corneal "nipples" that provide a gradual transition between the refractive index of air and that of the cornea, effectively eliminating front-surface reflection for many wavelengths of light. One of the functions of this coating may be to minimize reflection of images from the tapetum lucidum back into the eye.

The rhabdomeres within an ommatidium in the compound eyes of flies (Diptera) have a microvillar structure (*i.e.*, one involving minute hairlike projections; see below *Morphological features*) and are completely separate from one another, in contrast to the apposition eyes above, in which the rhabdomeres are in contact. Each ommatidium contains eight retinular cells and eight rhabdomeres.

From any relatively distant point in space, a small cluster of ommatidia in a fly's eye appears darker than other areas because these ommatidia are the ones that are best aligned to absorb light coming from the direction in which the ommatidia are pointed. This cluster, called the pseudopupil, is often found in insect eyes; it moves, appearing to follow an observer viewing different parts of the compound eye. In the compound eye of the butterfly it is the pseudopupil that lights up with colours when observed and viewed from the same direction as the illumination with the ophthalmoscope. The pseudopupil of flies usually consists of seven ommatidia. In this apposition eye a distant point forms an image on the distal end of the rhabdomere, not within the crystalline cone as in the apposition eye of the bee. From any distant point in space only one rhabdomere in each ommatidium of the pseudopupil is illuminated by a point source; *i.e.*, one rhabdomere in each ommatidium of the pseudopupil is directed toward the same point.

Spatial resolution in the apposition eye

Experimental results suggest that the resolution of the compound eye (*i.e.*, the capability of the eye to distinguish between two separate but adjacent objects) is determined by the divergence angle between the ommatidia. If a compound eye views a pattern of alternating black and white bars in which the angle formed by one bar is the same as that between the ommatidia, the image of one black bar falls on one ommatidium, and the image of the flanking white bars falls on its nearest neighbours. This defines the limit of resolution; if the image of more than one bar falls in an ommatidium, the bars are not resolved.

The image is out of focus at the rhabdom in the bee eye. The angle between fly rhabdomeres is the same as the angle between ommatidia. The fused rhabdom of the bee and the separate rhabdomeres of the fly probably convey polarization and colour information rather than spatial information.

The array of corneal lenses on the surface of the eye can be considered as sources of light that interfere deep within the eye to give high-order diffraction images. The structure of the eye suggests that this effect is not detectable by the photoreceptors. Substantial shielding pigment between the ommatidia reduces the intensity of such patterns, which thus should be relatively ineffective in exciting the rhabdom.

The superposition eye. In the compound eyes of nocturnal arthropods, the rhabdoms are deep within the eye, far from the cornea and crystalline cone. In 1891 Sigmund Exner, an Austrian physiologist, reported experiments that showed how these eyes function. He demonstrated that there is pronounced pigment migration within the iris cells. In eyes adapted to darkness, the pigment of these cells migrates upward and into spaces between the crystalline cones of neighbouring ommatidia. In the light-adapted eye, on the other hand, the pigment migrates into the region between the cones and the rhabdom, in effect isolating each ommatidium by surrounding it with a light-absorbing tubular pigmented structure.

The pigment migration has the effect of changing the sensitivity of the eye. When a light-adapted eye is placed in the dark, there is an initial increase in sensitivity of about ten times with no accompanying pigment movement. In the following 25 minutes in the dark, the sensitivity increases about 100 times, as the pigment gradually withdraws from between the cones.

Exner found that the corneal process of the firefly *Lampyris*, a structure that resembles the crystalline cone of butterflies and other nocturnal insects, has a higher refractive index at the centre than near the periphery and effectively bends incoming light. When the eye is dark-adapted, light from many facets converges in the rhabdoms of many ommatidia forming a so-called superposition image (Figure 8). This eye differs from the apposition eye in that light from many facets is involved in forming an image in the rhabdom of an ommatidium; in the apposition eye, on the other hand, light from its own corneal lens reaches the rhabdom within a particular ommatidium. The mosaic image of the superposition eye, although less sharply defined than that of the apposition eye, is brighter and thus a valuable adaptation for nocturnal insects. In the light-adapted condition, inward migration of pigment in the iris cells effectively prevents the spread of light to adjacent ommatidia, and the superposition eye acts somewhat like the apposition eye; *i.e.*, only the light from one facet reaches the rhabdom within an ommatidium.

Important differences between apposition and superposition eyes

THE PROPERTIES OF PHOTORECEPTORS

The photoreceptor cell absorbs light energy and transforms it into a nervous response. The actual photoreceptor component, or organelle, of the photoreceptor cell contains a coloured substance (visual pigment) that absorbs light and initiates the chain of chemical reactions leading to nervous excitation. The vertebrate photoreceptor cells, the rods and cones, are so called because of the shape of their photoreceptors, which are found in the outer segments of the cells. Invertebrate photoreceptor cells, the photoreceptors of which are the rhabdomeres, show greater diversity of structure than do those of vertebrates.

The function of photoreceptors is dependent on the visual pigment they contain; the identification of photoreceptors ultimately depends, therefore, on correlating the light-absorbing properties of the pigments within them and the physiological response of the receptor cells. Such evidence exists for a number of invertebrates and vertebrates; the identity of some vertebrate and many invertebrate photoreceptors, however, has been inferred from their location, colour, and structure after comparison with known photoreceptors.

When viewed with the light microscope, the outer segments of vertebrate receptors and the rhabdomeres of invertebrates can be recognized by their shape, location, and high refractive index. In the electron microscope, all photoreceptors are seen to consist of a dense collection of membranes. The outer segments of vertebrate photoreceptors are lamellate (layered). Invertebrate rhabdomeres usually consist of microvilli (minute projections of the receptor cell membrane) and resemble a tiny honeycomb in three dimensions. In certain invertebrates, however, the rhabdomeres are lamellate, resembling those of vertebrates. Some features of photoreceptors from various animal groups are discussed below.

Morphological features. *Invertebrate photoreceptors.* The photoreceptor of the jellyfish *Polyorchis penicillatus* (Figure 9) is a modified cilium, or hairlike structure, as are the photoreceptors of many higher animals. Cilia are found in the cells of most animals; their motion is used either to move the animal or to move materials within or outside the body. As seen in the electron microscope, a cilium characteristically consists of nine double, peripherally arranged filaments (very fine threadlike structures) and a central pair of filaments in a so-called 9 + 2 pattern. A cilium arises from a cell structure known as a centriole, near which is often found a second centriole; ciliary rootlets extend from the centrioles to deep within the cell. In sensory cells a cilium often develops into a

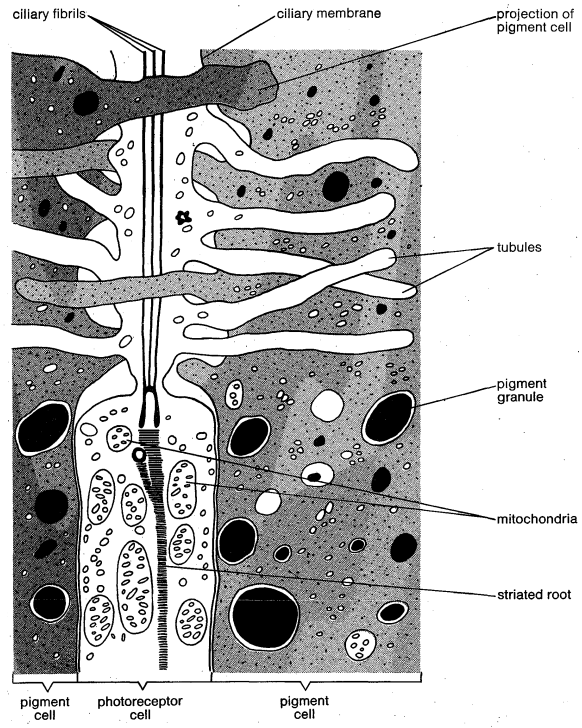


Figure 9: The distal part of a photoreceptor cell and two adjacent pigment cells in the ocellus of the hydromedusan *Polyorchis penicillatus*.

From *Proceedings of the U.S. Academy of Sciences* (1962) in D. Mazia and A. Tyler, *General Physiology of Cell Specialization*, Copyright 1963; used by permission of McGraw-Hill Book Company

sensory organelle. Sensory cilia invariably lack the central pair of filaments, having what is called a 9 + 0 pattern. The sensory cilium of *Polyorchis* contains numerous coarse microvilli, which constitute the sensory portion; centrioles and a ciliary rootlet are also present. Near the base of the photoreceptors are frequently found large numbers of cell components called mitochondria. Mitochondria manufacture the compound adenosine triphosphate, which seems necessary for the photoreception process.

The microvilli found in the primitive photoreceptor of *Polyorchis* are atypical in two respects: first, it is unusual for the microvilli of rhabdomeres to be derived from cilia; second, the microvilli of the rhabdomere, which are usually packed together very densely (hexagonal packing), are, in the jellyfish, mingled with the microvilli of the pigment-containing cells that envelop the rhabdomere. Migration of pigment into or out of the pigment-cell microvilli provides a mechanism for regulating the amount of light absorbed by the receptor. The identification of photoreceptor cells in the jellyfish has been based on anatomical and behavioral evidence. The cells are located in ocelli, the removal of which renders the jellyfish incapable of responding to illumination.

Although the ctenophorans (comb jellies) usually have been assumed to lack both a response to light and photoreceptor cells, recent evidence has raised some doubt about such a conclusion. Electron microscopic examination of *Pleurobrachia pileus* has revealed a radial arrangement of four groups of lamellate bodies; the lamellae are composed of membranes of about 12 sensory cilia (*i.e.*, 9 + 0 type). The sensory cilia, instead of developing microvilli as in *Polyorchis*, become individual platelets. These lamellate structures, which are located in infolded regions of the presumed photoreceptor cell, resemble similar lamellate structures in certain molluscan and vertebrate photoreceptors; it is presumed by some investigators, therefore, that the lamellate structures of *Pleurobrachia* may be photoreceptors.

The planarian flatworm *Dendrocoelum lacteum* (phylum Platyhelminthes) has photoreceptor cells with well-developed rhabdomeres (see Figure 10). The cytoplasm of the receptor cell, which has a tubular extension, contains many mitochondria; the surface membrane, which

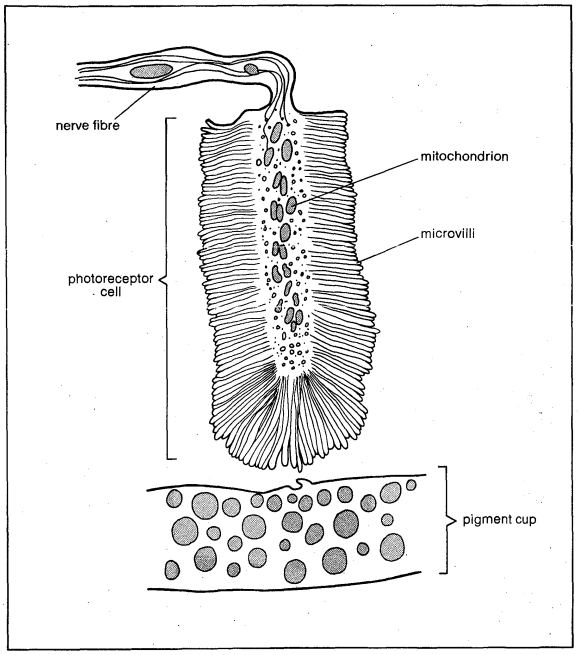


Figure 10: The photoreceptor cell in the flatworm *Dendrocoelum lacteum.*

From *Zeitschrift fur Zellforschung und mikroskopische Anatomie* (1961) in D. Mazia and A. Tyler (ed.), *General Physiology of Cell Specialization* (Copyright 1963); used by permission of McGraw-Hill Book Company

forms the tightly packed microvilli that constitute the photoreceptor organelle, is presumed to be the site of the visual pigment. No evidence has been found that either this photoreceptor or those of other flatworms are derived from cilia.

The nemertine worms (phylum Nemertea) have photoreceptor cells that resemble those of planarians when viewed with the light microscope. The photoreceptor cells on the eye of *Lineus ruber* resemble those of planarians in the electron microscope in that both have a rhabdomere consisting of microvilli. An important difference exists, however: the photoreceptors of *L. ruber* possess a filament that resembles a ciliary rootlet, even though no evidence exists that the rhabdomere is derived from cilia; in fact, cilia have not been observed in this cell. The function of the filament is not yet known.

Photoreceptors in the eye of the rotifer *Asplanchna brightwelli* (phylum Aschelminthes) are lamellate. Thin lamellae, arranged as leaves of a cabbage, apparently are folds of the receptor membrane. There is no evidence that the receptor is derived from cilia.

The photoreceptor cells in the eyes of the arrowworm *Sagitta scrippsae* (phylum Chaetognatha) have rhabdomeres consisting of structures of ciliary derivation that are called microtubules. The microtubules arise from a cone-shaped body filled with granules and cordlike structures. The cone-shaped body is derived from a sensory cilium. The microtubules comprising the rhabdomere are 50 nanometres in diameter and 20 long.

Photoreceptors of annelid worms

The segmented worms (Annelida) are the first invertebrate animal group in which the photoreceptors of a considerable number of species have been investigated. Studies indicate that annelids generally have rhabdomeres consisting of microvilli and are not of ciliary origin; however, at least one exception (*Branchiomma*) to this pattern has been observed.

The photoreceptor cell of the earthworm *Lumbricus terrestris* is an example of a rhabdomere comprised of microvilli; as mentioned previously, the microvilli of the membrane of the rhabdomere border on a vacuole within the receptor cell rather than on its outer surface. A number of sensory cilia are mixed with the microvilli and extend from the cell cytoplasm into the vacuole. The microvilli are not derived from cilia, and their function is as yet unknown. Although rhabdomeres consisting of microvilli have been found in several other annelids, no evidence of cilia has been reported. In the polychaete worm *Nereis vexillosa,* however, the microvillar rhabdomere arises from a tubular extension of the cell containing a centriole and a small fibre (fibril) with the structure of a ciliary rootlet; no direct evidence has been found thus far to link these ciliary vestiges to the rhabdomere.

The photoreceptor of the polychaete *Branchiomma vesiculosum* is a well-documented example of an annelid photoreceptor derived from cilia. The photoreceptor cell has a large invaginated (inpocketed) cavity filled with about 450 flattened lamellar sacs, which are the expanded flattened membranes of cilia having the 9 + 0 configuration. A collection of mitochondria and the nucleus of the receptor cell are displaced from the light path.

The Onychophora represent a transitional group that combines features of both the annelids and the arthropods. The photoreceptor cells of several species of onychophorans have been studied with the electron microscope. The photoreceptor cell of *Peripatus* closely resembles that of some annelids. The rhabdomere, which consists of microvilli, has a sensory cilium near its base. In studies of the *Peripatus* eye during development, no connection has been found between the sensory cilium and the developing microvilli. The function of the sensory cilium in this and in other photoreceptors in which no clear developmental relationship exists between the two is obscure.

Arthropods have few cilia, although they do occur in a few organs and provide the developmental basis for one type of arthropod ear, the chordotonal organ. Cilia have not been found in association with arthropod photoreceptors, however. The photoreceptor cells of all arthropods studied thus far are generally similar. Part of the cell surface forms densely packed microvilli, which is the photoreceptor containing the visual pigment.

The presumed photoreceptor in echinoderms is a collection of loosely packed microvilli extending from each receptor cell into the central cavity of the ocellus. In some species, each receptor cell has at least one cilium with the 9 + 0 pattern; this pattern is presumed to exist in other species. No relationship exists between the cilia and the microvilli; thus, the same condition prevails as in certain mollusks, annelids, and *Peripatus.* The echinoderms studied thus far appear to have a rhabdomere composed of microvilli and in close association with a sensory cilium.

Photoreceptor cells in mollusks and cephalopods

The mollusks (*e.g.,* scallops, clams, squid) have two morphologically distinct types of photoreceptor cells. One contains microvilli; of these, some photoreceptors have rudimentary cilia and fibrils such as those of certain annelids. Most mollusks appear to have this type of photoreceptor. The other type contains cilia; about 100 cilia on one surface of the cell develop into flattened sacs that are either packed tightly, like slices of bread, or curl on one another, like cabbage leaves. The ciliary type of rhabdomere is found in *Onchidium verruculatum* and in *Cardium edule.* The eyes of the bivalves *Pecten, Chlamys, Spondylus,* and *Amussium* contain double retinas, each with one of the types of photoreceptor. Evidence indicates that both types are photoreceptors and that they have different functions. The distal retina, with a ciliary photoreceptor, responds to a decrease in illumination; the proximal retina, with a microvillar photoreceptor, responds to increasing illumination.

The cephalopod retina has long cylindrical photoreceptor cells with rhabdomeres consisting of microvilli. In some cells the long axes of the microvilli are horizontal. The other type has vertical microvilli. The axes of both types of microvilli are parallel to a plane that is tangent to the retinal surface. This orthogonal arrangement of the microvilli in different cells—*i.e.,* in which they are perpendicularly disposed to one another—forms the physical basis for polarized light detection in the cephalopods.

Vertebrate photoreceptors. The photoreceptor of such primitive chordates as the urochordate ascidian tadpole is a modified cilium with the 9 + 0 pattern; the cilium

has developed into numerous lamellae. The cephalochordate *Branchiostoma californiense* may have both microvillar and ciliary photoreceptors. The Hesse cells, shown experimentally to be photoreceptor cells, have typical microvillar rhabdomeres with no cilia. Two other types of cells that may have a photoreceptor function are the dorsal ependymal cells (*i.e.*, special cells lining the cavities of the brain) and cells of the infundibulum, a ventral portion of the brain. Both cells have appendages containing cilia that form lamellae.

Other vertebrates have the two types of photoreceptor cells mentioned before: rods and cones. The rods are the photoreceptors for vision under conditions of dim illumination; the cones mediate daylight vision and colour sensation in many animals. The photoreceptors in both the rods and cones are composed of stacks of disks derived from cilia; the lamellae are stacked at right angles to the long axis of the cell and constitute the outer segments of rod and cone cells.

Most of the lamellae of the outer segment of the rod consist of free-floating disks within the limiting membrane of the outer segment (see Figure 11). The free-floating disks are formed at the base of the outer segment by a process of continuous infolding of the membrane of the outer segment. The process of new disk generation proceeds continuously during the life of the animal, and the entire outer segment of the rod is replaced in about one week. As new disks are being formed at the base of the rod, disks at the other end of the outer segment are being broken off and ingested by pigment cells; the rod length thus remains constant. As infolded regions of the membrane at the base of the rod move outward, they are pinched off to form free-floating disks. The outer segments of the cones, sometimes conical, or cone-shaped, and shorter than the outer segments of the rods, do not have free-floating disks. The process of infolding stops just after formation of the outer segments of the cones; infoldings occur along the entire length of the cones (see Figure 11). Because the membrane of the outer segments of the cones is not formed continuously during life, membrane components are continuously replaced. It has been shown that renewal takes place continuously along the entire surface of the membrane in cones. In rods, on the other hand, the renewal occurs only at the base, as new disks form.

Structural differences between rods and cones

The terms rod and cone are misleading, because these receptors cannot always be identified by shape; considerable variation exists between species. There are, nevertheless, a number of morphological differences between the rods and cones. The outer segments of the rods of many species are deeply incised—*i.e.*, lengthwise clefts in the disks give the outer segment a scalloped appearance in cross section. Both the rods and cones contain dense aggregations of mitochondria in the inner segment, or ellipsoid (see Figure 12). In the same region the cones of many species have an oil droplet, sometimes coloured, that filters the light before it reaches the receptor. An important difference between the photoreceptors of vertebrates and those of most invertebrates is that the latter give rise to nerve fibres that synapse either with cells in the central nervous system or with cells in a nerve ganglion within the eye; the vertebrate receptors, on the other hand, end in a region called a foot piece, into which nerve cells of the retina send synaptic projections.

From R.W. Young, "Visual Cells," Copyright © 1970 by Scientific American, Inc., all rights reserved

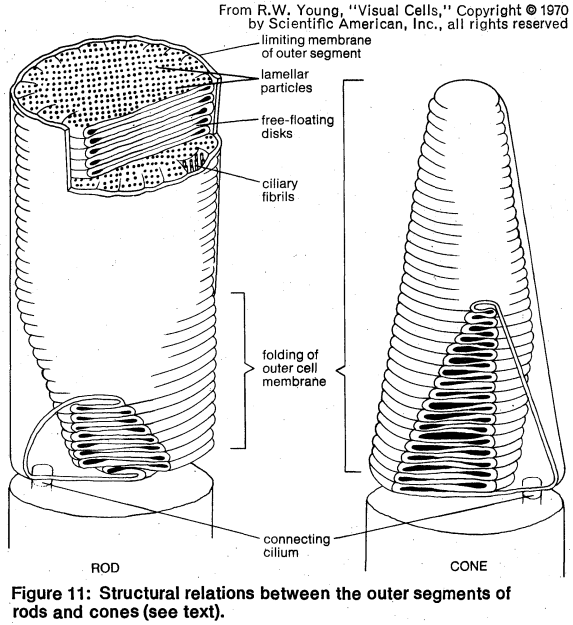


Figure 11: Structural relations between the outer segments of rods and cones (see text).

From F.S. Sjostrand, in G.K. Smelser (ed.), *The Structure of the Eye*; Academic Press, Inc., in D. Mazia and A. Tyler (eds.), *General Physiology of Cell Specialization* (Copyright 1963); used by permission of McGraw-Hill Book Company

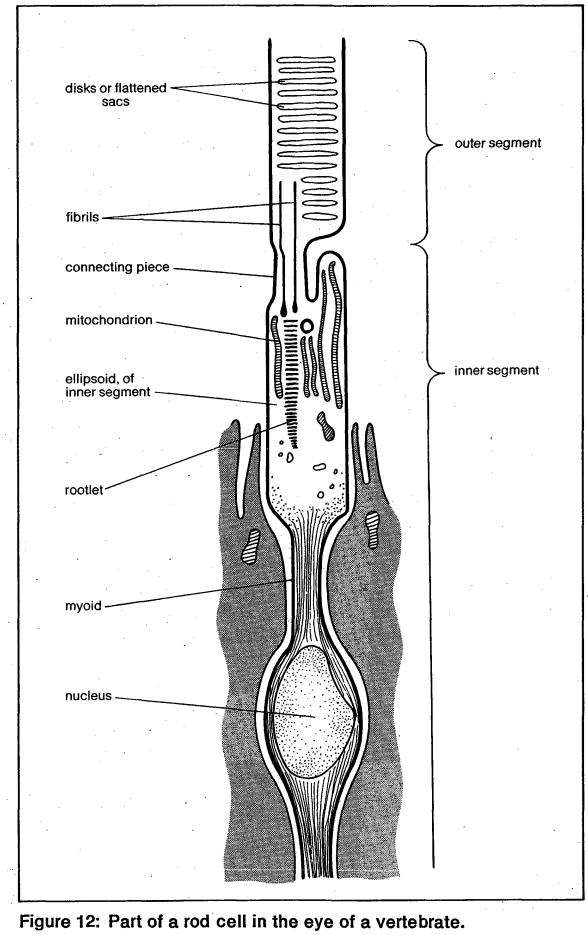


Figure 12: Part of a rod cell in the eye of a vertebrate.

Embryology and evolution of photoreceptors. There are two types of vertebrate eyes. The more familiar is the highly developed lateral eye. The other is the primitive median, or pineal, eye (in the top of the head). Only cyclostomes, reptiles, and amphibians have a median eye. It is best developed in lizards, in which the cornea is translucent and the number of receptors is small. The photoreceptor cells of the pineal eye are homologous (*i.e.*, structurally or developmentally related) with those of the lateral eye. They are, in fact, cones, the outer segments showing the characteristic infoldings along their entire length. In the median eye, the ends of the outer segments face the lens and light; in the lateral eye the retina is inverted, so that the ends of the outer segments face away from the light and toward the back of the eye.

The vertebrate retina is literally a part of the brain. The photoreceptor cells are derived from cells that line the neural tube, a hollow, dorsal structure that appears early in the embryo. The median eye develops from an outpocketing of the neural tube, and the cilia that develop into outer segments face the lens.

Invertebrates lack a neural tube; their photoreceptor cells develop from embryonic epithelial cells that send

out axons—threadlike extensions—that grow into the central nervous system. Often the epithelial cells contain cilia, and the photoreceptors develop from the cilia. The ciliary photoreceptors of invertebrates do not appear to be homologous with those of vertebrates. As mentioned above, it is of particular interest that photoreceptors of diverse structure and in many different phyla develop either from cilia or in close association with them. Still unknown is the developmental and functional role of the sensory cilia that are frequently found in association with photoreceptors of apparently nonciliary origin.

Optical properties of photoreceptors. *Rhabdomeres and outer segments as light guides.* The densely packed membranous structures of photoreceptors have a higher refractive index than the surrounding substances. The refractive index of the rhabdomere of the blowfly (*Calliphora*), for example, is 1.349, and that for the surrounding substance is 1.336; any light reflected from the inside surface of the rhabdomere is called dense to rare reflection. When the angle at which the light strikes the inside surface (angle of incidence) is such that the change of the angle of the light when entering the photoreceptor (*i.e.*, the angle of refraction) is 90°, all of the light is reflected back into the rhabdomere. The angle of incidence at which total internal reflection occurs is called the critical angle. The rhabdomere and outer segment trap the light entering it at angles equal to or greater than the critical angle and propagate it, essentially without loss of energy. The rhabdomere and outer segment thus act as an optical wave guide, or light pipe (see OPTICS, PRINCIPLES OF). Although all light entering the photoreceptor organelle—*i.e.*, all light incident at angles equal to or less than the critical angle—is propagated by successive internal reflection, only certain angles of reflection occur; these are determined by various physical characteristics of the organelle. As a result, the light within the photoreceptor light guides is propagated in modes, or patterns, of energy. This is important for the photoreceptive function of the rhabdomere and outer segment because the amount of information the organelle can carry is related to the number of modes propagated. The thinnest photoreceptor organelles, such as those of certain fly rhabdomeres with a diameter of 0.5 nanometre, can, for example, support only one mode, called the lowest order mode. The amount of energy propagated in the rhabdomere in this mode depends on the wavelength of light, so that the shorter wavelengths are best propagated. Thus reddish light is not made available to the rhabdomere's visual pigment as readily as bluish light; the light guide properties of the organelle, therefore, can influence the physiological response of the photoreceptor. The light guide also profoundly affects the amount of image information transmitted because the greater the number of modes, the more image information. This last consideration is of particular importance for fused rhabdoms, in which image information apparently is not transmitted (although the question is not yet entirely settled). These light guide properties are of importance for still another reason. A small amount of the modal energy propagated by a light guide is actually outside the structure. Thus when, as described in the next section, pigment granules come near the photoreceptor organelle, they control the amount of light in the organelle by absorbing the energy propagating outside, so that it can no longer be transmitted by the photoreceptor. Such pigment granules may exert an even more profound effect by raising the refractive index outside the organelle, thus destroying its guiding properties and causing the light to spread into adjoining structures where it is absorbed, but not by visual pigment; absorption thus does not result in sensation.

Migration of pigment granules

Photomechanical light and dark adaptation. A mechanism that may partly control the amount of light within the photoreceptor under widely different illumination conditions involves the migration of photoprotective pigment granules either within the receptor or within neighbouring cells so as to envelop the photoreceptor; it has been observed in both vertebrates and invertebrates. Pigment migration in response to illumination takes place in the retinas of most vertebrates except mammals. The most pronounced effects are found in fishes, frogs, and toads; they also occur in reptiles and in diurnal and nocturnal birds.

Pigment migration is combined with photomechanical movements of the photoreceptor cells of many lower vertebrates (Figure 13). In darkness the pigment granules

From Dr. Samuel R. Detwiller, *Vertebrate Photoreceptors* (1943); Macmillan & Co.

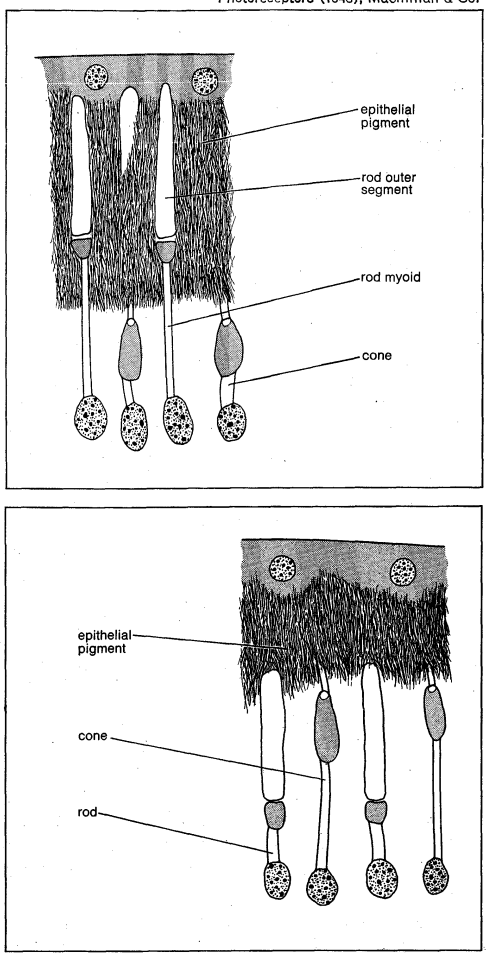


Figure 13: *Adaptation of retina.*
(Top) Theoretical light-adapted retina, showing migrated epithelial pigment, elongated rods, and contracted cones. (Bottom) Theoretical dark-adapted retina, showing contracted pigment, elongated cones, and contracted rods.

of the pigment epithelium are withdrawn, as shown in the figure. The rod myoids (contractile elements) are contracted in darkness so that the rod outer segments are withdrawn from the pigment and exposed fully to the available light. On the other hand, the cone myoids are extended, so that their outer segments are within the pigment, which functions effectively to reduce the amount of light in the outer segment, as explained in the previous section. The cone outer segments are enveloped by pigment, presumably to help suppress responses from the daylight receptors under conditions of dim illumination. Under the influence of light the rod myoids extend within several minutes, and the pigment envelops the rod myoids to help suppress the diurnal activity of the rod system (pigment surrounds only the rod myoids, not the outer segments [see Figure 13, top]). At the same time the myoids of the cones contract to withdraw the cone outer segments from the pigment. As mentioned above, neither photomechanical nor pigment movements occur in mammalian eyes; switching between rod and cone systems is accomplished neurally.

The envelopment of rhabdomeres by pigment migration is widespread in invertebrates. An example of this phenomenon occurs in the ommatidium of the amphipod *Gammarus ornatus*, in which the rhabdomeres of five radially arranged receptor cells form a star-shaped rhabdom. Extensions of the receptor cells are directed toward

the cornea next to the lenslike crystalline cone and away from the cornea to cell bodies beneath a so-called basement membrane. In the dark-adapted condition the pigment is located around the cone and in the cell body. The rhabdom functions as a light guide within which is the visual pigment, which absorbs light. In the light-adapted condition, the pigment migrates from the receptor cell body to surround the rhabdom completely. Similar cell pigment migrations occur within the receptors of many invertebrates. In the compound eye of the horseshoe crab *Limulus*, the rhabdom is also star-shaped in cross section, but the rays of the star are much longer than are those of *Gammarus*. In the eye of *Limulus* the pigment migrates radially within the receptor cell, completely enveloping the rays in a matter of minutes in bright light and requiring about one hour to withdraw entirely from the rhabdom area in darkness.

The rhabdom of the migratory locust (*Schistocerca*) shows an interesting variation. In the light-adapted retina, mitochondria migrate and become tightly packed around the rhabdom. This may be a device for providing energy for the rapidly metabolizing light-activated photoreceptor. There is physiological evidence, however, that, in the light-adapted condition, the mitochondria, which have a high refractive index, increase the critical angle by coming close to the rhabdom. They thereby cause loss of light from the rhabdom and a decrease in the "field of view" of the rhabdom. When the eye is dark-adapted, the space around the rhabdom is replaced by fluid-filled spaces, with a low refractive index; they return the rhabdom to an efficient light guide mode of operation.

Birefringence and dichroism. The physical properties of certain crystals, glass, liquid, and gas, are the same regardless of the direction of the light propagated through them; for example, the speed and thus the refractive index of light are the same regardless of its direction of propagation. Such a medium is said to be isotropic. In other crystals, the atoms are arranged so the refractive index varies with the direction of propagation of light; such crystals are said to be anisotropic. If a medium has a crystalline arrangement that retards light in a particular direction, it is birefringent. In such an arrangement an entering ray of light is divided into two rays that are polarized in planes at right angles to one another.

When birefringence can be made to disappear by immersing a substance in a liquid with an appropriate refractive index, the birefringence is caused by an orderly arrangement of isotropic particles submicroscopic in size, such as the limiting membranes surrounding cells. Such birefringence is called form birefringence to distinguish it from the intrinsic molecular crystalline birefringence described above. Very small rods, lying parallel to each other, cause positive birefringence; parallel platelets cause negative birefringence. Studies have shown that rods in frogs have a negative form birefringence caused by the platelets and a positive intrinsic birefringence. These phenomena are thought to result from the ordered arrangement of a lipid (fat) layer two molecules thick in the platelet membranes.

Birefringence in the rods of frogs

Variation in the colour of light absorbed dependent on the direction of polarization of the light is termed dichroism. This property is a sensitive indicator of the orientation of molecules in a structure; dichroism in photoreceptors, for example, results from the ordered arrangement of the visual pigment molecules. The visual pigment of outer segments is dichroic, as are the outer segments. Dichroism of the outer segment can be demonstrated only by measuring the absorption of polarized light shone through the side of the outer segment. Light that propagates in the usual direction, that is down the long axis of the outer segment, does not show different absorption properties that depend on the direction of polarization. The fact that there is not dichroism for light propagating along the long axis, but there is for light propagating perpendicular to the long axis, indicates that the visual pigment molecules are oriented at random but with the long axes in the plane of the disks.

Rhabdoms consisting of microvilli can detect the plane of polarization of polarized light. Invertebrates with such rhabdoms use polarization properties of the blue sky opposite the Sun for navigation. In animals with the ability, the rhabdomeres comprising a rhabdom show orthogonal orientation. It has been shown that the microvilli are dichroic and that the dichroism is caused by the visual pigment. The orientation of the microvilli provides the mechanism for sensation of polarized light. Vertebrates cannot detect polarized light because of the random orientation of visual molecules within the plane of the platelets.

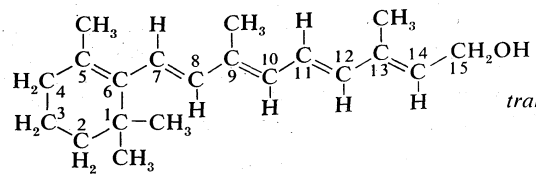
Stiles–Crawford effect. In the Stiles–Crawford effect, first observed in humans, the luminous efficiency of light has been shown to be greater when entering the centre than when entering the edge of the pupil. This effect may be considered an optical property of the cone receptor cells. The energy acceptance, or radiation, pattern of photoreceptors is a consequence of the optical wave guide properties resulting from the small size of the receptors. The Stiles–Crawford effect is not observed with dim illumination (rod vision), probably because the acceptance angle of the rods is wider, and it essentially fills the pupil (see also EYE AND VISION, HUMAN).

Visual pigments. As light propagates through the photoreceptor, the visual pigment absorbs photons (light particles). The absorption of one photon by a visual pigment molecule can initiate events that lead to nervous excitation. The photoreceptors in the eyes of both vertebrates and invertebrates consist of tightly packed membranes (lamellae), the function of which is to support the visual pigment and other molecules necessary for excitation. The photoreceptor membrane is a lipid layer two molecules thick, with protein either between the layers, coated over the layers, or both. The outer segment of the frog rod is 40 percent lipid and 60 percent protein; more than 80 percent of the protein is visual pigment.

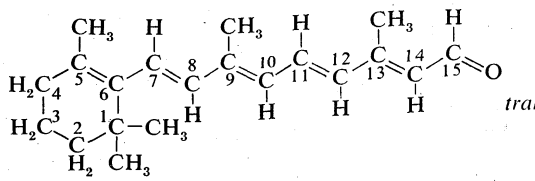
Components. Visual pigment has two components: the light-absorbing chromophore (a chemical group that produces colour) and the protein moiety (opsin) to which it is chemically attached. The structure of the naturally occuring chromophore, retinal, is thought to be the same in all animals. Before it has been exposed to illumination, the chromophore is 11-*cis*-retinal, an alternate structure of vitamin A aldehyde. Upon exposure to light each chromophore molecule absorbs one photon of light. The absorption of a photon changes 11-*cis*-retinal to *trans*-retinal. (See structural formulas.) The resulting change in the shape of the molecule causes the separation of retinal from its protein component and gives rise to a number of coloured intermediates. In the dark, 11-*cis*-retinal combines spontaneously with the protein component to regenerate the visual pigment.

Characteristics of retinal

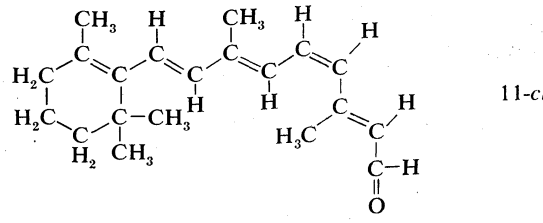
Hundreds of visual pigments have been identified and



trans-retinol



trans-retinal



11-*cis*-retinal

characterized by the wavelength of light that is best absorbed by them (absorption maximum, or λ max); the wavelength maxima range from 432 nanometres for the green rod of the frog to 625 nanometres for the red-absorbing cone of the goldfish. The reason for such a great range when the chromophore is the same in all visual pigments lies in part in properties of the chromophore; much of the variation, however, resides in the protein component. Variation in protein structure among different animals gives rise to different absorption maxima.

There are two forms of vitamin A: retinol A_1 and retinol A_2. The structure of A_1 differs only slightly from that of A_2. The effect of this slight structural difference is to increase the wavelength for maximum absorption by as much as 20 to 60 nanometres, depending on the pigment involved. The pigments with the shortest λ max are A_1 pigments; those with the longest are A_2. There is a broad area of overlap within the two groups, however. In general, marine fishes, land vertebrates, arthropods, and mollusks have A_1 pigments; freshwater fishes have A_2 pigments.

The protein components, or opsins, of visual pigments have not yet been fully characterized. The visual pigment of rods, a combination of retinal and opsin, is called rhodopsin, or visual purple. Estimates for the molecular weights of the opsins found in rods centre around 40,000. Squid opsin, however, has a molecular weight of about 70,000. Analyses of the visual pigments from the rods of squid and vertebrates suggest that the protein components are quite different. The nature of cone pigment proteins is thus far virtually unknown.

Important to an understanding of the photochemical reaction is the nature of the attachment of the chromophore to the opsin. Although controversy still exists on this point, it is clear that the visual pigment is part of the photoreceptor membrane and is very closely attached to its lipid and protein.

The role of visual pigments in vision. The sensitivity of the dark-adapted human eye can be measured as a function of wavelength of the stimulating light falling on that part of the retina in which the density of rods is greatest. When such a curve is compared with a curve of the absorption spectrum (the entire range of wavelengths of light absorbed) of human rhodopsin in the outer segments of the rods, the two curves correspond exactly. This is strong evidence that rhodopsin initiates the visual process for light absorbed in the outer segments.

The problem of cone pigments is an important one. More than 100 years have passed since it was concluded after a comparative study of retinas that the rods of vertebrate retinas mediate vision in dim light and the cones mediate daylight and colour vision. Recently, a method called microspectrophotometry has been developed for the study of cone pigments and other visual pigments. By projecting a minute beam of light onto a single photoreceptor, the absorption spectra of a variety of vertebrate and invertebrate photoreceptors have been measured. The absorption spectra of single human and other primate cones have thus been characterized. There are three types: a blue absorbing cone, a green absorbing one, and a yellow absorbing one. Originally suggested in 1802, these three cone types, now positively identified, can quantitatively explain human colour sensation.

Absorption spectra of photoreceptors

Such methods have been helpful in identifying receptors and their visual pigments in animals other than man; for example, in the retina of the frog *Rana pipiens*, a green rod and a red rod have been identified. A double cone with principal and accessory pigments and a single cone have also been characterized. Microspectrophotometry has been used to obtain evidence of visual pigments in rhabdomeres of the fly *Calliphora*.

PHYSIOLOGICAL RESPONSE OF PHOTORECEPTORS

The initial molecular events of photoreception can be summarized: after the absorption of a photon of light, 11-*cis*-retinal is converted into *trans*-retinal; the process disrupts the binding of retinal to opsin and results in the sequential formation of a number of coloured photoproducts.

The final physiological effects of photoreception have also been well defined. Excitation of visual pigments results in a change in certain properties (*i.e.*, permeability and potential) of the limiting membrane of the photoreceptor cell. Because the initial and final molecular events of photoreception are now understood, there is considerable hope that the entire chain of molecular events between the absorption of light by the visual pigment and the changes in potential of the cell membrane can eventually be reconstructed. This challenging problem has stimulated research concerned with the nature of the photoreceptor membrane and the chemistry of the photoproducts.

The internal environment of cells is generally different from that outside. The main cation (positively charged atom) inside cells is usually potassium; the main cation outside is sodium. In the resting, or unexcited, state, the cell membrane of most excitable cells (such as receptor, nerve, and muscle cells) is impermeable to sodium ions but permeable to potassium ions. A separation of charges occurs on either side of the membrane because the positively charged potassium ions diffuse outside of the cell, leaving a surplus of nondiffusing negatively charged atoms (anions) inside the cell. The result is a difference in potential across the cell membrane; this difference is proportional to the logarithm of the ratio of the concentration of potassium ions outside to that inside the cell. Photoreceptor cells are negative by some 30 to 60 millivolts (1 mV = 0.001 volt) inside relative to the outside, the exact value depending on the specific cell. This negative potential difference is termed the resting potential.

Receptor potential. The absorption of a single photon by the visual pigment in a photoreceptor may result in a physiological response of the receptor cell. Such a response of the cell can be measured as the receptor potential—that is, a change in the potential of the cell membrane caused by a change in the permeability of some part of the membrane to certain ions. The permeability change is in turn related to the absorption of light by the chromophore through a chain of molecular events as yet unknown. The receptor potentials of the eyes of animals from a number of phyla have been investigated. In certain photoreceptors the receptor potential is positive in sign, or depolarizing; in others it is negative, or hyperpolarizing.

Definition of receptor potential

Depolarizing receptor potentials. The depolarizing receptor potential, a change in potential response to illumination that is positive in sign, is characteristic of photoreceptor cells in many invertebrate phyla. One of the best known photoreceptors is the retinular cell in the compound eye and in the ventral eye of the horseshoe crab *Limulus*, an arthropod. The membrane potential of this cell is easily measured by means of a very small electrode (a device that can be placed inside a cell to measure the potential across the cell membrane). When the electrode penetrates the cell in the absence of light, the potential falls to some negative value, perhaps −40 millivolts, which is the resting potential. Even in complete darkness, however, variation in the resting potential occurs in the form of slow fluctuations. Although these spontaneously and randomly occurring fluctuations presumably seldom result in a physiological visual response, they are miniature receptor potentials. Each is accompanied by a decrease in the resistance of the membrane (*i.e.*, it becomes a better conductor) and an increase in the permeability of the membrane to sodium ions. When the receptor is illuminated with dim light, the slow potential fluctuations increase in frequency, and, with brighter light, they become briefer and higher in frequency, finally fusing to form a more or less steady depolarizing receptor potential. Thus, a short time after the cell is exposed to light, the membrane becomes depolarized; when the light is removed, the receptor potential decays, and the resting potential is restored. The extent of the receptor potential increases with increasingly stronger illumination; it has a linear relationship to the logarithm of the intensity of illumination. Whatever properties are responsible for this transformation are responsible for the photoreceptor's enormous dynamic range: despite the fact that the recep-

Relationship between receptor potential and illumination

tor potential can change at most on the order of 100 millivolts, the receptor can respond over a range of more than ten orders of magnitude, a range of intensities of over 100,000,000,000.

In all species with this depolarizing receptor potential, the fact that the resistance of the photoreceptor cell membrane decreases during the receptor potential suggests that it becomes more permeable to some ion or ions during this time. Evidence as to the nature of the permeability change suggests that it is principally an increase in permeability to sodium ions that causes the receptor potential. There is additional evidence for this conclusion (see *Bibliography*).

The depolarizing receptor potential is characteristic of the arthropod photoreceptor consisting of microvilli and the receptor cell. Species from most of the orders of Arthropoda have now been investigated, and exceptions have not been reported. Usually, the photoreceptor cell gives rise only to a receptor potential. The receptor potential is conducted passively a short distance to higher order nerve cells via a nerve fibre. The higher order nerve cells convert the receptor potential to propagated action potentials (nerve impulses) for transmission of the information over longer distances.

The proximal sense cell of the double retina of *Pecten* is similar to the above in that illumination gives rise to a depolarizing receptor potential and a decrease in resistance. The morphology of the proximal sense cell is also similar to that of the arthropod photoreceptors in that it has coarse microvilli. But the proximal sense cell of *Pecten* has a long nerve fibre that carries nerve impulses. The depolarizing receptor potential activates the mechanism that initiates an action potential, and steady depolarization of the axon gives rise to a series of nerve impulses; the stronger the intensity of illumination, the greater the extent of the receptor potential and the higher the frequency of nerve impulses. Nerve activity of this nature is rarely found in arthropod photoreceptor cells.

In addition to the arthropods and *Pecten*, a proximal sense cell in the leech (*Hirudo*), an annelid, has the depolarizing type of receptor potential.

Hyperpolarizing receptor potential with decrease in resistance. As noted above, the proximal sense cells of the *Pecten* eye have a depolarizing receptor potential, and the nerve responds to illumination with a series of nerve impulses. The distal sense cells, which also have nerve fibres, respond to a decrease in illumination with nerve impulses. This is called an off response. Recent studies indicate that the receptor potential of the distal sense cell is a hyperpolarizing potential, a potential change that is negative in sign. When the cell is depolarized, the extent of the receptor potential increases; when the cell is sufficiently hyperpolarized, the receptor potential reverses polarity and becomes depolarizing. The receptor potential of the distal sense cell is accompanied by a decrease in resistance. The hyperpolarizing potential is believed to be caused by an increase in permeability to potassium and perhaps to other ions.

The photoreceptor of the *Pecten* distal sense cell is a ciliary derivative. A primitive chordate, the tunicate *Salpa*, which has a microvillar photoreceptor, has the same type of hyperpolarizing photoreceptor potential. There thus does not seem to be any association of physiological potential types with structurally or embryologically similar photoreceptors.

Hyperpolarizing receptor potential with increase in resistance. The rod and cone photoreceptors of vertebrate eyes have a third type of receptor potential. In the dark, sodium ions flow steadily into the outer segment; light decreases this dark current—*i.e.*, the cell becomes hyperpolarized, and its resistance increases. If the cell is sufficiently hyperpolarized under experimental conditions, depolarization occurs; the receptor potential thus is reversed, as would be expected if the permeability change involves principally sodium ions.

Interpretation of the physiological response. As a nerve cell is depolarized, it discharges nerve impulses; as it becomes hyperpolarized, the membrane potential returns to the resting potential. The fact that certain photoreceptor cells are hyperpolarized by illumination is thus confusing and raises the possibility that light is inhibiting the receptors. In the generation and transmission of information about illumination, however, the sign of the response may actually be unimportant. It has been known for some time that there are several types of optic nerve fibres in vertebrate eyes; some discharge impulses during illumination, others discharge either only at the onset and cessation of illumination or merely at the cessation. In other words, information about the cessation of illumination is at least equal in importance to information about its onset and duration. Such observations of optic nerve activity prove that the hyperpolarizing response of the receptor, the only response of the receptor, is translated into a code that includes the discharge of nerve impulses both during and after illumination. Thus, the hyperpolarizing response of the receptor, although it may be important in emphasizing the off response, does not result in the loss of information about excitation. Because the hyperpolarizing response is definitely used to produce excitation at a later stage, it is misleading to consider it an inhibitory response. The same is true in reverse for the depolarizing response.

The way in which initial photoreception is processed by the retina and higher centres to cause visual perception is beyond the scope of this article. It may be mentioned, however, that the initial information used in this processing is the direct result of the physical and chemical properties of the photoreceptor.

BIBLIOGRAPHY. M.H. PIRENNE, *Vision and the Eye*, 2nd ed. (1967), optics and physiology of vision (vertebrate and invertebrate eyes) for the beginner and non-specialist; *Handbook of Sensory Physiology:* vol. 7, pt. 1, H.J.A. DARTNALL (ed.), *The Photochemistry of Vision*; vol. 7, pt. 2, M.G.F. FOURTES (ed.), *Physiology of Photoreceptor Organs* (1971–72), authoritative treatises by leading scientists; C.G. BERNHARD (ed.), *The Functional Organization of the Compound Eye* (1966), on the optics, morphology, photochemistry, and physiology of the compound eye and other primitive eyes; GORDON L. WALLS, *The Vertebrate Eye and Its Adaptive Radiation* (1942, reprinted 1963), a classic and encyclopaedic store of knowledge of the vertebrate retina and eye; TSUNEO TOMITA, "Electrical Activity of Vertebrate Photoreceptors," *Q. Rev. Biophys.*, 3:179-222 (1970), a summary of vertebrate photoreceptor physiology; T.H. GOLDSMITH and G.D. BERNARD, "The Visual System of Insects," in MORRIS ROCKSTEIN (ed.), *The Physiology of Insecta* (1972), on the visual system of insects, optics, visual pigments, and physiology; M.F. MOODY, "Photoreceptor Organelles in Animals," *Biol. Rev.*, 39:43–86 (1964); RICHARD M. EAKIN, "Lines of Evolution of Photoreceptors," in DANIEL MAZIA and ALBERT TYLER (eds.), *General Physiology of Cell Specialization* (1963), and "Structure of Invertebrate Photoreceptors," in H.J.A. DARTNALL (ed.), *Photochemistry of Vision* (1972); three articles on photoreceptor structure and function; BRADLEY R. STRAATSMA *et al.* (eds.), *The Retina* (1969), on the morphology, function, and clinical characteristics of the vertebrate retina; HUGH DAVSON (ed.), *The Eye*, vol. 2, *The Visual Process* (1962), a textbook of photoreception and function of the visual system; the *Journal of the Optical Society of America*, vol. 53, no. 1 (1963); the Cold Spring Harbor Symposia on Quantitative Biology, vol. 30, *Sensory Receptors* (1965); and the Proceedings of the International School of Physics, "Enrico Fermi," course 43, ed. by W. REICHARDT, *Processing of Optical Data by Organisms and by Machines* (1969), all contain recent research on photoreception and retinal function; H.J.A. DARTNALL, *The Visual Pigments* (1957); and GEORGE WALD, "Molecular Basis of Visual Excitation," *Science*, 162:230–239 (1968), reviews of visual pigment chemistry; T.H. BULLOCK and G.A. HORRIDGE, *Structure and Function in the Nervous Systems of Invertebrates*, vol. 2 (1966), contains a broad survey of invertebrate sensory receptors; H.K. HARTLINE, "Visual Receptors and Retinal Interaction," *Science*, 164: 270–278 (1969); FLOYD RATLIFF, *Mach Bands: Quantitative Studies on Neural Networks in the Retina* (1965), physiological effects on vision of neural activity in the retina.

(W.H.M.)

Photosynthesis

Photosynthesis means putting together with light energy. During photosynthesis in green plants, light energy is captured and used to convert water, carbon dioxide, and minerals into oxygen and energy-rich organic compounds.

GENERAL CHARACTERISTICS

Importance. Without photosynthesis, the earth's atmosphere would be devoid of oxygen. All living cells convert food into energy and structural components necessary for life; if plant photosynthesis were to stop, most living things would disappear from the earth in a few years. Growth, repair, reproduction, and other dynamic processes performed by organisms require light energy that is absorbed by chlorophyll pigments in green plants and stored as chemical energy in organic products, especially carbohydrates. The chemical energy and electrical energy used by brain cells, for example, were once sunlight that was absorbed by the chlorophyll in green plants.

In the past, green plants, growing in warm climates in the presence of more carbon dioxide than the atmosphere contains now, increased in numbers faster than they were consumed; as a result, abundant supplies of "fossil fuels" (*i.e.*, coal, oil, and gas) were deposited in the earth's crust. The combustion of these fuels provides much of the energy now used in factories, in homes, and in transportation.

Increasing requirements for food in a world in which human population is increasing exponentially requires not only an increase in the amount of photosynthesis on earth but also an increase in the efficiency of converting photosynthetic products into food for man. The possibility of developing methods for synthesizing food commercially depends in part on the costs involved and on the population pressure. If the cost of energy were reduced, the synthesis of food from water, minerals, and carbon dioxide might become economically feasible. Food synthesis is a remote prospect, however, because the cost of power plants cannot compete with photosynthesis, in which the energy supply is free.

Overall reaction of photosynthesis. In many green plants, carbohydrates are the most important direct organic product of photosynthesis (see Figure 1). The formation of a simple carbohydrate, glucose, is indicated by a chemical equation,

$$6CO_2 + 12H_2O \xrightarrow[\text{green plants}]{\text{light}} C_6H_{12}O_6 + 6O_2 + 6H_2O$$

carbon dioxide — water — glucose — oxygen — water

Little free glucose is produced in plants; instead, glucose units are linked together to form starch, or are joined with fructose, another sugar, to form sucrose (see CARBOHYDRATE).

Not only carbohydrates, as was once thought, but also amino acids, proteins, lipids (or fats), pigments, and other organic components of green tissues are synthesized during photosynthesis. Minerals supply the elements (*e.g.*, nitrogen, N; phosphorus, P; sulfur, S) required to form these compounds. A generalized equation that expresses the total photosynthetic process is:

$$kCO_2 + lH_2O + mNO_3^- + nSO_4^= + \text{other materials} \xrightarrow[\text{green plants}]{\text{light}} \text{organic compounds} + xO_2$$

carbon dioxide — water — nitrate — sulfate — oxygen

The specific organic compounds formed and the quantities (k,l,m,n,x in the equation above) of minerals, water, and other necessary components required vary with the species, physiological state, and environment of a plant. In the overall reaction, chemical bonds are broken between oxygen (O), carbon (C), hydrogen (H), nitrogen, and sulfur, and new bonds are formed in products that include oxygen and organic compounds containing the elements. More energy is required to break bonds (*e.g.*, in water, sulfate) than is released when new bonds form in the products. This excess bond energy, provided by light, accounts for a large part of the light energy stored as chemical energy in the organic products formed during photosynthesis.

Evolution. Although life and the quality of the atmosphere today depend on photosynthesis, it is likely that green plants evolved long after the first living cells appeared on earth. When the earth was young, complex molecules probably were synthesized from hydrogen (both gaseous and combined, as hydrocarbons and water) and the action of ultraviolet radiation from the sun. The first living cells probably evolved from these molecules (see LIFE). The accidental joining together (condensation) of glycine, an amino acid, and acetate, a fatty acid, for example, may have formed complex organic molecules known as porphyrins; these molecules, in turn may have evolved further into coloured molecules called pigments; *e.g.*, chlorophylls of green plants, bacteriochlorophyll of photosynthetic bacteria, hemin (the red pigment of blood), and cytochromes, a group of pigment molecules essential in both photosynthesis and respiration.

Primitive coloured cells then had to evolve mechanisms for using the light energy absorbed by their pigments. At first, the energy may have been used only to initiate reactions useful to the cell. As the process for utilization of light energy continued to evolve, however, increasingly larger amounts of light energy probably were stored as chemical energy, to be used as required in the processes necessary to maintain life. Green plants, which have a mechanism for utilizing light energy in cell organelles called chloroplasts and are capable of storing up to 30 percent of absorbed light energy, are the culmination of the evolutionary process.

Bacterial photosynthesis

The evolutionary pathway that led to the photosynthetic bacteria is distinct from the one that resulted in green plants. Photosynthetic bacteria use absorbed light energy less efficiently than do green plants, although both can synthesize organic compounds from carbon dioxide and inorganic minerals. Bacterial photosynthesis always differs from photosynthesis in green plants in two important ways: first, no oxygen is produced; second, "far-red light" can be absorbed and used by bacterial chlorophylls. This light has a longer wavelength and less energy than the light used by green plants. Although bacterial photosynthesis is of little quantitative importance, its study and comparison with photosynthesis in green plants have been valuable in understanding both these processes.

From *Biological Sciences: An Inquiry into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich, Inc., New York. By permission of the Biological Sciences Curriculum Study

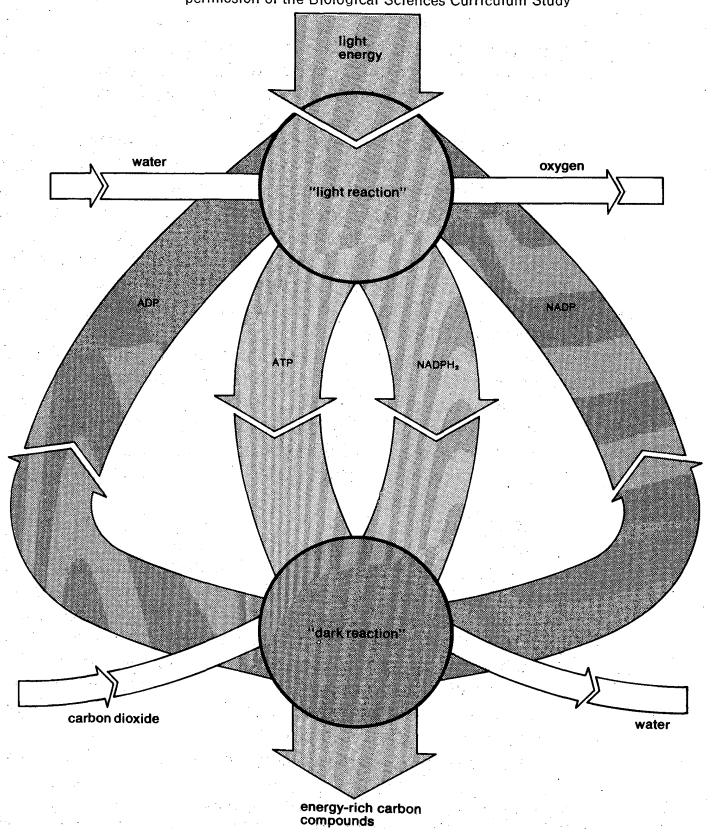


Figure 1: Summary of photosynthesis (see text).

FACTORS THAT INFLUENCE THE RATE OF PHOTOSYNTHESIS

The rate of photosynthesis is defined in terms of the rate of oxygen production either per unit mass (or area) of green plant tissues or per unit weight of total chlorophyll. Active, healthy plants growing in air, which contains 0.03–0.04 percent carbon dioxide, at 25° C are able to produce several litres of oxygen per hour for each gram of chlorophyll. This amount of chlorophyll is contained in about one kilogram (2.2 pounds) of wet tissue of an active sun-loving plant such as spinach. The single-celled alga *Chlorella pyrenoidosa* contains about four grams of chlorophyll per kilogram of wet tissue, or four times the amount in a kilogram of spinach. Oxygen production can be raised severalfold by increasing the carbon dioxide level to 1 percent and, sometimes, by raising the temperature to 35° C or more. The algae must be suspended in liquid 100 times their volume or spread as a thin film, however, so that the organisms obtain sufficient air and light. These requirements make more difficult the task of astronautical engineers who want to design spaceship systems that use photosynthesis to renew oxygen supplies to remove carbon dioxide, and to provide food during flights in outer space. For similar reasons, the culture of algae has not yet been developed as an economic process for large-scale food production on earth. As the human population increases, however, it is likely that methods will be developed for conversion of algae into foodstuffs for man.

Environmental factors

The amount of light, the carbon dioxide supply, and the temperature are the three most important environmental factors that directly affect the rate of photosynthesis; water and minerals in sufficient quantities also are necessary. The rate of photosynthesis in green plants is determined by the plant species and its physiological state; *e.g.*, health, maturity, whether or not it is in flower.

Light intensity and temperature. The complex mechanism of photosynthesis includes a photochemical, or light-dependent, stage and an enzymatic, or dark, stage that involves chemical reactions. These stages can be distinguished by studying the rates of photosynthesis at various degrees of light saturation (*i.e.*, intensity) and at different temperatures. Over a range of moderate temperatures and at low to medium light intensities (relative to the normal range of the plant species), the rate of photosynthesis increases as the intensity increases and is independent of temperature. As the light intensity increases to higher levels, however, the rate becomes increasingly dependent on temperature and less dependent on intensity; light "saturation" is achieved at a specific light intensity, and the rate then is dependent only on temperature if all other factors are constant. In the light-dependent range before saturation, therefore, the rate of photosynthesis is determined by the rates of photochemical steps (*i.e.*, light absorption by chlorophyll, energy transfer between pigments, and the conversion of the energy of excited pigments to chemical energy). At high light intensities, some of the chemical reactions of the dark stage become rate-limiting. At light saturation, rate increases with temperature until a point is reached beyond which no further rate increase can occur.

Carbon dioxide. Included among the rate-limiting steps of photosynthesis are the chemical reactions by which organic compounds are formed using carbon dioxide as a carbon source. The rates of these carboxylation reactions can be increased slightly by increasing the carbon dioxide concentration. An increase in carbon dioxide concentration of 100 times (from 0.03 to 3.0 percent), however, results only in a two to fivefold increase in the rate of photosynthesis in algae such as *Chlorella pyrenoidosa* at light saturation.

Internal factors. Each plant species has become adapted to a range of environmental factors. Complex regulatory mechanisms within a plant's chloroplasts automatically adjust the activities of enzymes (*i.e.*, organic catalysts) involved in important photosynthetic reactions; these adjustments maintain a balance in the overall photosynthetic process. When light intensity, temperature, or some other factor varies beyond the normal environmental range of a plant, therefore, the regulatory mechanisms no longer can compensate adequately, and the overall rate of photosynthesis decreases.

ENERGY EFFICIENCY OF PHOTOSYNTHESIS

The energy efficiency of photosynthesis is defined as the ratio of energy stored (as organic compounds) to the energy of light absorbed. The chemical energy stored during photosynthesis is difficult to calculate precisely because many products are formed, and the amount of energy stored varies with the plant species and environmental conditions. If the equation for glucose formation given above is used to approximate the actual storage process, each mole (*i.e.*, 6.02×10^{23} molecules) of oxygen produced (equal to one-sixth mole of glucose formed) results in the storage of a large quantity (about 117 kilocalories) of chemical energy.

Light is thought to consist of a wave of particles known as photons; these are units of energy, or light quanta. The light energy, or number of photons, in one mole of light is called an einstein. The energy of a light quantum varies inversely with the length of the photon waves; *i.e.*, the shorter the wavelength, the greater the energy content. Blue light has a shorter wavelength and therefore more energy than red light; the same number of einsteins of both blue light and red light, however, are required in photosynthesis per mole of oxygen formed. Because this "quantum requirement" of photosynthesis is about the same for any wavelength of light absorbed, the excess energy must be discarded as heat before conversion of light to chemical energy occurs. In this way, only one mechanism for energy conversion is required in the chloroplast, regardless of the wavelength of light absorbed.

Quantum requirement

After years of study and controversy, the quantum requirement for photosynthesis has been reliably established as about nine einsteins absorbed per mole of oxygen formed. The energy efficiency for glucose production with red light is 31%; this value coincides with the theoretical one calculated from the mechanism of photosynthesis, as the product of the efficiencies of the "light stage" (38%) and the "dark stage" (83%).

The so-called "light stage" includes not only the photochemical steps but also some dark reactions that occur in or on chloroplast membranes. These dark reactions involve breaking the bonds of the water molecule (*i.e.*, the oxidation of water), forming oxygen gas, transferring hydrogen (as electrons and protons), and storing chemical energy in the energy-rich molecule adenosine triphosphate (ATP). The maximum efficiency of the photochemical steps, which have an upper limit imposed by the second law of thermodynamics, is about 75%. The remaining steps of the light stage have an overall efficiency of about 54%.

The optimal efficiency of the light stage of photosynthesis has been measured only under carefully selected laboratory conditions, usually with algae such as *Chlorella pyrenoidosa* or *Scenedesmus*. These aquatic single-celled plants have fewer energy-requiring functions and can be supplied with the most desirable conditions more easily than can higher plants. The energy efficiency of the light stage in higher plants generally is much less than 54% for several reasons: the variation in the amount of light that reaches the leaves, the possibility of low concentrations of carbon dioxide in regions with many growing plants, the energy requirements for synthesizing non-photosynthetic tissues (*e.g.*, stems, roots, fruit), and the other physiological functions that require energy. If agricultural products (*e.g.*, fruit) alone are considered as the end product of the energy conversion process of photosynthesis, the efficiency may decrease to 1% or less.

THE STUDY OF PHOTOSYNTHESIS

Early studies. The study of photosynthesis began in 1771, when it was observed that mint produced some substance (later recognized as oxygen) that supported combustion. The requirement for light was recognized in 1779. In 1782 it was demonstrated that the combustion-supporting gas (oxygen) was formed at the expense of another gas, or "fixed air," which had been identified the year before as carbon dioxide. Gas-exchange experiments

in 1804 showed that the gain in weight of a plant growing in a carefully weighed pot was the sum of carbon, which came entirely from absorbed carbon dioxide, and water taken up by plant roots. Almost half a century passed before the concept of chemical energy developed sufficiently to permit the discovery (in 1845) that light energy from the sun is stored as chemical energy in products formed during photosynthesis.

Recent studies. *Sulfur bacteria.* Comparisons between photosynthetic processes in green plants and in certain photosynthetic sulfur bacteria provided important information about the photosynthetic mechanism. Sulfur bacteria, which use hydrogen sulfide as a source of hydrogen atoms instead of water, produce sulfur instead of oxygen during photosynthesis. In the 1930s, a Dutch biologist, Cornelis van Niel, recognized that the utilization of carbon dioxide to form organic compounds was similar in the two types of photosynthetic organisms. Suggesting that differences existed in the light-dependent stage and in the nature of the compounds used as a source of hydrogen atoms, he proposed that hydrogen was transferred from hydrogen sulfide (in bacteria) or water (in green plants) to an unknown acceptor (called A), which was reduced to AH_2. During the dark reactions, which are similar in both bacteria and green plants, the reduced acceptor (AH_2) reacted with carbon dioxide (CO_2) to form carbohydrate (CH_2O) and to oxidize the unknown acceptor to A. Van Niel's proposal was important because the popular (but incorrect) theory had been that oxygen was removed from carbon dioxide (rather than hydrogen from water) and that carbon then combined with water to form carbohydrate (rather than the hydrogen from water combining with CO_2 to form CH_2O).

Importance of isotopes of oxygen

Isotope studies. By 1940 chemists were using radioactive isotopes of the elements as markers to follow the reactions of photosynthesis. Water marked with an isotope of oxygen (^{18}O) was used in early experiments. Plants that photosynthesized in the presence of water containing $H_2{}^{18}O$ produced oxygen gas containing oxygen-18; those that photosynthesized in the presence of radioactive carbon dioxide and normal water, however, produced normal oxygen gas. These results provided strong support for van Niel's theory that the oxygen gas produced during photosynthesis is derived from water.

Radioactive isotopes of carbon (^{14}C) and phosphorus (^{32}P) have been valuable in identifying the products of photosynthesis. A photosynthesizing plant uses the radioactive carbon (^{14}C) from radioactive carbon dioxide ($^{14}CO_2$) to synthesize complex organic compounds. If green plants photosynthesize in the presence of radioactive carbon dioxide, the compounds formed early in the pathway by which the carbon is reduced, or "fixed" into organic compounds, become labelled with radioactive carbon. The amount of radioactivity in each of the intermediate compounds in the pathway indicates its concentration under conditions of continuous (or steady-state) photosynthesis. If, on the other hand, algae photosynthesizing in the presence of radioactive carbon dioxide are subjected suddenly to some change in environment (*e.g.*, removal of light and, therefore, cessation of the light dependent reactions), changes occur in the concentrations of certain intermediates, compared with their concentrations during steady-state photosynthesis. When light is removed, for example, reactions that require chemical energy originating with light energy stop; reactions that do not require this chemical energy, however, continue. As a result, intermediate compounds that require light energy to undergo further change "pile up" (or increase in concentration), while intermediates that can undergo further change in the dark decrease in concentration. The reactions that require converted light energy were established in this way, as was the step in which carbon dioxide enters the pathway for fixation into organic compounds. This technique also has aided in identification of enzymes that function in regulating the reactions of the carbon dioxide fixation (or reduction) pathway. Radioactive phosphorus has been used to study this pathway because many of the intermediates are sugar phosphates (see below *The carbon reduction cycle*).

Electron flow studies. The earliest stages of light absorption by the chlorophyll pigments have been studied with fluorescence techniques. The ability of the chlorophyll molecule to absorb light resides in its molecular structure. The light energy aborbed by a chlorophyll molecule excites some electrons (negatively charged particles) within the structure of the molecule to higher energy levels. In the case of one type of chlorophyll molecule, called chlorophyll *a*, an excited state is created when the molecule absorbs red light. The return of an excited chlorophyll *a* molecule to its original low energy state (called the ground level) is accompanied by the release of the light energy originally absorbed by the electrons. Part of the energy is released by the re-emission of red light (called fluorescence), and the remainder is converted into chemical energy. Anything that blocks the photosynthetic conversion of light to chemical energy, therefore, causes an increase in fluorescence. If the temperature, for example, is lowered so that enzymatic reactions no longer can occur, fluorescence increases.

When a high-energy electron in an excited chlorophyll molecule leaves chlorophyll during photosynthesis, it is transferred through a series of molecules, called electron carriers, which can gain electrons (be reduced) or lose them (be oxidized). As a result, the chlorophyll molecule returns to the ground state, and the carrier molecule, also called an electron acceptor, is reduced. The electron then is passed successively to other carriers in the electron transport chain. If specific chemical inhibitors are added to green plant tissue to block the transfer of electrons between known pairs of electron carriers, one carrier will be more reduced than the other. Information about the sequence of electron carriers in the electron transport chain has been obtained in this way; in addition, the relationship of each carrier to other carriers and to the site of the photochemical steps has been determined.

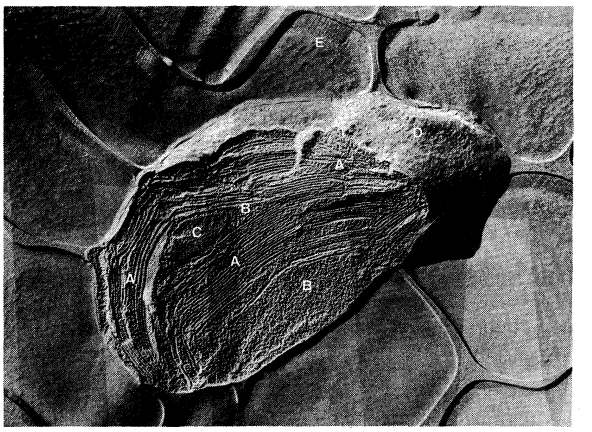
Mutants

Algae whose genetic material has been changed (*i.e.*, algal mutants) have been used to follow electron flow during photosynthesis. Certain algae can be subjected to treatment (*e.g.*, high-energy radiation or specific chemicals) that induce genetic alterations (*i.e.*, mutations). Single mutations in the photosynthetic apparatus may occur. Such mutants, which frequently cannot synthesize one of the key electron carrier molecules of the photosynthetic apparatus, no longer can carry out normal photosynthesis. These algal mutants, which can survive in a growth medium that contains certain necessary organic substances, can be tested for their ability to photosynthesize in the presence and absence of selected electron carrier molecules. Algae are grown, for example, in the presence of light and in a growth medium with or without added plastoquinone, one type of organic molecule involved in electron transfer. If the algae can photosynthesize only if plastoquinone is present, the mutant (assumed to be incapable of synthesizing this molecule) is used to determine whether or not any other electron carrier is more oxidized (*i.e.*, has fewer electrons) in the absence of plastoquinone than in its presence; *i.e.*, algae containing a sequence of carriers lacking plastoquinone molecules cannot transfer electrons past the plastoquinone site in the electron chain. Such carriers presumably would follow plastoquinone in the sequence of carriers in the electron transport chain.

Electron-flow studies have revealed that hydrogen is not transferred directly from water to carbon, nitrogen, and sulfur. Instead, as bonds in the water molecule are broken by the chemical energy generated from light in the photochemical reactions, electrons and hydrogen ions (also called protons) move separately through complex membrane structures called lamellae in the photosynthetic apparatus, or chloroplast. The electrons, which eventually are transferred to electron acceptors (also called coenzymes), are used during the reduction of carbon dioxide to organic products such as glucose. As carbon is reduced, therefore, the protons that result from breaking the bonds of the water molecule are used to balance the chemical reactions involved in the pathway in which organic products are formed.

Chloroplasts. Isolation of the entire photosynthetic apparatus, called the chloroplast, and fractionation of this

cell organelle into its component parts also have provided information about photosynthesis. Around 1940 a British biochemist, Robert Hill, discovered that green particles obtained from broken cells could produce oxygen from water in the presence of light and an electron acceptor compound (*e.g.*, ferric oxalate). This reaction, commonly known as the Hill reaction, provided additional support for the van Niel theory that water provides the oxygen that forms oxygen gas during photosynthesis in green plants. During the 1950s Daniel Arnon and other American biochemists prepared plant cell fragments in which not only the Hill reaction but also the synthesis of the energy storage compound ATP (adenosine triphosphate) occurred. In addition, the coenzyme NADP (nicotine adenine dinucleotide phosphate), the electron acceptor in the carbon dioxide fixation pathway in intact chloroplasts, was used as the final acceptor of electrons, replacing other electron acceptors (*e.g.*, the ferric oxalate used by Hill). Hill's procedures were refined further so that individual small pieces of isolated chloroplast membranes, or lamellae, could perform the Hill reaction. These small pieces of lamellae were then fragmented into pieces so small that they performed only the two light reactions (see below *The light reactions*) of the photosynthetic process. It is now possible also to isolate the entire chloroplast so that it can carry out the complete process of photosynthesis, from light absorption, oxygen formation, and the reduction of carbon dioxide to the formation of glucose and other products.



Electron micrograph of an isolated spinach chloroplast prepared by the freeze-etching method. (A) Lamellae, which are layers of lipid and protein with chlorophyll, other pigments, and enzymes, are the sites of light absorption and splitting of the water to give oxygen and hydrogen. (B) Stroma, containing soluble enzymes, is the site of conversion of carbon dioxide and hydrogen to carbohydrates. (C) Starch granule is a storage product of photosynthesis. (D) Surface of the chloroplast, and (E) frozen solution in which the chloroplast was suspended.
By courtesy of the University of California, Lawrence Radiation Laboratory, Berkeley

CHLOROPLASTS, THE PHOTOSYNTHETIC UNITS OF GREEN PLANTS

General features. Intricate structural organization of the photosynthetic apparatus is essential for the efficient performance of a complex process such as photosynthesis. In addition, regulation of the rates of formation of the various organic products (*e.g.*, amino acids, fatty acids, simple carbohydrates such as glucose) is dependent on organization with the chloroplast.

Structural features

The chloroplast is enclosed in a double membrane, and its size approximates a spheroid about 2.5μ thick and 5μ long (1μ = 0.001 millimetre). Some single-celled algae have one chloroplast that occupies more than half the cell volume. Higher plants (*e.g.*, spinach) contain many chloroplasts, each approximately the size of the one in some algal cells.

When thin sections of a chloroplast are examined under the electron microscope, several separate regions are apparent. The stroma area contains a few membranes and also soluble—*i.e.*, dissolved—enzymes, including those involved in the conversion of carbon dioxide to glucose and other organic compounds. A lamella, which contains all the photosynthetic pigments, is approximately 0.010–0.015μ thick and can extend throughout the chloroplast. Examination of cross sections of lamellae under the electron microscope shows that their edges are joined to form closed hollow discs that are called thylakoids ("sacklike"). It is probable that the spaces within the thylakoids are separate and that the spaces outside are continuous with the stroma regions.

The chloroplasts of some higher plants have another interesting feature. In chloroplasts such as those from spinach, there are certain regions, called grana, in which the thylakoids are very tightly stacked. Since the lamellae contain chlorophyll and all the other pigments involved in photosynthesis, grana regions are highly coloured; when viewed through a light microscope, therefore, grana appear as tiny green bodies within the chloroplasts. The function of grana is not yet known with certainty.

Chemical composition. *Proteins.* Both the structure and the chemical composition of lamellae are complex. Lamellae consist of approximately equal amounts of lipids and proteins. The proteins include structural proteins, enzymes, and protein-containing coenzymes. Enzymes are required as organic catalysts for specific reactions within the lamellae. Structural proteins either bind together or provide a matrix for the other components of the lamellae. Protein coenzymes, also called cofactors, include important electron carrier molecules called cytochromes, which are iron-containing pigments; the pigment portions of cytochromes are attached to protein molecules. During electron transfer, an electron is accepted by an iron atom in the pigment portion of a cytochrome molecule, which thus is reduced; then the electron is transferred to the iron atom in the next cytochrome carrier in the electron transfer chain, thus oxidizing the first cytochrome and reducing the next one in the chain. The cytochrome cofactors of importance in photosynthesis are called cytochrome f, cytochrome b_6, and cytochrome b_3.

In addition to the metal atoms found in the pigment portions of cytochrome molecules, metal atoms also are found in other protein molecules of the lamellae. In proteins with a total molecular weight of 900,000 (based on the weight of hydrogen as one), there are two atoms of manganese, ten atoms of iron, and six atoms of copper. These metal atoms are required for the catalytic activity of some of the enzymes important in photosynthesis. The manganese atoms may be involved in oxygen formation. Both copper- and iron-containing proteins may be involved in electron transport between water and the final electron acceptor molecule of the light stage of photosynthesis, (an iron-containing protein called ferredoxin); this molecule is in part firmly fixed in the lamallae but also occurs as a soluble component in the chloroplasts. The electrons accepted by ferredoxin may be passed to the electron acceptor coenzyme NADP, which is the carrier of electrons required to reduce carbon dioxide in the pathway that culminates in the formation of glucose.

Lipids. About one-fourth of the lipid portion of the lamellae consists of pigments and coenzymes; the remainder consists of various lipids, including phosphate-containing ones called phospholipids. Although the functions of lipids in the photosynthetic process are not yet fully understood, lipids in the lamellar membranes can function as electrical insulating material and permit a charge, or potential difference, to develop across the membrane. The charge is a source of either chemical or electrical energy. The lipids in lamellar membranes also serve as solvents for certain small organic molecules that do not dissolve in water (*i.e.*, hydrophobic molecules) and are necessary for the various photochemical reactions that occur during photosynthesis. Lamellar phospholipid molecules provide stability for the intricate structural arrangement of the lamellae.

Pigments

The most prominent pigment molecules of the lamellae are the chlorophyll molecules. Approximately one-fifth of the lamellar lipids are chlorophyll; one type, chlorophyll *a*, is more abundant than the second type, called chlorophyll *b*. Although chlorophylls are the main light-absorbing molecules in green plants, other pigments such as carotenes and carotenoids (which are responsible for the

yellow-orange colour of carrots) also can absorb light and may supplement or replace chlorophyll as the light-absorbing molecules in some plant cells; the light energy absorbed by these pigments, however, must be passed to chlorophyll before conversion to chemical energy can occur. Various plastoquinone compounds found in the lamellae are involved, as are the cytochromes, in the electron-transport process.

Lamellae. Although the detailed structure of the lamellae is not yet known with certainty, it is known that structural organization is required for their proper function. Intact thylakoids (*i.e.*, lamellar discs) apparently are necessary for the formation of the energy-rich storage compound adenosine triphosphate (ATP). Lamellae that have been broken down to small units can no longer form ATP as the conversion of light into chemical energy occurs during the electron transport process. Lamellae can be disrupted mechanically under carefully controlled conditions into small fragments that absorb light energy and break the bonds of water molecules (*i.e.*, oxidize water) to produce oxygen, as electrons are transferred to ferredoxin, the final electron acceptor of the light stage. Further disruption of mechanically broken lamellae, however, results in complete loss of photosynthetic activity.

Lamellae that are treated with certain detergents can be broken down into two types of small particles. One type can absorb light energy, oxidize water, and produce oxygen, as electrons are accepted by a dye (which thus is reduced) that is added to the particles; these particles use light energy to generate "oxidizing power," which is necessary for the production of oxygen. This type of particle contains one or more of the photochemical units from the lamellae (see light reaction II in Figure 2), which are composed of pigment molecules and enzymes necessary for using light energy to oxidize water and to produce oxygen. A large number of chlorophyll molecules, probably several hundred, in this photochemical unit are arranged around a reaction centre in which the actual conversion of light to chemical energy occurs. Chlorophyll molecules found in this reaction centre can either lose electrons (*i.e.*, be oxidized) or gain electrons (*i.e.*, be reduced) and are known to be associated intimately with the following electron carrying cofactors: cytochrome b, cytochrome f, and plastoquinone. The chlorophyll molecules that do not function at the reaction centre trap light energy that otherwise might not be absorbed. The light energy absorbed by one of these chlorophyll molecules is transferred from one pigment molecule to another until it reaches the reaction centre; there it is converted to chemical energy. In this energy conversion centre a cofactor that has lost an electron during the photochemical step reacts with enzymes and water to produce oxygen. Another cofactor, reduced during the light reaction, transfers an electron to one of a series of electron carriers.

In the presence of electron donors, such as the reduced dye added to the unit above, and ferredoxin (an electron acceptor), the second type of lamellar particle, obtained by treating lamellae with detergents, can absorb light and transfer electrons from the electron donor (*i.e.*, the reduced dye) to ferredoxin; these particles utilize light energy to generate "reducing power," which is necessary for the reduction of carbon dioxide. This type of particle may contain the second type of photochemical reaction centre (see light reaction I, Figure 2), in which a photochemically oxidized cofactor (from the energy conversion centre described above), reduced when it accepts an electron from an electron carrier, is reoxidized as it transfers an electron to ferredoxin, the final electron acceptor of the light stage of photosynthesis. The two types of photochemical reaction centres found in the lamellae of green plants generate sufficient chemical energy from light to carry out all the reactions of photosynthesis, including the transfer of electrons from water through a series of electron carrier molecules to carbon dioxide.

Importance of lamellar particles

THE PROCESS OF PHOTOSYNTHESIS

The role of light energy in electron transfer. The general features of a widely accepted mechanism for photoelectron transfer, in which two light reactions occur during the transfer of electrons from water to carbon dioxide, were proposed by Robert Hill and F. Bendall, in 1960. A scheme for this mechanism is shown in Figure 2. In this figure the vertical scale represents the relative ability (in volts) of various cofactors of the electron-transfer chain to be oxidized or reduced. Molecules with the strongest affinity for electrons (*i.e.*, strong oxidizing agents) are near the bottom of the scale. Molecules that readily give up electrons (*i.e.*, strong reducing agents) are near the top of the scale. The zero point indicates the tendency for hydrogen ions (H^+) to gain electrons (*i.e.*, be reduced) to form hydrogen gas (H_2).

By courtesy of the University of California, Lawrence Radiation Laboratory, Berkeley

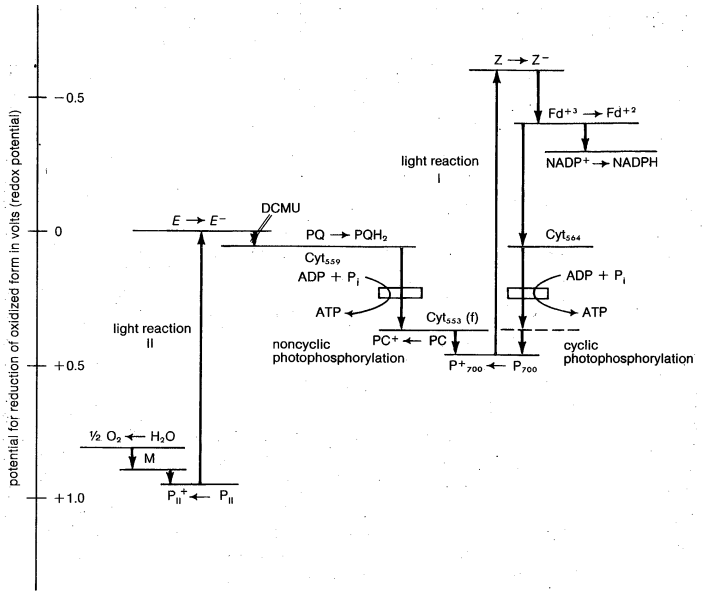


Figure 2: Flow of electrons during the light stage of photosynthesis. Upward pointing arrows represent light reactions that increase the chemical potential; downward pointing arrows represent the flow of electrons in the membrane, in which the release of energy is coupled with the storage of energy in the formation of ATP (see text).

The light reactions. In light reaction II (Figure 2), the energy of an absorbed photon of light is used during the enzyme-catalyzed transfer of an electron from an unknown molecule (designated M) to a carrier (called E). The electron added to carrier E (thus reduced to E^-) is transferred to the next carrier, plastoquinone (PQ), which is reduced to PQH_2. An electron from reduced PQ is transferred through a series of electron carrier molecules to cytochrome f. The energy released during the flow of two electrons from plastoquinone to cytochrome f is used to form at least one molecule of the energy-rich storage compound ATP. In another reaction, electrons are removed from water (thus oxidizing it) and transferred to the unknown molecule M (very easily reduced; *i.e.*, a strong oxidizing agent), thereby reducing M once again; as a result of this reaction, oxygen gas is formed.

In light reaction I (Figure 2) an electron is transferred from a pigment molecule (possibly a molecule of chlorophyll *a*), discovered in 1961 and called P_{700}, to an unknown acceptor molecule (designated Z). P_{700} oxidizes reduced cytochrome f (from light reaction II) and a copper-containing protein molecule (called plastocyanin, or PC). An electron is transferred from the reduced acceptor molecule (Z^-) to the iron-containing protein molecule, ferredoxin (Fd). The reducing power in reduced ferredoxin (Fd^{+3}) either can be used to reduce NADP (the initial step in carbon fixation) or nitrate (to ammonium ion) or sulfate (to sulfide).

Evidence to support the existence of two light reactions. Evidence in support of the above scheme of two light reactions during the light stage of photosynthesis (Figure 2) has been obtained. Using *Chlorella*, the American biochemist Robert Emerson increased the quantum efficiency (*i.e.*, ratio of light absorbed to oxygen formed)

of light with a wavelength longer than .68μ by using it in the presence of light with a shorter wavelength (.58μ). The addition of light of shorter wavelength resulted in oxygen production that exceeded the sum of oxygen produced when each wavelength of light was used alone.

Accessory pigments in algae

In some algae, with coloured molecules (called accessory pigments) in addition to chlorophyll, the effect of light absorbed by accessory pigments is different from that absorbed by chlorophyll *a*. Light absorbed by the accessory pigments in these algae is involved more directly in the oxidation of water and the production of oxygen (reaction II in Figure 2); on the other hand, light absorbed by chlorophyll *a* is used in the transfer of electrons that are used during the reduction of carbon dioxide (reaction I in Figure 2).

When ferredoxin and NADP are added to broken spinach chloroplasts in the presence of light, the following events take place: electrons are transferred from water to NADP, NADP is reduced, and oxygen is formed. None of these events occurs if the plant killer dichlorophenylmethylurea (DCMU; see Figure 2) is present. Since the addition of certain reduced dyes (*i.e.*, electron donors), however, results in the reduction of NADP (light reaction I) without production of oxygen (light reaction II), it is known that DCMU blocks light reaction II. During light reaction I, electrons are transferred from the donor to NADP by way of reduced carrier Z and ferredoxin (Fd).

Quantum requirement. The quantum requirements of the individual light reactions of photosynthesis are based on the theory that one photon of light is absorbed for each electron transferred in each of the two light reactions. The total number of quanta required, therefore, to transfer four electrons that result in the formation of one molecule of oxygen should be four times two, or eight. It appears, however, that additional light is absorbed and used to form ATP by a cyclic photophosphorylation pathway (see next section). The actual quantum requirement, therefore, probably is nine to ten.

The conversion of light energy to ATP: Photophosphorylation. In contrast to electron transfer in light reactions I and II, in which transfer can occur with small particles of chloroplasts or membrane fragments, intact thylakoids are required for photophosphorylation. Enzymes, attached to the outer surface of the thylakoids and shaped like spherical particles, catalyze the final reaction in the formation of ATP from ADP and inorganic phosphate (abbreviated P_i). When these enzymes are removed by chemical agents, electron flow occurs without phosphorylation; photophosphorylation then is said to be uncoupled from electron transfer.

Photophosphorylation in chloroplasts occurs either by a cyclic or a noncyclic process. During noncyclic photophosphorylation, the transfer of each pair of electrons from water to NADP during light reactions I and II (see Figure 2) results in the formation of at least one molecule of ATP. During cyclic photophosphorylation, electrons, which have been transferred to the primary electron acceptor (Z) of light reaction I, are returned through a phosphorylation site to the donor of electrons (P_{700}) in light reaction I. The number of molecules of ATP formed during the transfer of a pair of electrons through this pathway for each light activation step has not yet been established with certainty.

The mechanism by which energy is used to form ATP during electron transfer is not yet known with certainty. One theory is that energy liberated during electron transfer is used to form some energy-rich compound (called X~I), which then reacts with inorganic phosphate (P_i) to form an energy-rich compound, called X~P. An enzyme-catalyzed reaction would result in the formation of ATP and X. An alternative mechanism is that electron flow is accompanied by a flow of hydrogen ions (H^+), also called protons, through the lamellar membrane. In this way, a gradient of protons, which is a form of energy, could be built up across the membrane. Two protons from a specific gradient could provide the energy necessary to form one molecule of ATP. In this mechanism the proton gradient replaces X~I above. The membrane itself could contain the structures and molecules necessary to convert the energy of the proton gradient to either X~P or ATP.

The carbon reduction cycle. The reduction cycle in which carbon dioxide is fixed, reduced, and utilized involves the formation of intermediate compounds, called sugar phosphates; the sequence in which the intermediates are formed has been determined. During the reduction cycle, the energy-rich compounds formed during the light reactions (*i.e.*, ATP and reduced NADP) are utilized.

From the University of California, Lawrence Radiation Laboratory, Berkeley

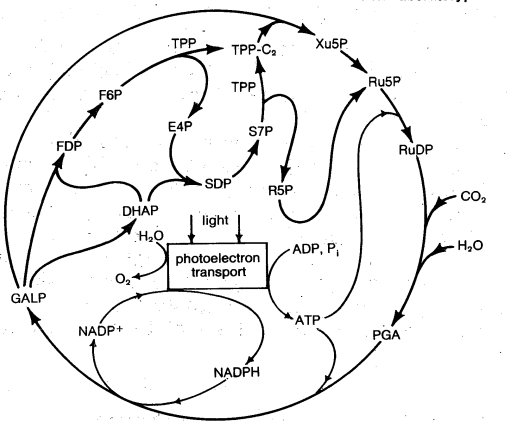


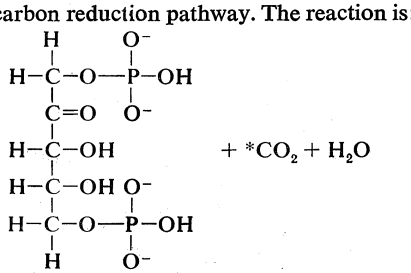
Figure 3: Pathway of carbon reduction in photosynthesis, the reductive pentose phosphate cycle. Thick arrows indicate steps catalyzed by regulated enzymes; the box labelled photoelectron transport represents the reactions shown in Figure 2 (see text).

The fixation of carbon dioxide. Early investigations of the photosynthetic process showed that the first radioactive products formed from carbon dioxide labelled with radioactive carbon included organic acids known as carboxylic acids.

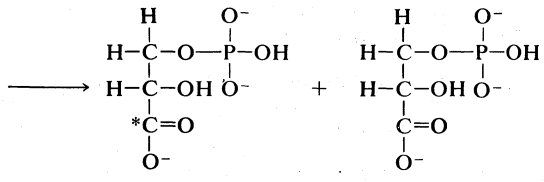
Identification of the first compound formed

Melvin Calvin, a Nobel Prize recipient for his work on the carbon reduction cycle, allowed green plants to photosynthesize in the presence of radioactive carbon dioxide for a few seconds under various experimental conditions. Products labelled with radioactive carbon included a three-carbon compound called 3-phosphoglyceric acid (abbreviated PGA, see Figure 3), sugar phosphates, amino acids, sucrose, and carboxylic acids. When photosynthesis was stopped after two seconds, the principal radioactive compound was PGA, which therefore was identified as the first compound formed during carbon dioxide fixation in green plants. Some plants, however, including many tropical grasses have a different pattern (see below *Carbon fixation in tropical grasses*).

If light is removed from plants photosynthesizing in the presence of radioactive carbon dioxide, the amount of PGA formed increases rapidly; simultaneously, however, the amount of another sugar, called ribulose-1,5,-diphosphate (RuDP), declines rapidly. Energy-rich compounds derived from the light reactions, therefore, are utilized in converting PGA to other products and in forming RuDP. If the carbon dioxide level is decreased suddenly in the presence of light, there is a simultaneous increase in the amount of RuDP and a decrease in PGA formation; in other words, RuDP reacts with radioactive carbon dioxide (indicated below as $*CO_2$) to form PGA; this is the first step in the carbon reduction pathway. The reaction is:



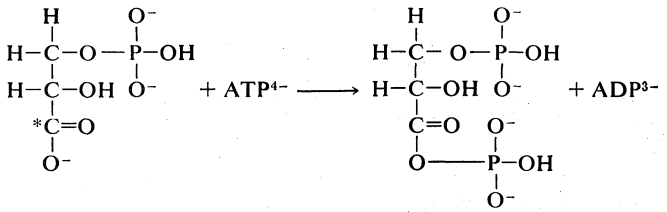
RuDP
(ribulose-1.5-diphosphate)



PGA (labelled 3-phosphoglyceric acid) PGA (3-phosphoglyceric acid)

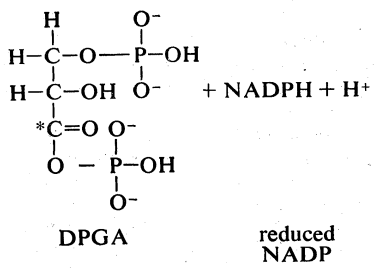
The carboxylation enzyme (*i.e.*, the enzyme that catalyzes the addition of carbon dioxide to RuDP), called ribulose diphosphate carboxylase, has been purified from spinach leaves.

The reduction of carbon. During the formation of ATP in the light stage of photosynthesis, about 13 kilocalories (or one-third of the energy available in the photons contained in one mole—an einstein—of red light) of chemical energy are stored in this energy-rich molecule. An enzyme (3-phosphoglycerate kinase) catalyzes the transfer of the terminal phosphate of an ATP molecule, as well as a large part of the chemical energy stored in it, to the carboxyl carbon (labelled * below) of PGA. The reaction is:

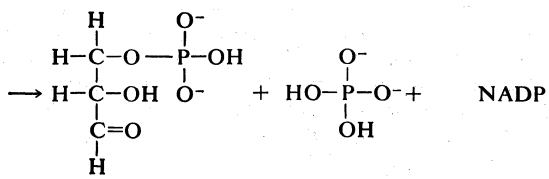


PGA DPGA (diphosphoglyceric acid)

The unstable energy-rich compound formed in this reaction (called phosphoryl-PGA, and abbreviated to DPGA) immediately accepts two electrons from reduced NADP to form a stable three-carbon compound called glyceraldehyde-3-phosphate (GALP; see Figure 3). The reaction (presented below), catalyzed by the enzyme glyceraldehyde phosphate dehydrogenase, is the step in which carbon is reduced.



DPGA reduced NADP

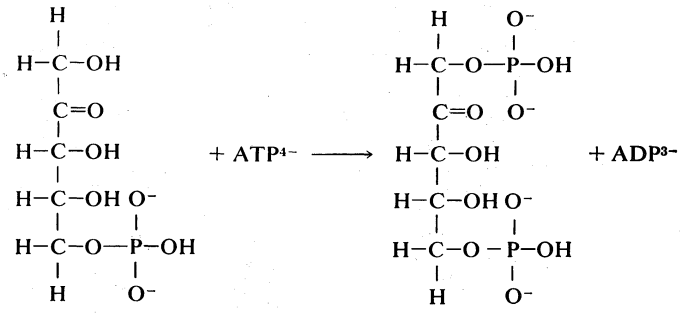


GALP (glyceraldehyde-3-phosphate) inorganic phosphate (oxidized NADP)

Two reactions (*i.e.*, those in which DPGA and GALP are formed) are the reverse of reactions that occur in the Embden-Meyerhof-Parnas (EMP), or glycolytic pathway by which sugars are converted to a three-carbon compound called pyruvate (see METABOLISM).

The reductive pentose phosphate cycle. Many reactions (in addition to the two cited above) of both the EMP pathway, in which oxygen is not utilized, and an oxidative (*i.e.*, oxygen-requiring) five-carbon sugar phosphate cycle known as the HMP shunt, also are involved in the cycle known as the reductive pentose phosphate cycle in photosynthesis. Several of the enzymes (*i.e.*, triose phosphate isomerase, pentose phosphate isomerase, aldolase, transketolase, and ribulose phosphate 3-epimerase) that catalyze the reactions, therefore, also are important in the photosynthetic reductive cycle (shown in Figure 3). Two additional enzymes, fructose diphosphatase and sedoheptulose diphosphatase, however, are important only in the photosynthetic reductive cycle; each removes a phosphate group, from fructose diphosphate (FDP, a six-carbon phosphate-containing sugar) and sedoheptulose diphosphate (SDP, a seven-carbon phosphate-containing sugar), respectively. The net result of the reactions catalyzed by the enzymes in the photosynthetic reductive cycle is the conversion of five molecules of the three-carbon sugar phosphate, glyceraldehyde-3-phosphate (GALP), to three molecules of the five-carbon sugar phosphate, ribulose-5- phosphate (Ru5P; see Figure 3).

ATP formed during the light stage also is used to convert Ru5P to RuDP, the compound involved in carbon dioxide fixation. The reaction, catalyzed by the enzyme ribulose phosphate kinase, is:

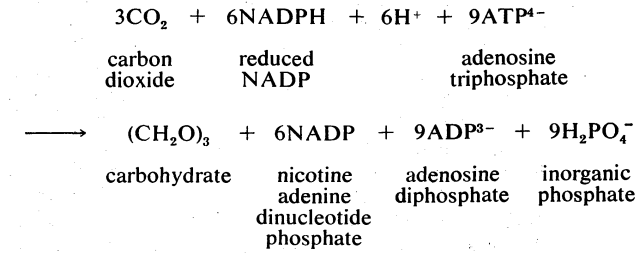


Ru5P ribulose-5-phosphate RuDP (ribulose-1,5-diphosphate)

Utilization of the products of the cycle. During a complete carbon reduction cycle, three molecules of carbon dioxide are added to three molecules of ribulose-1,5-diphosphate (RuDP); *i.e.*, RuDP is carboxylated, and six molecules of phosphoglyceric acid (PGA) are formed. If the carbon reduction cycle is to be self-sustaining (*i.e.*, require no carbon source other than carbon dioxide), five PGA molecules must be reduced to five triose phosphate molecules (*i.e.*, glyceraldehyde-3-phosphate, GALP); these molecules then must be converted to three molecules of RuDP, which are required to continue the carboxylation reaction, in which carbon dioxide is fixed. The sixth molecule of PGA is used in the synthesis of other products; *e.g.*, carbohydrates such as starch (see below *Regulation of the cycle*).

Summary of the energy requirements of the cycle

During the carbon reduction cycle, six molecules of PGA are reduced; this reduction requires the converted light energy contained in six molecules of ATP and six molecules of reduced NADP. The phosphorylation (*i.e.* addition of phosphate) of three molecules of ribulose-5-phosphate requires an additional three molecules of energy-rich ATP. The overall reaction of the cycle can be represented as:



carbon dioxide; reduced NADP; adenosine triphosphate

carbohydrate; nicotine adenine dinucleotide phosphate; adenosine diphosphate; inorganic phosphate

The reduction of one molecule of carbon dioxide to a sugar, therefore, requires the energy contained in three molecules of ATP and two molecules of reduced NADP.

Regulation of the cycle. For every three molecules of carbon dioxide fixed in the carboxylation reaction and then reduced, one three-carbon compound (*e.g.*, phos-

phoglyceric acid, glyceraldehyde-3-phosphate) can be utilized in the synthesis of products such as carbohydrates, fats, and proteins. Three-carbon compounds usually are utilized in the synthesis of fats and proteins; these compounds include phosphoglyceric acid and the triose phosphates (that is, glyceraldehyde-3-phosphate, GALP, and a derivative, dihydroxyacetone phosphate, DHAP; see Figure 3).

During the synthesis of carbohydrates, however, two triose phosphate molecules must combine to form a six-carbon compound called fructose diphosphate (FDP); this molecule then loses one phosphate group and forms fructose-6-phosphate (F6P; Figure 3). The fructose diphosphatase enzyme that catalyzes this reaction is called a regulatory enzyme because its activity can be increased or decreased by certain factors; for example, the activity of the enzyme is regulated not only by the presence or absence of light but also during photosynthesis in light. This enzyme controls the rate of the reaction in which three-carbon compounds that form fats and proteins are changed into the six-carbon compound required to form carbohydrates (*e.g.*, sucrose, starch, cellulose). Fructose diphosphatase, therefore, is a very important enzyme in regulating the fate of the sixth sugar phosphate molecule, formed during the photosynthetic cycle, as it enters the biosynthetic pathways of metabolism. For additional information about regulation of metabolic pathways, see METABOLISM.

Another important regulatory enzyme, ribulose diphosphate carboxylase, controls the rate of carbon dioxide fixation. By limiting the rate of fixation and thus the amount of carbon dioxide reduced, this enzyme provides a mechanism by which ATP and reduced NADP, which otherwise would be used in the carbon reduction cycle, can be used in the biosynthetic pathways that result in the formation of carbohydrates, proteins, and fats. By means of the regulatory effects of this enzyme, the reduction of carbon dioxide is balanced with the conversion of reduced carbon to organic products.

A third regulatory mechanism involves four enzymes of the carbon reduction cycle: ribulose diphosphate carboxylase, phosphoribulokinase, fructose diphosphatase, and sedoheptulose diphosphatase; these enzymes are not active in the absence of light. The dependence of these enzymes on light for activity is part of a regulatory mechanism by which the chloroplasts can change from photosynthetic to oxidative metabolism at night.

Carbon dioxide fixation in tropical grasses. When some tropical grasses (and certain other plants) photosynthesize for a few seconds in the presence of carbon dioxide labelled with radioactive carbon, the important radioactive compounds formed are malate (a carboxylic acid) and aspartate (an amino acid). The plants contain an enzyme that catalyzes the formation of phosphoenolpyruvate (PEPA) from pyruvate, an important intermediate in the EMP or glycolytic pathway. A mechanism has been proposed for the incorporation of carbon dioxide in these photosynthetic plants. It is postulated that phosphate first is added to pyruvate to form PEPA; carbon dioxide is incorporated into PEPA (*i.e.*, PEPA is carboxylated) to form a four-carbon acid called oxaloacetic acid (OAA). Oxaloacetic acid can be converted either into aspartate or malate. By some mechanism not yet known with certainty, the carbon atom incorporated into the four-carbon acids may be transferred to some acceptor (perhaps ribulose-1, 5-diphosphate); 3-phosphoglyceric acid eventually is formed and reduced by way of the carbon reduction cycle.

An explanation for this unusual pathway may be that the plants store carbon, and perhaps transport it from one kind of cell to another, as malic acid. Chloroplasts in cells around vesicles of the leaf called parenchyma cells contain enzymes that catalyze the reactions of the reduction cycle. Chloroplasts in a second type of green cells called mesophyll cells apparently contain the enzymes that catalyze the formation of PEPA and its carboxylation to oxaloacetic acid. The tropical grasses and other plants that utilize this pathway may have evolved under conditions of limited water supply and high light intensity; in this case, carbon dioxide would have been the limiting factor in photosynthesis (see above *Factors that influence the rate of photosynthesis*). It is possible that, in such circumstances, a special mechanism for the fixation of carbon dioxide might represent an evolutionary advantage, even though some chemical energy is sacrificed. The basic process of photosynthesis, therefore, may have been adapted or altered during the evolution of the process to allow certain plant species to survive under available ecological conditions.

BIBLIOGRAPHY. J.A. BASSHAM and M. CALVIN, *The Path of Carbon in Photosynthesis* (1957), an account of the mapping of the path of carbon in photosynthesis using radiocarbon; R.K. CLAYTON, *Molecular Physics in Photosynthesis* (1965), an excellent introduction to photosynthesis, especially biophysical problems involved in energy absorption and conversion; G.E. FOGG, *Photosynthesis* (1968), a textbook-level exposition of the principles of photosynthesis, with emphasis on biochemical aspects; E.I. RABINOWITCH and GOVINDJEE, *Photosynthesis,* (1969) a textbook-level exposition of the principles of photosynthesis with the emphasis on photochemical and photophysical processes.

(J.A.B.)

Phthiraptera

Lice (order Phthiraptera) are small, wingless, parasitic insects divisible into two main groups: the Mallophaga, or chewing or biting lice, which are parasites of birds and mammals, and the Anoplura, or sucking lice, parasites of mammals only. One of the biting lice, the human louse, is the carrier of typhus and louse-borne relapsing fever; it thrives in conditions of filth and overcrowding. Outbreaks of louse-borne diseases were frequent by-products of famine, war, and other disasters before the advent of powerful insecticides (see INFECTIOUS DISEASES). Heavy infestations of lice may cause intense skin irritation, and scratching for relief may lead to secondary infections. In domestic animals rubbing and damage to hides and wool may also occur, and meat and egg production may be affected; in badly infested birds the feathers may be severely damaged. One of the dog lice is the intermediate host of the dog tapeworm, and a rat louse is a transmitter of murine typhus among rats.

General features. The flattened bodies of lice range from ⅓ millimetre to 11 millimetres in length and are whitish, yellow, brown, or black. Probably all species of birds have chewing lice, and most mammals have either chewing or sucking lice, or both. There are about 2,900 known species of Mallophaga, with many others still undescribed, and about 400 species of Anoplura. No lice have been taken from the duckbilled platypus or from anteaters and armadillos; and none are known from bats or whales. The size of louse populations varies enormously on different individuals, sometimes seasonally. Sick animals and especially birds with damaged bills, probably because of the absence of grooming and preening, may have abnormally large numbers: over 14,000 on a sick fox and over 7,000 on a cormorant with a damaged bill; the numbers found on healthy hosts are usually considerably smaller. Apart from grooming and preening by the host, lice and their eggs may be controlled by predatory mites, dust baths, intense sunlight, and continuous wetting.

Natural history. *Life cycle.* With the exception of the human body louse, lice spend their entire life cycle, from egg to adult, on the host. The females are usually larger than the males and often outnumber them on any one host; in some species males are rarely found, and reproduction is by unfertilized eggs (parthenogenetic). The eggs are laid singly or in clumps, usually cemented to a feather or hair; the human body louse lays its eggs on clothing next to the skin. The eggs may be simple ovoid structures glistening white among the feathers or hairs or may be heavily sculptured or ornamented with projections that assist in the attachment of the egg or serve in gas exchange. When the nymph within the egg is ready to hatch, it sucks in air through its mouth; this passes down the alimentary canal and accumulates behind the nymph until sufficient pressure is built up to force off the cap (operculum), helped by an armed, platelike structure, the hatching organ, at the upper end of the prelarval skin. The emergent nymph is similar to the adult but is

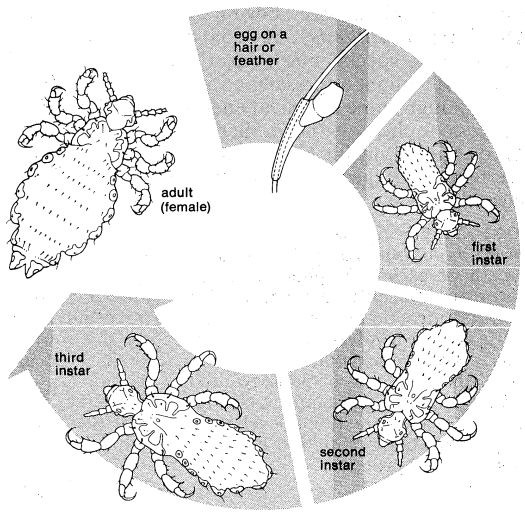


Figure 1: Life cycle of a typical louse.

smaller and uncoloured, has fewer hairs, and differs in certain other morphological details.

Metamorphosis in the lice is simple, the nymphs molting three times, each of the three stages between molts (instars) becoming larger and more like the adult (see Figure 1). The duration of the different stages of development varies from species to species and within each species according to temperature. In the human louse the egg stage may last from six to 14 days and the stages from hatching to adult, eight to 16 days. The life cycle may be closely correlated with the particular habits of the host; *e.g.*, the louse of the elephant seal must complete its life cycle during the three to five weeks, twice a year, that the elephant seal spends on shore.

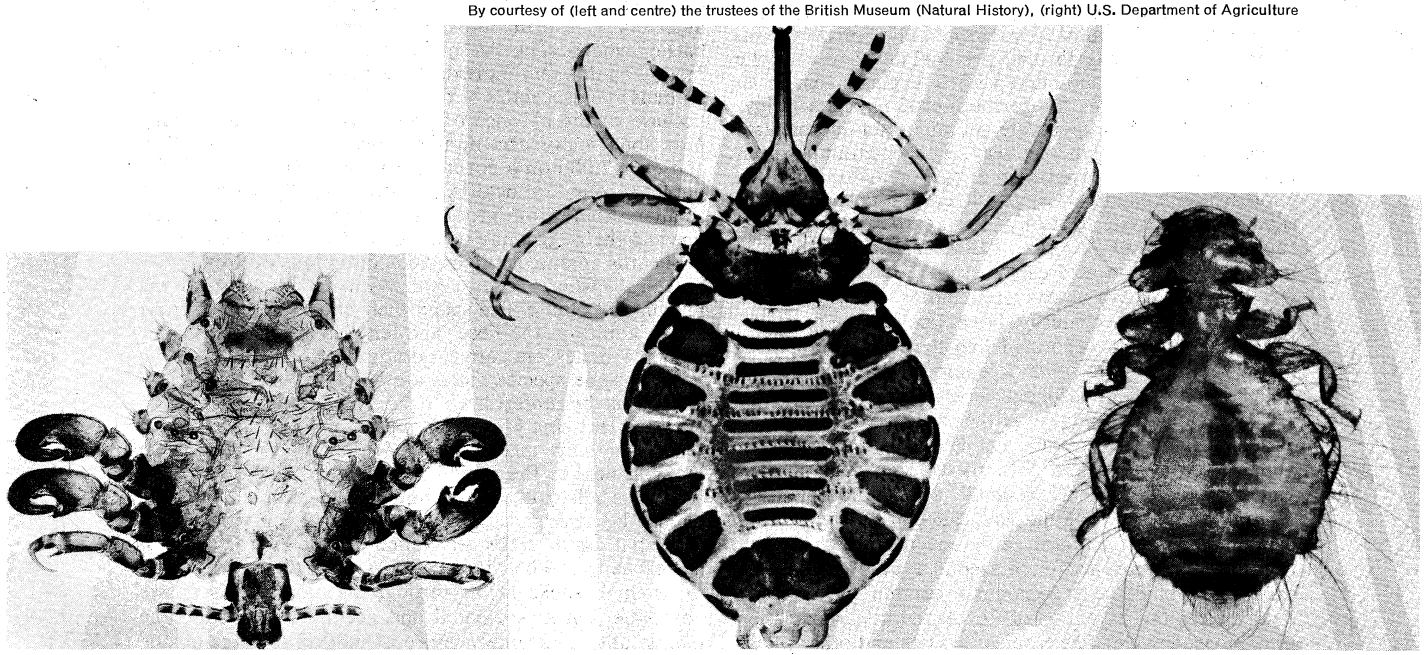
Ecology. Sucking lice live exclusively on blood, their mouthparts being well adapted for this purpose. The delicate stylets, retracted into the head when the louse is not feeding, are used to pierce the skin, and a salivary secretion is injected to prevent coagulation while the blood is sucked into the mouth. The chewing lice of birds feed on the feathers, or on feathers, blood, and tissue fluids, or on fluids only. The fluids are obtained either by gnawing the skin or, as in the poultry body louse, from the central pulp of a developing feather. The chewing lice of hummingbirds have delicate mandibles that can pierce the skin of the host and a modified eversible "throat" (hypopharynx). The feather-eating Mallophaga are able to digest the keratin of feathers. It is probable that the chewing lice of mammals do not feed on wool or hairs but on skin debris, secretions, and perhaps sometimes blood and tissue fluids.

Many birds and mammals are infested by more than one species of lice, most of the birds having at least four or five; these species may inhabit different parts of the body, with adaptations to the particular habitat. Among the avian chewing lice, some species occupy different regions for resting, feeding, and egg laying. A louse is unable to live for more than short periods away from its host, and adaptations serve to maintain its close contact: being attracted by body heat and repelled by light, the louse stays within the warmth and darkness of the host's plumage or pelage. It is also probably sensitive to the smell of its host and the peculiarities of feathers and hairs that help the louse orient itself. A louse may leave its host temporarily to pass to another host of the same species or to a host of another species, such as from prey to predator. Chewing lice have often been found attached to louse flies (Hippoboscidae), also parasitic on birds and mammals, and on other insects by which they may be transferred to a new host. They may not be able to establish themselves on the new host, however, perhaps because of chemical or physical incompatibility with the host as food or habitat; some mammalian lice, for example, can lay their eggs only on hairs of a suitable diameter.

The infrequency of transfer from one host species to another leads to host specificity, or host restriction, in which a species of louse is found only on one species of host or a group of closely related host species. It is probable that some host-specific species have developed through isolation because there is simply no opportunity for the transfer of lice. Domestic and zoo animals sometimes have established populations of lice from different hosts, and pheasants and partridges often have flourishing populations of chicken lice. *Heterodoxus spiniger*, which is parasitic on domestic dogs in tropical regions, was most likely acquired relatively recently from an Australian marsupial.

Form and function. The louse body is flattened dorsoventrally with the long axis of the head horizontal, enabling it to lie close along the feathers or hairs for attachment or feeding. The shape of the head and body varies considerably, especially in the avian chewing lice,

By courtesy of (left and centre) the trustees of the British Museum (Natural History), (right) U.S. Department of Agriculture



Body types of lice.
(Left) Pubic or crab louse *(Phthirus pubis)*, a sucking louse; (centre) elephant louse *(Haematomyzus elephantis)*, a louse intermediate to the Anoplura and Mallophaga; and (right) pigeon biting louse *(Menacanthus latus)*, a chewing louse (all magnified about 30×).

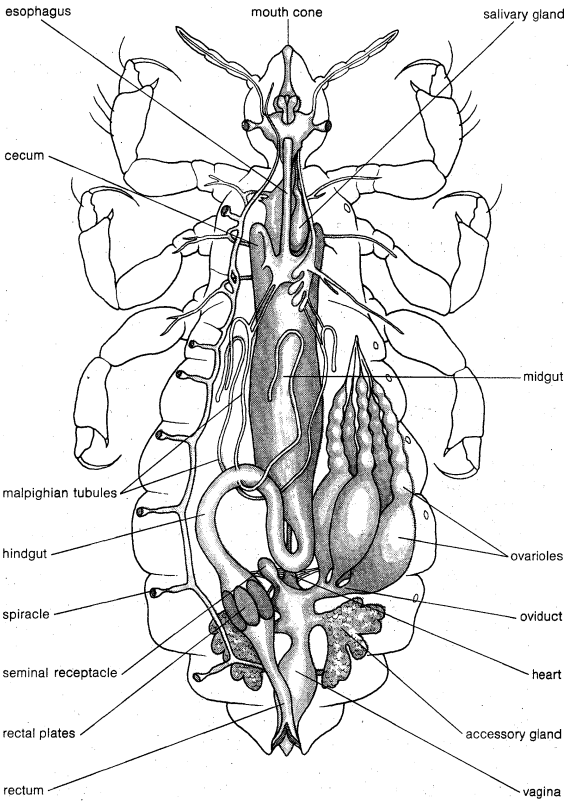


Figure 2: Internal anatomy of a typical female louse
From H. Weber, *Grundriss der Insektenkunde* (1967); Gustav Fischer Verlag, Stuttgart

in adaptation to the different ecological niches on the body of the host. Birds with white plumage, such as swans, have a white body louse, while the dark-plumaged coot has an almost black body louse. The antennae are short, three- to five-segmented, sometimes modified in the male as clasping organs to hold the female during copulation; the eyes are compound; the mouthparts are biting (mandibulate) in the Mallophaga, strongly modified for sucking in the Anoplura. The Anoplura have three stylets enclosed in a sheath within the head, and a small proboscis armed with recurved toothlike processes, probably for holding the skin during feeding. The elephant louse, morphologically intermediate between the Mallophaga and Anoplura, has chewing mouthparts, with the modified mandibles borne on the end of a long proboscis. The thorax may have three visible segments, the mesothorax and metathorax fused, or all three fused into a single segment as in the Anoplura. The legs are well developed, the tarsus one- to two-segmented, with two claws in the avian Mallophaga and a single claw in some of the mammal-infesting families; the Anoplura have a single claw opposed to a tibial process forming a hair-clasping organ. The abdomen has eight to ten visible segments. There is one pair of thoracic breathing pores (spiracles) and a maximum of six abdominal pairs. The eversible male genitalia provide important characters for the classification of species. The female has no well-defined ovipositor, but various lobes present on the last two segments of some species may act as guides to the eggs during laying. The alimentary canal in the Mallophaga is composed of the esophagus, well-developed crop and midgut, a smaller hindgut, four malpighian tubules, and a rectum with six papillae. The crop is either a simple swelling between esophagus and midgut or a diverticulum from the esophagus. In the Anoplura the esophagus passes straight into the large midgut with or without a swelling forming a crop; there is also a strong pump for sucking up the blood. Members of the superfamily Amblycera have well-developed, comblike structures at the base of the crop, which prevent undigested feather parts or other particles from passing into the midgut; in the family Philopteridae these combs are smaller and lie at the anterior part of the crop; the Trichodectidae and Anoplura have no crop teeth. Apart from the eyes, which are sensitive to light, the other sensory structures are the tactile hairs and the sense organs in the mouth and on the antenna, some of which function as taste and smell organs.

Evolution and paleontology. It is generally accepted that the lice are derived from the book lice (order Psocoptera). It is also accepted that the Anoplura are related to the Mallophaga, some authorities believing that they evolved from an ancestral stock before the division into the Amblycera and Ischnocera, others that they diverged from those Ischnocera already parasitic on mammals. The origins of the elephant louse are obscure.

Apart from a louse egg found in Baltic amber, there are no fossils that might provide information on the evolution of the lice; however, their host distribution is in some ways analogous to a fossil history. Mallophagan genera frequently have a number of species that are restricted to one species of bird or to a group of closely related birds, suggesting that the stock ancestral to the bird order was parasitized by an ancestral mallophagan stock that diverged and evolved along with the divergence and evolution of its bird hosts. This relationship between host and parasite may throw some light on the relationships of the hosts themselves. The flamingos, which are usually placed with the storks, are parasitized by three genera of Mallophaga found elsewhere only on ducks, geese, and swans and may therefore be more closely related to those birds than to storks. The louse most nearly related to the human body louse, is that of the chimpanzee, and to the human pubic louse that of the gorilla. However, a number of factors have obscured the direct relationship between louse species and host species, the most important being secondary infestation, which is the establishment of a louse species on a new and unrelated host. This may have happened at any stage during the evolution of host or parasite so that subsequent divergence will have obscured all traces of the original change of host.

Classification. *Distinguishing taxonomic features.* The important characters used in classifying lice at the subordinal level are mainly based on the mouthparts. Features separating the lower categories are the special modifications of mouthparts, crop, antennae, sutures, and internal thickening of the head capsule; the number and form of claws; the segmentation of thorax and abdomen; the form of body plates; the number of spiracles; the pattern of bristles (or setae); and features of the male genitalia and terminal segments of the abdomen.

Annotated classification. Lice can be included in one order, the Phthiraptera, being separated by the characters of the mouthparts into three suborders: Mallophaga, Rhynchophthirina, and Anoplura.

ORDER PHTHIRAPTERA

Small dorso-ventrally flattened parasitic insects. Eyes reduced or absent, ocelli absent, antenna three- to five-segmented, mouthparts mandibulate or piercing. Obligate permanent ectoparasites of birds and mammals.

Suborder Mallophaga (chewing or biting lice)

Mandibulate mouthparts. Parasites of birds and mammals. About 2,900 species.

Superfamily Amblycera

Antenna four- to five-segmented, third segment pedunculate; articulation of mandibles horizontal; two- to five-segmented maxillary palpus; crop simple.

Family Menoponidae. Widespread parasites of birds, contained in about 60 genera.

Family Boopidae. Confined to marsupials of Australasia, except for one species found on domestic dogs.

Family Laemobothriidae. Contains some of the largest Mallophaga up to 11 millimetres in length. Parasites of birds of prey, rails, and some storks.

Family Ricinidae. Parasites of passerines and hummingbirds.

Family Trimenoponidae. Parasites of New World marsupials and rodents.

Family Gyropidae. New World, parasitic mainly on rodents, with one species on primates and one on peccaries.

Superfamily Ischnocera

Third antennal segment filiform; articulations of mandibles vertical; maxillary palpus absent; crop as diverticulum of esophagus; parasites of birds and mammals.

Family Philopteridae. Parasites of birds, except for one genus (*Trichophilopterus*) on lemurs; contains approximately 130 genera recorded from all orders of birds, except the swifts and hummingbirds.

Family Trichodectidae. Approximately 14 genera parasitic on mammals, rarely on primates, sloths, and rodents; more widespread on land carnivores, hyraxes, horses, donkeys, and artiodactyls, except pigs.

Suborder Rhynchophthirina

Modified mandibles borne at end of long proboscis; filiform five-segmented antennae; meso- and metanotum fused; thoracic spiracle ventral; single tarsal claw; crop absent; one genus with two species: *Haematomyzus elephantis* on the African and Indian elephants and *H. hopkinsi* on the African wart hog.

Suborder Anoplura (sucking lice)

Piercing mouthparts in the form of three fine eversible stylets; filiform four- to five-segmented antennae; all three segments of thorax fused together; thoracic spiracle dorsal; single tarsal claw, at least on second and third legs; crop absent, or if present is a simple enlargement; about 400 species.

Family Echinophthiriidae (seal lice). Parasitic on seals.

Family Haematopinidae (wrinkled sucking lice). Includes two genera parasitic on pigs, cattle, deer, and horses.

Family Hoplopleuridae (small mammal-sucking lice). A large family containing approximately 27 genera parasitic mainly on rodents but also on insectivores, primates, and one on an ungulate.

Family Linognathidae (smooth sucking lice). Parasitic on artiodactyls and hyraxes, except for two species parasitic on carnivores.

Family Neolinognathidae. Two species parasitic on insectivores.

Family Pediculidae (human lice). Two genera parasitic on man, the great apes, and on some of the New World monkeys, the last perhaps being secondarily acquired from man.

Critical appraisal. The classification given above is a currently accepted one. Other classifications rank the Anoplura and Mallophaga as separate orders, with the elephant louse placed in one or the other according to personal opinion. Some workers consider the Phthiraptera as comprising four suborders: the Amblycera and Ischnocera (both now included in the Mallophaga), Anoplura and Rhynchophthirina, dropping the name Mallophaga entirely. This represents the view that the Ischnocera, Anoplura, and Rhynchophthirina are more closely related to each other than are any of them to the Amblycera. There are considerable differences in the number and extent of the mallophagan families recognized and also on the generic limits within the families. In the Anoplura, controversy continues on the relations of the genera to each other and in which families they should be included.

The lice of man are referred to by various names, depending on whether the head louse is considered as a distinct species or as a variety or subspecies of the body louse. At present they are probably best referred to under one name, *Pediculus humanus,* but if separated subspecifically they must be called *Pediculus humanus humanus* (the body louse) and *Pediculus h. capitis* (the head louse).

BIBLIOGRAPHY. G.F. FERRIS, "The Sucking Lice," *Mem. Pacif. Cst. Ent. Soc.*, vol. 1 (1951), an outline of the morphology, growth, and identification of the Anoplura, with a host-parasite list; G.H.E. HOPKINS, "The Host-associations of the Lice of Mammals," *Proc. Zool. Soc. Lond.*, 119:387–604 (1949), an account of the biology and methods of preservation, examination, and distribution of the mammalian lice, with a list of mammals and their lice; and with T. CLAY and G. TIMMERMANN, *First Symposium on Host Specificity among Parasites of Vertebrates* (1957), papers on the distribution of the Anoplura and Mallophaga of birds and mammals, with a discussion on host-parasite relationships; S. VAN KELER, "Bibliographie der Mallophagen," *Mitt. Zool. Mus.* 36:147–403 (1960), a full bibliography of the literature on the Mallophaga arranged chronologically, together with an author and subject index; M. ROTHSCHILD and T. CLAY, *Fleas, Flukes and Cuckoos: A Study of Bird Parasites,* 3rd ed. (1957), a popular account of the Mallophaga of birds; H. ZINSSER, *Rats, Lice and History* (1935, reprinted 1965), a readable account of the scientific and historical aspects of *Typhus.*

(T.C.)

Phylogeny

Phylogeny may be defined as the history of the evolution of a species or group. While the term can be applied to a particular species or limited group, it is commonly used to refer to lines of descent and relationships within the broader groups of plants or animals.

Phylogeny is, thus, clearly an evolutionary concept, for it is based upon the proposition that organisms of different species and higher groups are, in fact, related by descent from common ancestors. Unfortunately, however, judgments of relationships among organisms are almost always based upon incomplete evidence, for the plants and animals of today are merely the current stage in a continuing process of diversification that has been in progress for more than 3,000,000,000 years. The overwhelming majority of species that have ever lived have long since been extinct and with them the connecting links necessary for the direct demonstration of the descent of modern organisms from common ancestors. Phylogenetic judgments, therefore, are based largely upon indirect evidence and cautious speculation. Different biologists, using much the same data, may arrive at wholly different and sometimes conflicting phylogenies. Biologists are confident that the world of life is the product of organic descent from earlier ancestors and that the true relationships, in principle at least, are discoverable. The various systems of phylogeny are simply each author's best approximations of that true but unattainable phylogeny.

The speculative nature of phylogeny

Phylogeny is the foundation for taxonomy, the science of the classification of organisms. The necessity for classification into species, genera, families, and higher groups derives from the enormous number and variety of species and man's desire to find some convenient and orderly way of grouping them. Were it not possible to classify them into larger groups concerning which generalizations can be made, study of the world of life would be severely limited. The early systems of classification were purely empirical attempts to group organisms according to degree of similarity, without any theoretical basis. Since the publication of the *Origin of Species* by Charles Darwin in 1859, organic descent and relationship have been accepted as the basis for the taxonomic system, and a good taxonomy is one that expresses these relationships best (see CLASSIFICATION, BIOLOGICAL).

THE BASIS OF PHYLOGENY

In principle, every aspect of biology is a product of phylogeny and may shed some light upon phylogenetic relationships, but, in practice, some kinds of data are more useful than others. The more valuable kinds of evidence derive from paleontology, comparative anatomy, comparative embryology, and biochemistry; geographical distribution, study of ultrastructure (electron-microscopic structure of cells), and computer analysis of data may also be useful.

Fossils, the preserved remains of the life of the past, are usually contained in sedimentary rocks, with ancient fossils in the deeper layers and more recent fossils in the superficial layers. Most commonly, only the hard parts of the body (wood, skeleton, and teeth, for example) are preserved; soft parts are generally lost. As a result, the fossil record is often very helpful in determining the phylogeny of groups with extensive hard parts, such as corals, mollusks, starfishes, and vertebrates, but is generally of little help in the study of the numerous soft-bodied groups.

A closely graded series of fossils leading from an ancestral to a descendant group is conclusive evidence of relationship. One of the best known examples of such a series is that of the horses, the fossil record of which begins nearly 60,000,000 years ago with the four-toed *Hyracotherium* (or eohippus), less than a foot high at the shoulder and having low-crowned teeth adapted to browsing on

tender leaves and shoots. The series proceeds by moderate steps to the modern horses, *Equus*, which may stand five feet tall at the shoulder, have only a single toe, and have high-crowned teeth adapted to grazing on hard, gritty grasses. The evidence of direct relationship is clear. (The horse phylogeny also includes several lines of descent that became extinct.)

The importance of comparative studies

Comparative anatomy has been used more extensively than any other source of data for phylogenetic studies. A comparison of the skeletons of the forelimbs of a mole, a bat, and a seal shows that, in spite of their radically different uses, these limbs are all built on the same plan, with variations in size and proportions suited to the life habits of the animals. In each, there is a single, relatively long bone, the humerus ("upper arm"), which joins the shoulder at one end and connects at the other end with two parallel bones, the radius and ulna ("forearm"); next is a cluster of small bones, the carpals ("wrist bones"), which in turn connect with the metacarpals ("palm bones"), and these finally attach to the phalanges ("finger bones").

As it is difficult to believe that such different functions (burrowing, flying, and swimming) would coincidentally be subserved by such similar structures, the inference is strong that these animals are related by descent from a common ancestor. Such structures, in which the same materials have been used in much the same ways, are said to be homologous, and homology is considered to be strong evidence of relationship.

Comparative embryology is essentially the comparative anatomy of development, and it provides similar phylogenetic evidence. The embryos of all terrestrial vertebrates, for example, develop a typical, fishlike set of gill pouches, only to refashion them into such unlike structures as the ear tubes, various glands of the neck region, and the lungs. While this seemingly roundabout pattern is inexplicable on any other grounds, it follows quite logically if it is assumed that the higher vertebrates are descendants of ancient fishes.

Biochemical evidence of relationship originally dealt with such things as enzymes or hormones held in common by organisms of a series. Currently, however, detailed comparisons of the structures of proteins and of deoxyribonucleic acid (DNA, which carries hereditary information) are possible, and these are expected to contribute valuable phylogenetic data.

THE EARLY STAGES OF PHYLOGENY

The earliest stages of phylogeny are intertwined with the origin of life and its early diversification (see LIFE).

Chemical evolution. Fossil remnants of the origin of life are not and presumably never can be known, but much experimentation indicates that the earliest living things were probably products of a long chemical evolution, based upon random reactions of the chemicals of the crust of the earth and of the early atmosphere. As a result, the primeval seas must have become a thin "soup" of many compounds, including some of those characteristic of living organisms. Some of these probably formed complex aggregates, or protein-containing droplets, which, by binding molecules to their surface, could form a membrane, thus permitting the beginnings of individuality. If such a protein droplet with a limiting membrane were to produce more of its own substance—that is, to reproduce—it could rightly be called an elementary organism. In present-day organisms such a process is associated principally with the class of complex molecules called nucleic acids. For this reason, many investigators believe that the beginnings of life could not have preceded the association of nucleic acids (especially DNA) with proteins; others believe that this stage may have been preceded by one based upon self-reproducing proteins alone.

A strand of DNA with a protein sheath would be essentially a virus. Although all known viruses are parasitic, it is conceivable that a free-living virus system could have existed at a time when the waters of the world were rich in many organic compounds that are now found almost exclusively within organisms. As larger self-reproducing units were formed, mutations, or chance variations in DNA, could produce genetically still more complex units, whose increasing numbers would rapidly deplete the primitive stockpile of nutrients in the surrounding water. As each nutrient became scarce, natural selection would favour any units that had become able, by a mutation, to produce the scarce nutrient. Thus, an increasingly complex system of enzymes that could synthesize nutrients would be built up, and with it increasing cell size and structural complexity. The earliest organisms were heterotrophic—that is, they depended upon outside sources of nutrition, primarily by absorbing organic matter from the primitive seas—but later organisms acquired the ability to make their own food (*i.e.*, became autotrophic). These organisms were the earliest bacteria, complete cells with many enzymatic pathways of intermediary metabolism and the capability to store energy as food and release energy when needed by the burning of that food (respiration). The bacteria evolved several nutritional and respiratory systems. Many became heterotrophic, obtaining their nutrients from other organisms, usually by parasitism; others became saprophytic, absorbing dissolved organic matter from decomposing organisms; and still others became autotrophic, synthesizing some or all of their nutritive requirements. They evolved the ability to use several different energy sources, based upon the oxidation of compounds of sulfur, nitrogen, or iron (chemotrophic nutrition). Some of the "sulfur-eating" bacteria contain the green pigment chlorophyll, which absorbs sunlight and provides energy. Bacterial chlorophyll is chemically related to the chlorophyll of higher plants, and it is possible, but by no means certain, that it is genetically related. It differs markedly in that the bacterial chlorophyll is not contained in saclike plastids but is dissolved in the cellular fluid, or cytoplasm; and it does not evolve free oxygen during the energy-storing process. The next step probably was the appearance of the cyanophytes (blue-green algae), typically with a larger mass of cytoplasm than bacteria. Although single-celled, they may form colonies of considerable size. Their green pigment is true chlorophyll, and they show typical plantlike photosynthesis with the evolution of free oxygen (a reaction that stores energy at nearly three times the rate of the best of the chemotrophic reactions). The advent of these blue-green algae, some 2,000,000,000 or 3,000,000,000 years ago, began the increase of the world's free oxygen and paved the way for the development of the higher forms of life. Photosynthesis greatly increased the store of organic compounds and the energy available, while the oxygen atmosphere made possible the highly efficient oxidative metabolism of higher organisms.

Earliest organisms

After the blue-green algae there appeared an extensive array of algae, molds, protozoans, and higher plants and animals. Three groups of algae can be dismissed with passing mention, as they arose from uncertain ancestors and have given rise to no further groups. These groups are the chrysophytes (yellow-green and golden-brown algae, chiefly diatoms); the pyrrophytes (cryptomonads and dinoflagellates); and the rhodophytes, or red algae. Three more have greater phylogenetic importance: the chlorophytes (green algae), which almost certainly gave rise to the land plants—the bryophytes (mosses and liverworts) and the tracheophytes, or vascular plants (including all of the higher plants); the euglenoids (unicellular, flagellate organisms), which suggest a broad connection between plants and animals at this primitive level; and the phaeophytes (brown algae), which some biologists have considered to be a probable source of the animal kingdom. Two groups of molds may have arisen from bacterial ancestors, but they share many characteristics with the protozoans, and they may well have been derived from protozoan ancestors. Finally, the protozoans (unicellular or acellular animals) were derived from unknown, more primitive ancestors, and one or more groups of protozoans have given rise to metazoans, the higher animals.

Algae of phylogenetic importance

The probable main lines of descent. Traditionally, the basic taxonomy of the living world has been dominated by its most conspicuous members—the higher plants and

animals, which contrast obviously in many ways. This dichotomy was formalized by the recognition of two kingdoms, Plantae and Animalia, which were at first thought to encompass the entire world of living beings in two sharply separated groups. When the multicellular algae were studied, they were assigned to the plant kingdom, and many of the protozoans were readily assigned to the animal kingdom. Yet other protozoans and many of the unicellular algae seemed difficult to separate, and it strained the definitions of plant and animal rather severely to include bacteria or blue-green algae in either kingdom, although doubts were generally suppressed in favour of assignment of both groups to the plant kingdom. As to the viruses, it was commonly debated whether they were living at all, and, hence, no assignment to kingdom was made.

Because of these problems, various biologists have suggested that a taxonomy based upon three or four kingdoms would be truer to nature. In one four-kingdom system, the kingdom Mychota comprises the bacteria and the blue-green algae (viruses were excluded on the ground that they were not considered organisms); the kingdom Protoctista comprises the protozoans and all the higher algae except the green algae; the kingdom Plantae comprises the green algae, the bryophytes, and the tracheophytes; and the kingdom Animalia includes only the metazoans, the multicellular animals. In a three-kingdom system, the Mychota (or Monera of some authors) includes the same groups, but the protoctists are considered to have sufficiently clear relationships, some to higher plants and others to higher animals, that they are best assigned to these kingdoms. In another version of a three-kingdom system the kingdom Protista includes monerans, algae, fungi, and protozoans; the kingdom Animalia includes organisms from sponges to echinoderms and chordates; and the kingdom Plantae includes bryophytes and tracheophytes.

Such ideas have long had theoretical interest, but they did not begin to gain wide acceptance until the 1960s, when growing knowledge of the cell structure and genetics of micro-organisms made it clear that there is a marked discontinuity between the bacteria and blue-green algae on the one hand and the cells of all higher organisms on the other. Because of the contrast in structure and behaviour of the nucleus (Greek *karyon*), these most primitive organisms are designated as procaryotes; all higher organisms are designated as eucaryotes. The differences are many and profound. The nuclear material of the procaryotes is not separated from the cytoplasm by a membrane, and, when the cell divides, it does so by simple fission, without the complex division stages (mitosis) so striking in all eucaryote cells. The genetic material is represented by strands of DNA that, unlike the DNA of eucaryotes, is not combined with protein to form typical chromosomes. Genetic evidence, however, proves that the simple division is qualitatively equal; that is, complete sets of genes are passed to the two daughter cells. A sexual process does not occur in procaryotes, although there are several parasexual alternatives whereby genes are transferred from one cell to another. The cytoplasm is similarly very simple in procaryotes. The ribosomes, microscopic globules on which protein synthesis occurs, drift freely in the cytoplasm. There are no storage areas (vacuoles), cellular components such as mitochondria, or plastids for photosynthesis. Flowing, or streaming, of the cytoplasm does not seem to occur. Some bacteria have flagella, extremely fine, whiplike projections by which they can move, but the structure of these is usually much simpler than that of eucaryote flagella.

Procaryotes and eucaryotes

In contrast, all eucaryote cells have a nucleus separated from the cytoplasm by a membrane, and the DNA is bound with large amounts of protein to form chromosomes. They divide by mitosis, a complex division in which chromosomes are duplicated, condensed, arranged on a spindle-shaped temporary structure, then distributed to each of the two resulting nuclei, one member of each duplicated set going to each nucleus. The fusion of sex cells is not universal among eucaryotes, but it does occur in most groups. The cytoplasm of eucaryote cells is filled with fine membranes (the endoplasmic reticulum, visible only with the electron microscope) to which the ribosomes are attached. The cytoplasm also contains a series of components, or organelles, not found in procaryotes: mitochondria, in which respiratory enzymes are located; the basal granules and centrioles, structures upon which the locomotory hairs (cilia and flagella) and mitotic figures are based; and, in some cells, the plastids, which contain the photosynthetic pigment, chlorophyll. The fine structure of eucaryote cilia and flagella consists of a circlet of nine double filaments surrounding two central filaments.

Because of these extensive and fundamental differences, many biologists believe that the divergence between procaryotic and eucaryotic cells is the greatest single discontinuity among organisms. But how was this wide gap bridged? Surely, the eucaryotes must have been derived from procaryote ancestors. One theory of the origin of the eucaryote cell from the procaryote cell is based upon symbiosis, the invasion of one kind of cell by another to the mutual advantage of both cells, much as algae live within the cells of fungi in the lichen partnerships. Perhaps a primitive, heterotrophic procaryote, capable of anaerobic respiration (living without oxygen), ingested an aerobic procaryote, which required oxygen to live. As time passed, the partnership became obligatory, with the captured aerobes functioning as mitochondria, thus effectively conferring the advantages of aerobic respiration upon the host cell. Next, these protamoeba types of organism ingested flagellated procaryotes, which, in time, became flagella or cilia; in many lines of descent, only the basal bodies of the flagellated protozoan remained (with their DNA), and they became centrioles and centromeres, fundamental parts of the mitotic apparatus, the evolution of which thus became possible. Finally, such a cell ingested blue-green algae, which thus became plastids. The cell that was so constructed had effectively become a eucaryote.

The plausibility of such a sequence of symbioses is much enhanced by the fact that all of the organelles mentioned do, in fact, contain DNA, and they are self-reproducing within the eucaryote cell. Based upon this hypothesis of the origin of the eucaryotes, a basic phylogeny can be devised, with four kingdoms. The procaryotes comprise the kingdom Monera. Symbiotic modification of an anaerobic, heterotrophic procaryote would establish the eucaryote cell, the first members of the kingdom Protoctista. Some of these protoctists then formed symbiotic, multicellular associations with blue-green algae. In most of these, the cells of a colony all remained largely equivalent, thus giving rise to the several groups of algae. The green algae, however, showed a tendency to differentiate into specialized cells, and it seems very probable that the kingdom Plantae developed from such cells. Nonphotosynthetic eucaryotes (protamoebans) gave rise to the fungi (protoctists), to the protozoans (also protoctists), and, through one or another group of protozoans, to the kingdom Animalia, comprising all the multicellular animals with specialization of cells and division of labour among the cell types.

There has not been universal agreement on such a four-kingdom phylogeny. Some workers, using an array of biochemical data, derive the green algae and the red algae from the blue-greens by simple differentiation, and all higher plants they derive from the green algae. In such a scheme the single kingdom Plantae includes the procaryotes; by inference, all other organisms would belong to the kingdom Animalia, thus reaffirming the traditional two-kingdom system. A one-kingdom system, with all organisms assigned to the plant kingdom and animals deriving from the same phylogenetic stem as the brown algae, is also a possible arrangement. This interesting hypothesis has not received much support, because it separates widely some organisms that appear to be closely related on other grounds (as different groups of amoeboid protozoans), and it disregards the now well-established contrast between procaryotes and eucaryotes.

Perhaps a three-kingdom system represents a reasonable compromise and synthesis of present knowledge. Under one such system, the kingdom Mychota encom-

The three-kingdom system

passes the viruses, bacteria, and blue-green algae. (The viruses are here included because research has revealed a complex genetic system inconceivable except in an organism.) Viruses contrast with other procaryotes mainly by the absence of intermediary metabolism, for which they depend upon their hosts; however, it is quite usual for parasites to depend upon their hosts for some of the vital functions. The kingdom Plantae would include the eucaryote algae, all derived either directly or through an intermediate from the procaryotes via the symbiotic pathways described earlier; the fungi, which may be a connecting link to the animal kingdom, and the land plants—bryophytes and tracheophytes—both derived from the green algae. The kingdom Animalia would include the protozoans, with broad connections to the plant kingdom via the flagellate-euglenoid and flagellate-green-algan links, and the metazoans, the origin and relationships of which will be discussed below.

MAJOR TRENDS IN THE PHYLOGENY OF PLANTS

Algal predecessors. Whatever the origin of the eucaryote algae, all but the green algae were evolutionary blind alleys, having given rise to no further groups (Figure 1).

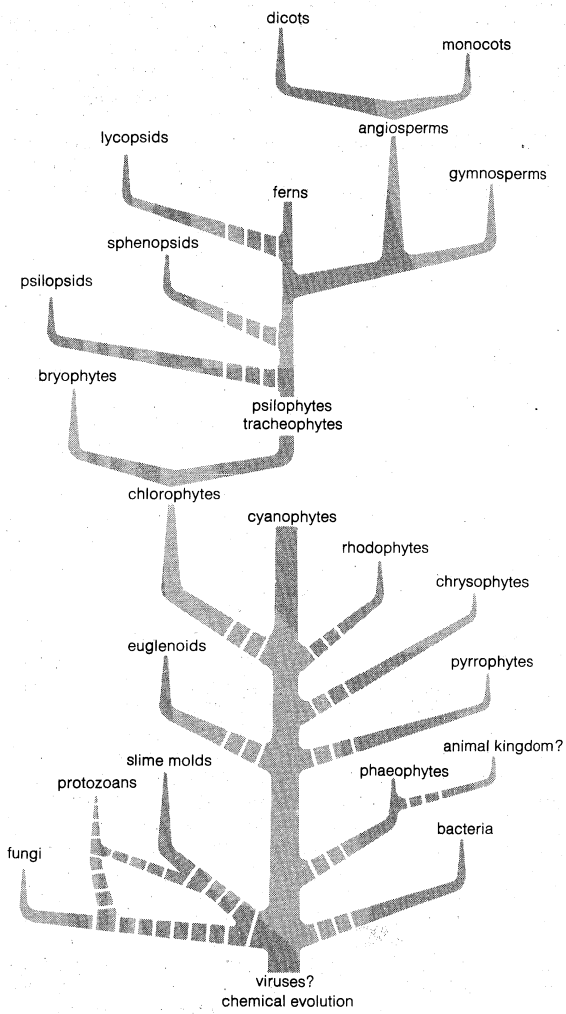


Figure 1: A summary of probable lines of plant evolution.

Although many algae are mainly single celled, most browns and reds are mainly multicellular; some, in fact, form seaweeds of very large size. Their multicellularity, however, has been achieved with a minimum of specialization of cells and of division of labour among them. The larger seaweeds form incipient organs, such as holdfasts and floats, but the body is best interpreted as a thallus, a simple aggregation of more or less similar cells, without specialized roots, stem, leaves, and conducting tissues, and lacking the specialized cell types that such structures require. Accordingly, the eucaryote algae may be grouped together as the subkingdom Thallophyta of the kingdom Plantae.

Connections between plants and animals

There are broad connections between these organisms, at the thallus level, and the animal kingdom, at the protozoan level. An amoeboid phase, for example, is prominent in the life cycles of the fungi, indicating that they probably arose from amoeboid protozoans or that both fungi and protozoans arose from the same primitive, amoeboid, ancestral eucaryote. Finally, it is difficult to draw a clear line between animal flagellates and euglenoids, on the one hand, and animal flagellates and green algae, on the other. To avoid this blurring of the lines between kingdoms, some biologists favour the interposition of the kingdom Protoctista to include the lower eucaryotes. Most algae, however, are so clearly allied to higher plants and most protozoans so clearly allied to animals that it seems better to recognize a certain amount of blurring of boundaries as a normal consequence of origin and diversification by evolution.

The range of variability of the green algae perhaps exceeds the combined ranges of the rest of the eucaryote algae. They may be unicellular (like *Chlamydomonas*), colonial (like *Pandorina*), or multicellular (like *Ulva*). Sexual reproduction is common among green algae and is always accompanied by an alternation of generations, a fundamental phenomenon in which a sexual phase, or gametophyte, containing one set of chromosomes (haploid), alternates with an asexual phase, or sporophyte, containing two sets of chromosomes (diploid).

Within this range of variability of the green algae are included most of the characters of the simpler land plants, although not all occur in any one alga. Indeed, the derivation of land plants from green algae seems to be one of the more secure conclusions of phylogeny.

Land plants. The two main groups of land plants, bryophytes and tracheophytes, are separated by many characteristics, but they also share certain features adaptive to dry-land habitats. Among these is the retention of the embryo within maternal tissue, for which these two groups are assigned to a subkingdom, Embryophyta.

The bryophytes, comprising mosses, liverworts, and hornworts, arose from green algae, have not progressed far beyond them, and have given rise to no further groups. Although they depend upon water for reproduction—for their sperm must swim to the eggs—a mere film of dew is sufficient. The fertilized egg develops within the maternal tissue, and the resulting sporophyte remains attached there even when mature, strictly dependent upon the gametophyte, for only the latter is supplied with chlorophyll. A waxy cuticle protects the exposed parts of the plant against drying.

The bryophytes were certainly derived from green algae. Many of the characteristics of the bryophytes, but not retention of the embryo within maternal tissues and possession of a cuticle, are found in green algae, although the complete ensemble is never found in any one species. Although one group of bryophytes, the hornworts, has progressive characteristics that suggest transitional stages to the vascular plants, it is certain that they did not give rise to the vascular plants, for the latter antedate the bryophytes in the fossil record by at least 100,000,000 years. The hornworts, therefore, must represent a terminal point of bryophyte evolution rather than a starting point for the higher plants.

The tracheophytes constitute a vast and richly varied assemblage that includes all of the dominant plants on earth today. They are characterized by having a dominant sporophyte phase, with the gametophyte reduced to microscopic proportions, and by having vascular elements that transport water and food within the plant. These properties have made possible the development of large plant bodies.

A further trend in the vascular plants has been the differentiation of tissues into roots, stems, and leaves, although this tendency is very tentative among the more primitive members. Finally, among the highest plants, the development of seeds and flowers is a major feature.

While it seems almost certain that the vascular plants arose from the green algae, specific details are not known,

Lignier's theory

and the fossil record has not been helpful. Around the turn of the century, a French botanist, O. Lignier, published a plausible theory that the ancestor of the vascular plants was a branching, filamentous green alga adapted to growing on tidal flats. Algae isolated in drying tide pools generally die, but one of the branches may have penetrated the soil and functioned as a root to supply the plant with water and minerals and, thus, secured its survival; meanwhile, another branch straightened to lift the reproductive parts into the air. A root system, however, could serve only a plant that had conducting tissue. Bundles of vascular tissue developed, forming transport pathways and strengthening the stems. The ends of the branches flattened and formed specialized organs of photosynthesis, the leaves. Finally, a cuticle developed to protect the exposed parts of the plant from drying.

Lignier's theory was purely speculative; nonetheless, fossils of early vascular plants have been found that accord with the theory sufficiently well that some enthusiasts have credited Lignier with describing the critical fossils before they were discovered. The oldest vascular plant known, *Cooksonia*, dates from Late Silurian time, more than 400,000,000 years ago. This simple plant approximates Lignier's hypothetical ancestral vascular plant quite closely. Later, during Early to Middle Devonian time, two other primitive plants appeared: *Psilophyton* and *Rhynia*, collectively called psilophytes. (Some authorities separate *Rhynia* as a group called rhyniophytes.) Fossils of these plants are so well preserved that even the microscopic structure is clearly apparent. The plants consisted largely of a branched, photosynthetic stem with small, spinelike "leaves" and terminal reproductive bodies (sporangia) on the aerial parts and rhizoids on the subterranean parts. The vascular tissue consisted of a simple core of large woody vessels (xylem) surrounded by smaller, thin-walled vessels (phloem).

The importance of the psilophytes derives not only from their antiquity but also from the fact that several genera displayed features that suggest other groups of vascular plants. *Psilophyton*, for example, had small, scalelike projections of the superficial layer of the stem, suggestive of the leaves of the lycopsids, or club mosses. *Hyenia* had whorled branches suggestive of the sphenopsids, or horsetails, and *Pseudosporochnus* had flattened branch tips, suggestive of the leaves of pteropsids, the ferns and flowering plants. It is probable, therefore, that all the principal groups of vascular plants may have arisen from psilophyte ancestors.

A different interpretation was widely advocated in the 1950s and later. Spores and fragments of vascular plants found in very ancient rock layers, or strata, suggest that the major groups of vascular plants must be much older than had been thought. The psilophytes appeared to be rather late offshoots of the most primitive plants rather than an ancestral group. Subsequent studies, however, have not supported this viewpoint. It appears that the fossils themselves, the strata in which they were found, or both, were misidentified; and hence the psilophytes have again taken a central position in the phylogeny of the vascular plants.

Club mosses and horsetails have not produced more progressive groups; however, the evolutionary branch leading from *Rhynia* through such genera as *Pseudosporochnus* and *Trimerophyton* to the ferns was more productive. These genera show a tendency towards flattening of whole branches to form frondlike leaves. Among the ferns, the differentiation into true roots, stem, and leaves is complete, but the stem is generally subterranean.

Seed plants. It will be recalled that one of the characteristics that permitted exploitation of the land was retention of the embryo within the maternal tissue. This trend culminated in the development of seeds. Not only is the embryo retained within the parent gametophyte but the gametophyte in turn is encased in and protected by tissue from the sporophyte that produced it. Thus, every seed includes tissues of three generations: the dominant sporophyte plant; the microscopic gametophyte phase; and the second generation sporophyte, or embryo. The seed ferns are an ancient group that originated in the Devonian Period (ending 345,000,000 years ago) and became extinct during the Mesozoic Era (ending 65,000,000 years ago). Whether they arose from true ferns or whether they developed along with true ferns from similar psilophyte ancestors is not known. In any event, they appear to have been the most primitive members of the gymnosperms.

Gymnosperms. Before they became extinct, the seed ferns gave rise to several distinct groups of gymnosperms, including the conifers (*e.g.*, pines, spruces, firs), which still form an important part of the world's vegetation.

The gymnosperms were the first plants to be independent of water for reproductive purposes. Pollen grains (male gametophytes) are released in large numbers and carried by the wind. Most released are lost, but those that alight on female cones of the same species may germinate and result in the formation of an embryo within the tissue of the female gametophyte. The embryo is encased in maternal tissue layers as the seed.

Angiosperms. The angiosperms, or flowering plants, must have arisen from gymnosperm ancestors, but little is known of their specific antecedents. The evolutionist Charles Darwin described their origin as "an abominable mystery," and it remains so today. The angiosperms appeared in great abundance among fossils of the Cretaceous Period (65,000,000–136,000,000 years ago), and it seems probable that their origin occurred much earlier. There are, indeed, fossils that possibly may be angiosperms of greater age. One of them is *Sanmiguelia*, a palmlike plant from the Triassic Period (190,000,000–225,000,000 years ago) for which no flowers have been uncovered. It is possible that it is not an angiosperm but rather a gymnosperm that approached the angiosperms in several features.

Monocotyledons and dicotyledons

Early in its history, the angiosperm group separated into two subgroups, the dicotyledons (dicots) and the monocotyledons (monocots). The dicots include the majority of flowering plants, and the monocots include lilies, palms, grasses, orchids, and a few related groups. The two subgroups contrast in many characters, of which a few easily observed ones may be mentioned. The flower parts of dicots commonly occur in fours and fives, whereas those of monocots occur in threes and sixes. The leaf veins form a network in the dicots, while those of monocots are more or less parallel. Finally, there are two cotyledons—seedling leaves in the embryo—in dicots and only one in monocots. In general, dicots appear to be more primitive than the monocots, and it is generally believed that monocots arose from some very primitive dicot by reduction of various parts.

It is generally believed—although fossil evidence is inadequate—that primitive flowers were large, with a branchlike axis on which numerous flower parts, rather leaflike in appearance, were spirally arranged. Colours of the petals served to attract insects, which pollinated the flowers. Among living plants, the magnolias approximate this plan most closely. A very general trend has been the reduction in number of floral parts. Among wind-pollinated plants especially, such as grasses, this trend is most obvious: sepals and petals may be reduced or missing altogether, and flowers may be unisexual, with pistils only (female) or stamens only (male). In some cases, flowers may be extremely reduced even though they are insect pollinated. In these, clusters, or inflorescences, of minute flowers (thus increasing the chances of many flowers being fertilized simultaneously) may be set within a whorl of white or coloured leaves that serve to attract insects. Familiar examples include the dogwood, poinsettia, and bunchberry.

Development of bilaterally symmetrical flowers is another important trend. In such flowers an insect enters only one way and is forced to brush against both stamens and stigma; thus, pollination is no longer left to chance. This trend reaches spectacular expression in the orchids among the monocots, but it is also well developed in such dicots as the snapdragon.

Fusion of parts is another common trend. The petals and sepals of trilliums are widely separate; those of Easter lilies are closely approximated to form a trumpetlike flower; and those of scilla are completely fused.

Finally, there is a trend toward aggregation of flowers into clusters. The extreme of this trend is found in the aster, or composite, family, as exemplified by asters, daisies, dandelions, and sunflowers. In this family, clusters of many flowers are gathered into compact heads, each of which resembles a single flower.

MAJOR TRENDS IN THE PHYLOGENY OF ANIMALS

Theories of the origin of animals

The problem of the origin of multicellular animals (metazoa) was long dominated by the gastrea theory of a German embryologist, Ernst Haeckel, who derived the ancestry of higher animals from the various embryological stages leading to their adult form. The higher animals, or metazoans, were believed to have come ultimately from some spherical, colonial protozoan comparable to *Volvox*. Such a hypothetical organism, called a gastread, was thought to have been structurally similar to the coelenterates (jellyfishes, corals), to which Haeckel accorded a key role in animal phylogeny. The theory, however, fell into disrepute, for it overvalued embryological evidence and undervalued other lines of evidence. In addition, the theory equated embryological stages of advanced animals with ancestral adults and assumed that embryos reflected ancestry undisturbed by adaptations to embryonic life; neither hypothesis is valid.

There are two major alternative theories. One derives the metazoans from colonial flagellates by integration of cells, leading to a flattened, ciliated animal (a planula) that gave rise to the coelenterates and, perhaps through them, to the flatworms. A minor group of parasitic worms, the mesozoans, retain planuloid structure even as adults. Whether mesozoan simplicity is primitive, or whether they have degenerated from flatworm ancestors is controversial.

The other theory derives the Metazoa from many-nucleated protozoans that subdivided into cells. This resulted first in a simple flatworm, to which is given the same importance that Haeckel gave the coelenterates. Some see close parallels between ciliates and flatworms in symmetry, excretion, reproduction, and in other respects; others feel that these similarities strain the facts. The theory suggests that flatworms gave rise to the coelenterates regressively and to all higher groups progressively.

Decisive data on these theories may never be available. Some biologists have suggested that perhaps neither is valid and that the earliest metazoans were an unknown group of planuloid worms that gave rise to coelenterates, flatworms, and mesozoans.

Several other possibilities deserve brief mention. Some biologists have suggested that the metazoans have arisen regressively from multicellular plants by loss of chlorophyll. The sudden appearance of most major groups of animals in the Cambrian Period (500,000,000–570,000,000 years ago) suggests to some that all or many of these groups may have arisen more or less simultaneously from some undefined ancestral form, probably at the procaryote level.

Changes in body plan: the lower metazoans. *Symmetry.* Symmetry, the geometrical distribution of parts about an axis, characterizes most animals. If there is a single longitudinal axis, so that any plane passing through it divides the animal into equivalent parts, radial symmetry obtains. The longitudinal axis usually includes the mouth. Radially symmetrical animals generally have a definite number of radii defined by structures (often four in jellyfishes, five in starfishes). In some protozoans and in the comb jellies, there are, in addition to the principal longitudinal axis, two secondary axes at right angles to each other and to the main axis. These axes differ structurally and are called the sagittal and transverse (*i.e.*, side-to-side) axes. A plane including the longitudinal axis and either of the other two will divide such a biradial animal into equal parts. Bilateral symmetry characterizes most animals. The same three axes are present as in radial symmetry, but the sagittal axis lies in such a way that dorsal (back) and ventral (belly) surfaces are defined. Only the plane that includes the longitudinal and sagittal axes divides the animal into equivalent parts, the right and left halves.

Some protozoans and most sponges (which are certainly derived from flagellate protozoans) are asymmetrical. Most protozoans, however are radial, without radii defined by specific structures, as in the mesozoans. Coelenterates, whether as jellyfishes or polyps, are radially symmetrical. Anemones and corals may show limited bilaterality defined only by the position of the mouthparts. This may be a minor overlay, of little significance, upon their basic radiality; but some regard it as a highly significant inheritance from flatworm ancestors, which were strongly bilateral. Their radial characteristics are considered to follow from their sedentary mode of life, for sessile animals meet the environment equally from all directions. There has been no loss of radiality in jellyfishes, however, which are highly mobile.

Asymmetrical animals

The flatworms are bilaterally symmetrical, and they are easily derivable from an ancestor with a single longitudinal axis, round in cross section, and with no defined radii. If such an animal were to become a bottom feeder, with a mouth on the under surface, the worm would tend to flatten against the surface, thus defining the secondary axes of bilateral symmetry.

Bilateral symmetry must have great selective value, for it characterizes all higher groups except the echinoderms (*e.g.*, starfishes, sea cucumbers). It is noteworthy that the larvae of echinoderms are bilateral, and it appears that their present radiality was a secondary acquisition.

The lower metazoans show a gradual progression to higher levels of organization. All metazoans are, of course, multicellular, but the sponges are at the cellular level in the sense that their cells are scarcely integrated into higher units. There are no clearly defined tissues, no organs, and scarcely more individuality than that of a protozoan colony. For these reasons and others, it is probable that the sponges are an isolated offshoot of colonial flagellates, unrelated to other metazoans. They are the sole members of the subkingdom Parazoa, the entire remaining assemblage of animals being members of the subkingdom Eumetazoa.

Development of tissues and organs. Three groups—the mesozoans, coelenterates, and ctenophores—are at the tissue level; that is, their cells are integrated into two tissues, an outer ectoderm, serving protective and sensory functions, and an inner entoderm, serving a digestive function (except in Mesozoa, in which the inner cells are reproductive). Generally, there are no organs in animals classified at the tissue level, but coelenterates and ctenophores may have fairly complex canal systems developed from the gastrovascular (digestive) cavity as well as an uncentralized nerve net and sensory organs of equilibrium. Their cell layers are separated by a greater or lesser amount of a jelly called mesoglea. There are generally some cells in the mesoglea, and in anemones they may be so abundant that they actually constitute a third cell layer, the mesoderm. There are no mesodermal organs as such, however, in these animals.

Animals organized at the organ level

The flatworms and all higher groups are at the organ level of construction. In general, the skin, sensory organs, and nervous system are formed from ectoderm; the lining of the digestive tract and the digestive glands from entoderm; and the rest of the organs from mesoderm. There is no body cavity in flatworms, but in most higher groups the principal organs are contained within a cavity (coelom) in the mesoderm. Scarcely advanced beyond the flatworms is a cluster of groups, sometimes treated as a rather loosely related group called aschelminths, in which the organs are held within highly vacuolated cells, forming what is termed a pseudocoel. In most higher groups, however, there is a true coelom, a cavity within the mesoderm and limited by a membrane called the peritoneum. Suspended within the coelom are most of the organ systems. The coelom is most striking in the annelids (segmented worms), the echinoderms, and the chordates (vetebrates and their lesser relatives).

Cephalization and segmentation. Cephalization (literally head formation) is another important trend in bilaterally symmetrical animals. The central nervous system, sense organs, and mouth are all clustered at the anterior end of the body. This tendency begins with the

flatworms, which have a simple head with an elementary brain, eye spots, and chemically sensitive lobes. In most invertebrate groups, this plan is varied in details without major improvements, but in annelid worms the central nervous system is more highly developed and the dominance of the brain more pronounced. In the arthropods (*e.g.*, insects, crustaceans), one of the most successful groups, cephalization is highly advanced, with the head including a well-developed brain, eyes, two pairs of sensory antennae, and, of course, the mouth and jaws. Some mollusks are retrograde in respect to cephalization, but one class, the Cephalopoda (octopuses, squids), have a highly developed brain and complex eyes; the head, however, is inseparably fused with the large fleshy portion called the foot. The head of chordates is the pinnacle of cephalization and includes a large brain, eyes, ears, nose, and mouth.

Segmentation of the body, with repetition of structures in every segment, is maximal in the annelids but is also prominent in the arthropods and chordates. Mollusks are generally unsegmented, but the Monoplacophora, a very ancient class of mollusks with but one known living member, are definitely segmented, suggesting that primitive segmentation was suppressed in the formation of the modern mollusks. Only the larvae of echinoderms show limited segmentation, as do some minor groups.

The origin of segmentation is conjectural. Possibly it arose as a response of the muscular system of an elongate animal to the demands of swimming, especially against the current of a stream. Serpentine swimming movements would be favoured by muscles arranged in blocks on either side of the body, capable of contracting in succession along the length of the body. This in turn would require segmental nerves from the nerve cord, and other systems might follow. In the annelids, almost every system is segmentally arranged; but in arthropods and chordates, segmentation is partially obscured by formation of complex structures from the rudiments of many segments and by suppression of some organs of many of the segments. It is probable that segmentation arose independently in the two lines of descent: annelid–arthropod and echinoderm–chordate.

Phylogenetic lines: the higher metazoans. *Monophyletic versus diphyletic theories.* With respect to lineage among higher animals, two major viewpoints are current. In the monophyletic theory, all of the metazoans are assigned to four groups, which arose in simple monophyletic, or linear, sequence from lower to higher. The theory is based on comparative anatomy, especially of the coelom and of segmentation. All of those organisms that are unsegmented and lack a true coelom, including flatworms, coelenterates, various minor groups, and even the mollusks, are assigned to the Ameria (meaning lacking segments). The strongly segmented annelids and arthropods, together with several minor groups, are presumed to have been derived from the Ameria and constitute the Polymeria. Some polymerians, such as the wormlike echiuroids and sipunculids, show a reduction of segmentation and thus lead to the Oligomeria, a group that includes insects, echinoderms, and several minor forms. The fourth group constitutes the chordates under the name of Chordonia.

The 18th-century chain of being

The originality of this theory is its greatest strength, but it goes back to the 18th-century idea of a *scala naturae*, or great chain of being, in which groups succeed one another on a linear ladder of life. These groups, however, include animals too highly diverse for most biologists to accept as being closely related. Much of the biology of mollusks and coelenterates, for example, would have to be overlooked to place them in one group under this definition.

The alternative—diphyletic—theory has been proposed by many zoologists. It envisions the higher metazoans as developing in two principal lines of descent—one culminating in the annelids, arthropods, and mollusks; the other in the echinoderms and chordates (Figure 2). The origin of these two lines has been referred to most of the lower noncoelomate groups, but it seems most probable that they took their origin from some ancient group of flatworms from which the modern flatworms also arose.

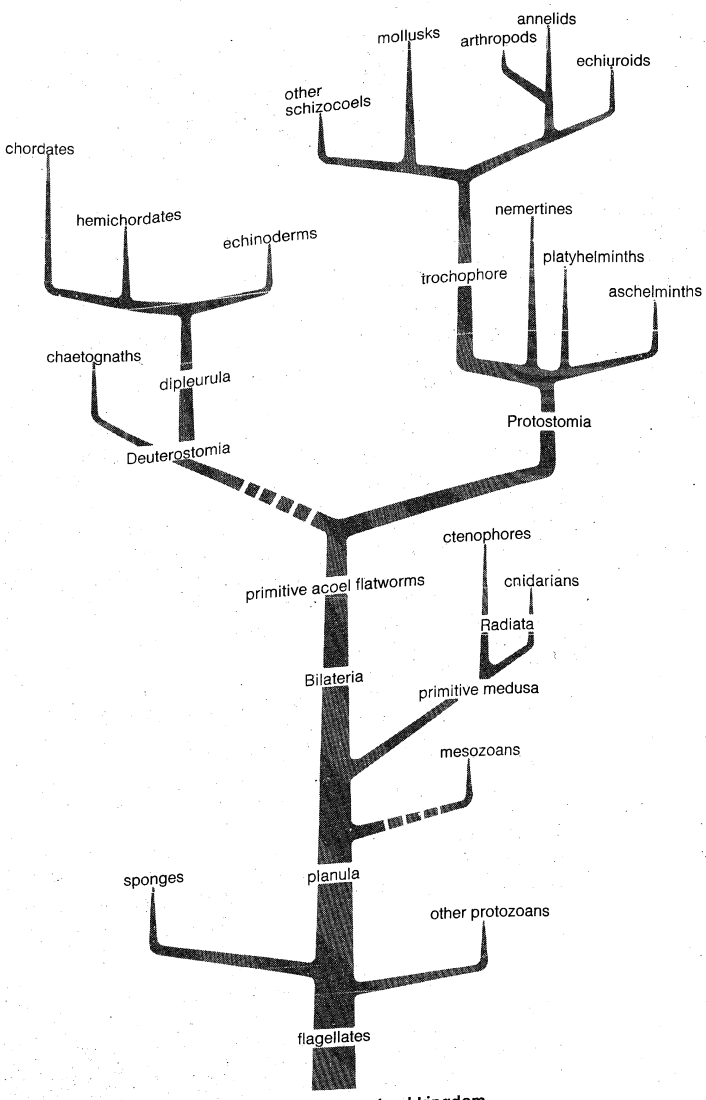


Figure 2: Hypothetical relationships in the animal kingdom.

The adult anatomies of animals in the annelid–arthropod line are strongly divergent, yet these organisms agree in many embryological traits; and a similar type of larva, called a trochophore, is common to all but the arthropods. The cleavage divisions (earliest divisions of the egg) are spiral, and they are determinate; *i.e.*, specific cells are destined to form predetermined structures. The coelom is called a schizocoel because it is formed by the splitting of the mesodermal cell mass. Finally, the embryonic mouth, or blastopore, becomes the mouth of the adult. This is the basis of the name of this line of descent, the Protostomia ("first mouth").

By contrast, the Deuterostomia show neither spiral nor determinate cleavage. Mesoderm and coelom are established at one stroke by outpocketings from the embryonic gut. The mouth is a new opening (hence the name Deuterostomia); the anus may be formed from the blastopore, or it too may form anew. Finally, there is no larval type common to all the deuterostomes (although one, the dipleurula, is common among lower deuterostomes).

Adaptive radiation. Both series show a radiation of lines of descent that exploit widely diverse structural plans and modes of life; for this reason the adults are difficult to compare, and relationships between the major groups must be proposed rather tentatively. The annelids and arthropods, however, are an exception. They share the protostome developmental characters as well as some adult traits, such as segmentation. As different as a typical worm and an arthropod such as an insect may appear, however, there is a minor group, the wormlike onychophorans, which is so nearly intermediate that the inference of close relationships is most suggestive.

The origin of chordates

The origin and phylogeny of the chordates hold especial interest, for this group climaxes the deuterostome series and includes man. Almost every invertebrate group has been considered as a source of the chordates, but the balance of evidence favours a common origin with the echinoderms, unexpected as this may seem. Chordates and echinoderms share the basic deuterostome traits. Further, the larvae of starfishes and of hemichordates (presumed lower chordates) are so similar that when the latter were first discovered, they were described as larval starfishes. The chordates are defined by three fundamental structures: a dorsal nerve tube (brain and spinal cord of vertebrates); a notochord (a supporting rod under the nerve tube); and a pharynx perforated by gill slits, at least in the embryo. The last of these is clearly developed in hemichordates, but the first two are equivocal. In balance, the hemichordates appear as a distinct group closely related to the chordates and sharing a common origin with them. Probably in Precambrian times a dipleurula-like bilateral deuterostome gave rise to two divergent lines of descent: one led to the aberrant echinoderms; the other to the more typical hemichordates and the very progressive chordates. In this line, a mutant with gill slits occurred and was favoured by selection, probably because the dipleurula was already a particle feeder; gill slits made it possible to sieve large quantities of water to extract food particles (still the principal function of gill slits in primitive chordates).

Which early chordates gave rise to the vertebrates is uncertain, but a good case has been made for the tunicates, the larvae of which show the three chordate characters with simple clarity. The earliest vertebrates were the ostracoderms, essentially "armoured lampreys" of the Ordovician Period (about 500,000,000 to 430,000,000 years ago) lacking both paired fins and jaws. They gave rise to the jawed placoderms, which had not only paired pectoral and pelvic fins but also, in some cases, a series of lesser fins between these.

Placoderms soon gave rise to sharks and to bony fishes in both salt water and fresh water. The bony fishes diversified greatly to dominate the waters of the world. They formed two groups: the ray-finned fishes, with basal bones and musculature of the fins entirely within the body wall; and the lobe-finned fishes, with the basal bones and muscles extending far into the fin itself. One group of the lobe-fins evolved into primitive amphibians during the Devonian Period. The amphibians had been preceded on land by plants and certain invertebrates. Feeding upon the latter, breathing by lungs and across the moist skin, and moving by means of lobe-fins that had become legs, the amphibians made their way upon the land. Still, because they had no protection against water loss across the skin, they were restricted to moist habitats and remained completely dependent upon water for reproduction.

Some of the early amphibians gave rise to reptiles, a group that completed the transition to dry land. A dry, hardened, outer layer of the skin restricted water loss, and the reptilian egg included membranes that enveloped the embryo and kept it bathed in fluid, thus freeing these animals of the water. Meanwhile, the reptilian limbs developed along the lines that were initiated by their amphibian ancestors.

Reptiles achieved their maximal development with the great array of dinosaurs during the Mesozoic Era, at the end of which the dinosaurs became extinct and the smaller types continued. During their greatest period of radiation, reptiles gave rise to birds and mammals. The former have many reptilian features, including hardened, scaly skin on their legs.

The mammals greatly modified reptilian structure. Three main examples may be cited. The limbs were rotated under the body, then modified adaptively in the many orders of mammals. Second, jaw structure was much simplified, and bones formerly part of the jaw suspension were relegated to the middle ear. Finally, the teeth, simple and uniform in other vertebrates, were differentiated into three types—incisors, canines, and molars—and these were further adapted to the widely different diets among the many mammalian orders.

SOME PHYLOGENETIC GENERALIZATIONS

The Precambrian fossil record consists largely of blue-green algae. Only late in the Precambrian do some invertebrates appear, and the dating of these has been disputed. Then suddenly in the Cambrian (beginning about 570,000,000 years ago) most major groups of invertebrates burst forth in profusion. Was their appearance really so sudden, or is the fossil record, which is always fragmentary, so extremely deficient at that great age that a once richly varied assemblage of life now appears almost monotypic? For more than a century, it was assumed that the Cambrian outburst was the product of a long development, the record of which had been lost. Recently, however, a highly plausible alternative theory has gained acceptance. As stated earlier, the primitive atmosphere was oxygen free, and oxygen began to accumulate only after the advent of photosynthesis. Perhaps it was only in Late Precambrian time that the oxygen level became high enough to support oxygen-breathing animals. Then, with almost all ecological positions (niches) for metazoans unoccupied, the major types evolved to fill those vacancies with startling rapidity. The Cambrian burgeoning may have been an unparalleled outburst of phylogenetic evolution.

The value of mutant types

Any group of organisms, if at all successful, produces an array of mutant types, which are tested by natural selection. The majority prove of no value in adapting the group to its shifting conditions of life, and so they tend to disappear. Occasional mutants, however, improve adaptive efficiency in some way, and these compete more successfully and contribute disproportionately more of their traits to succeeding generations. Different sets of mutant adaptations are favoured in different populations. As a result, an originally uniform group tends to fan out adaptively to occupy the widest possible range of available habitats. This positive drive toward diversification at all levels is called adaptive radiation.

After the origin of the eucaryotes, several groups of algae were quickly evolved and exploited a wide variety of aquatic habitats. One of these—the chlorophytes—was especially successful in its adaptive radiation, and its variants included protobryophytes and protopsilophytes. The latter radiated four principal adaptive types of vascular plants. One of these, the pteropsids, differentiated into three major types—ferns, gymnosperms, and angiosperms. The latter, especially, have radiated to form a large number of orders, families, and lower groups that collectively keep the world green.

More primitive groups, of course, give rise to more advanced groups, but it is noteworthy that it is not the advanced members of a group that produce new groups but, rather, the more primitive and less specialized members. For example, the early, primitive placoderms gave rise to the bony fishes and sharks. The lobe-finned fishes that blended into the amphibians were themselves primitive among the bony fishes. Similarly, the mammals arose from primitive, unspecialized reptiles.

The data and conclusions of phylogeny show clearly that the world of life is the product of a historical process, an evolution, and that degrees of resemblance within and between groups correspond to degrees of relationship by descent from common ancestors. A fully developed phylogeny is essential for the devising of a taxonomy that reflects the natural relationships within the world of living things.

BIBLIOGRAPHY

General: H.F. COPELAND, *The Classification of Lower Organisms* (1956), a summary statement by an early advocate of a four-kingdom system; E.O. DODSON, *Evolution: Process and Product* (1960), principles of evolution, with much phylogenetic information; and "The Kingdoms of Organisms" in *Systematic Zoology*, 20:265–282 (1971), a scientific paper that develops more fully the concepts provided in the above article; E.T. DRAKE (ed.), *Evolution and Environment* (1968), see esp. P.E. CLOUD, "Pre-Metazoan Evolution and the Origins of the Metazoa," and H.P. BANKS, "The Early History of the Land Plants"; G.A. KERKUT, *The Implications of Evolution* (1960), an evaluation of theories of the origin of the Metazoa, as well as other topics discussed in the present article.

The procaryote–eucaryote boundary: P. ECHLIN and I. MORRIS, "The Relationship between Blue-green Algae and Bacteria," *Biol. Rev.*, 40:143–187 (1965), an important review; L. MARGULIS, "Evolutionary Criteria in the Thallophytes: A Radical Alternative," *Science*, 161:1020–1022 (1968), a brief paper presenting a highly cogent synthesis of the problems of procaryote–eucaryote relationships and the kingdoms of organisms; R.Y. STANIER, M. DOUDOROFF, and E.A. ADELBERG, *The Microbial World*, 2nd ed. (1963), a textbook of microbiology that includes a good review of the traits that separate procaryotes and eucaryotes; R.H. WHITTAKER, "On the Broad Classification of Organisms," *Q. Rev. Biol.*, 34:210–226 (1959), one of the important presentations of a four-kingdom system.

Plant phylogeny: F.O. BOWER, *Primitive Land Plants, Also Known as Archegoniatae* (1935), a classic, and one of the bases of the psilophyte theory of the origin of vascular plants; W.C. DARRAH, *Principles of Paleobotany*, 2nd ed. (1960), a synthesis of "a reasonable phylogeny"; T. DELEVORYAS, *Plant Diversification* (1966), a concise introduction to plant phylogeny and paleobotany.

Animal phylogeny: N.J. BERRILL, *Origin of Vertebrates* (1955), a major presentation of the tunicate theory; E.O. DODSON, "Neo-Lamarckism, Modern Darwinism, and the Origin of the Vertebrates," *J. Paleont.*, 35:1065–1076 (1961), a critical review of theories of the origin of the vertebrates; E.C. DOUGHERTY, *et al.* (eds.), *The Lower Metazoa* (1963), many papers touching upon subjects germane to the present article; J. HADZI, "An Attempt to Reconstruct the System of Animal Classification," *Syst. Zool.*, 2:145–154 (1953), a good, brief presentation, and *The Evolution of the Metazoa* (1963), the major English language presentation of Hadzi's controversial views; C. HAND, "On the Origin and Phylogeny of the Coelenterates," *Syst. Zool.*, 8:191–202 (1958), perhaps the most telling criticism of Hadzi's theory; J. HUXLEY, A.C. HARDY, and E.B. FORD (eds.), *Evolution as a Process* (1954), see esp. the paper by G. DE BEER, on the evolution of the Metazoa in defense of Hadzi's theory.

(E.O.D.)

Physical Sciences, History of

The physical sciences are the refined and systematic results of man's efforts to understand his inorganic environment. At any given time, the character of these sciences is determined by and helps to determine the conceptual framework in which they are enmeshed: assumptions about the ultimate constitution of the world, what we are able to find out about the world, and what sorts of terms are appropriate for scientific explanations.

Problems in history of science

Because of the constantly changing foundations of science, the writing of its history presents several knotty problems. Although it is difficult not to judge the significance of past achievements in terms of currently acceptable theories, it is, nevertheless, profoundly ahistorical to do so. Obsolete theories that appear nonscientific and mystical today are usually better regarded as the result of an entirely different conception of nature rather than as aberrations of an unscientific frame of mind. To overlook this difficulty is to overlook the fact that the history of science should indeed be history. Since only a brief survey of the field can be made in this article, the problem of the selection of materials is most pressing. To portray something of the larger patterns of historical development, it is necessary to overlook many important and interesting chapters in the history of the several sciences.

ANCIENT SCIENCE

The physical sciences ultimately derive from the rationalistic materialism that emerged in classical Greece, itself an outgrowth of magical and mythical views of the world. The Greek philosophers of the 6th and 5th centuries BC abandoned the animism of the poets and explained the world order in terms of ordinarily observable natural processes. These early philosophers posed the broad questions that still underlie science: how did the world order emerge from pre-existing chaos? What is the origin of multitude and variety in the world? How can motion and change be accounted for? What is the relation between form and matter? Culminating in the philosophy of Aristotle in the 4th century BC, Greek philosophy answered these questions in terms that provided the framework for science for nearly 2,000 years.

Astronomy. Western astronomy had its origins in Egypt and Mesopotamia. Egyptian astronomy, which was neither a very well-developed nor an influential study, was largely concerned with time reckoning. Its main lasting contribution was the civil calendar of 365 days, consisting of 12 months of 30 days each and five additional festival days at the end of each year. This calendar played an important role in the history of astronomy, allowing astronomers to calculate conveniently the precise number of days between any two sets of observations.

Babylonian astronomy, dating back to about 1800 BC, constitutes one of the earliest systematic, scientific treatments of the physical world. In contrast to the Egyptians, the Babylonians were interested in the accurate prediction of several astronomical phenomena, especially the first appearance of the new Moon. Using the zodiac as a reference system, by the 4th century BC they developed a complex system of arithmetic progressions and methods of approximation by which they were able to predict first appearances. At no point in the clay tablets comprising the Babylonian astronomical literature is there the least evidence of the use of geometric models. The mass of observations they collected and their mathematical methods were important contributions to the later flowering of astronomy among the Greeks.

The Pythagoreans (5th century BC) were responsible for one of the first Greek astronomical theories. Believing that the order of the cosmos is fundamentally mathematical, they held that man is able to discover the harmonies of the universe by contemplating the regular motions of the heavens. Postulating a central fire about which all the heavenly bodies including the Earth revolve, they constructed the first physical model of the solar system. Subsequent Greek astronomy derived its character from a comment ascribed to Plato, in the 4th century BC, who is reported to have instructed the astronomers to "save the phenomena" in terms of uniform circular motion; that is, he urged them to develop predictively accurate theories using only various combinations of uniform circular motion. As a result, Greek astronomers never regarded their geometrical models as true or as being physical descriptions of the machinery of the heavens. They regarded them simply as instruments or tools, useful for predicting planetary positions.

Eudoxus of Cnidus (4th century BC) was the first of the Greek astronomers to rise to Plato's challenge. He developed a theory of homocentric spheres, a model that represented the universe by sets of nesting concentric spheres the motions of which combined to produce the planetary and other celestial motions. Using only uniform circular motions, Eudoxus was able to "save" the rather complex planetary motions with some success. His theory required four homocentric spheres for each planet and three each for the Sun and Moon. The system was modified by Callippus, a student of Eudoxus, who added spheres to improve the theory, especially for Mercury and Venus. Aristotle, in formulating his cosmology, adopted Eudoxus' homocentric spheres as the actual machinery of the heavens. The Aristotelian cosmos was like an onion consisting of a series of some 55 spheres nested about the Earth, which was fixed at the centre. In order to unify the system, Aristotle added spheres in order to "unroll" the motions of a given planet so that they would not be transmitted to the next inner planet.

Failures of Aristotelian astronomy

The theory of homocentric spheres failed to account for two sets of observations: (1) brightness changes suggesting that planets are not always the same distance from the Earth and (2) bounded elongations (that is, Venus is never observed to be more than about 48° and Mercury never more than about 24° from the Sun). Heracleides of Pontus (4th century BC) attempted to solve these problems by having Venus and Mercury revolve about the Sun, rather than the Earth, and having the Sun and other planets revolve in turn about the Earth, which he placed at the centre. In addition, to account for the daily motions of the heavens, he held that the Earth rotates on its axis. Heracleides' theory had little impact in antiquity except perhaps on Aristarchus of Samos (3rd century

BC), who apparently put forth a heliocentric hypothesis similar to the one Copernicus was to propound in the 16th century. Aristarchus' system was rejected for both astronomical and physical reasons.

Hipparchus (flourished 130 BC) made extensive contributions to both theoretical and observational astronomy. Basing his theories on an impressive mass of observations, he was able to work out theories of the Sun and Moon that were more successful than those of any of his predecessors. His primary conceptual tool was the eccentric circle, a circle of which the Earth is at some point eccentric to the geometric centre. He used this device to account for various irregularities and inequalities observed in the motions of the Sun and Moon. He also proved that the eccentric circle is mathematically equivalent to a geometric figure called an epicycle-deferent system, a proof probably first made by Apollonius of Perga a century earlier. Among his observations, one of the most significant was that of the precession of the equinoxes; that is, a gradual apparent increase in longitude between any fixed star and the equinoctial point (either of two points on the celestial sphere where the celestial equator crosses the ecliptic).

Claudius Ptolemy (flourished AD 140) applied the theory of epicycles to compile a systematic, detailed, and quantitative account of the whole realm of Greek astronomy. He elaborated theories for each of the planets, as well as for the Sun and Moon. His theory generally fitted the data available to him with a good degree of accuracy, and his book, the *Almagest*, became the vehicle by which Greek astronomy was transmitted to astronomers of the Middle Ages and Renaissance. It essentially molded astronomy for the next millennium and a half.

Greek physics. Several kinds of physical theories emerged in ancient Greece, including both generalized hypotheses about the ultimate structure of nature and more specific theories that considered the problem of motion from both metaphysical and mathematical points of view. Attempting to reconcile the antithesis between the underlying unity and apparent multitude and diversity of nature, the Greek atomists Leucippus (mid-5th century BC), Democritus (late 5th century BC), and Epicurus (late 4th and early 3rd century BC) asserted that nature consists of immutable atoms moving in empty space. According to this theory, the various motions and configurations of atoms and clusters of atoms are the causes of all the phenomena of nature.

The first atomists

In contrast to the particulate universe of the atomists, the Stoics (principally Zeno, of Citium, bridging 4th and 3rd centuries BC, Chrysippus [3rd century BC], and Poseidonius of Apamea [flourished *c.* 100 BC]) insisted on the continuity of nature, conceiving both space and matter as continuous and as infused with an active, air-like spirit—pneuma—which serves to unify the frame of nature. The inspiration for the Stoic emphasis on pneumatic processes probably arose from earlier experiences with the "spring" (*i.e.*, compressibility and pressure) of the air. Neither the atomic theory nor Stoic physics survived the criticism of Aristotle and the dominance of his theory over later developments.

In his physics, Aristotle was primarily concerned with the philosophical question of the nature of motion as one variety of change. He assumed that a constant motion requires a constant cause; that is to say, as long as a body remains in motion, a force must be acting on that body. He considered the motion of a body through a resisting medium as proportional to the force producing the motion and inversely proportional to the resistance of the medium. He used this relationship to argue against the possibility of the existence of a void, for in a void resistance is zero, and the relationship loses meaning. He considered the cosmos to be divided into two qualitatively different realms, governed by two different kinds of laws. In the terrestrial realm, within the sphere of the Moon, rectilinear up and down motion is characteristic. Heavy bodies, by their nature, seek the centre and tend to move downward in a natural motion. It is unnatural for a heavy body to move up, and such unnatural or violent motion requires an external cause. Light bodies, in direct contrast, move naturally upward. In the celestial realm, uniform circular motion is natural, thus producing the motions of the heavenly bodies.

Archimedes (3rd century BC) fundamentally applied mathematics to the solution of physical problems and brilliantly employed physical assumptions and insights leading to mathematical demonstrations, particularly in problems of statics and hydrostatics. He was thus able to derive the law of the lever rigorously and to deal with problems of the equilibrium of floating bodies.

SCIENCE IN THE MIDDLE AGES

Greek science reached a zenith with the work of Ptolemy in the 2nd century AD. The lack of interest in theoretical questions in the Roman world reduced science in the Latin West to the level of predigested handbooks and encyclopaedias that had been distilled many times. Social pressures, political persecution, and the anti-intellectual bias of some of the early Church Fathers drove the few remaining Greek scientists and philosophers to the East. There they ultimately found a welcome when the rise of Islām in the 7th century stimulated interest in scientific and philosophic subjects. Most of the important Greek scientific texts were preserved in Arabic translations. Although the Arabs did not alter the foundations of Greek science, they made several important contributions within its general framework. When interest in Greek learning revived in western Europe during the 12th and 13th centuries, scholars turned to Islāmic Spain for the scientific texts. A spate of translations resulted in the revival of Greek science in the West and coincided with the rise of the universities. Working within a predominantly Greek framework, scientists of the late Middle Ages reached high levels of sophistication and prepared the ground for the scientific revolution of the 16th and 17th centuries.

Arab contributions

Mechanics was one of the most highly developed sciences pursued in the Middle Ages. Operating within a fundamentally Aristotelian framework, medieval physicists criticized and attempted to improve many aspects of Aristotle's physics.

The problem of projectile motion was a crucial one for Aristotelian mechanics, and the analysis of this problem represents one of the most impressive medieval contributions to physics. Because of the assumption that continuation of motion requires the continued action of a motive force, the continued motion of a projectile after losing contact with the projector required explanation. Aristotle himself had proposed explanations of the continuation of projectile motion in terms of the action of the medium. The ad hoc character of these explanations rendered them unsatisfactory to most of the medieval commentators, who nevertheless retained the fundamental assumption that continued motion requires a continuing cause.

The most fruitful alternative to Aristotle's attempts to explain projectile motion resulted from the concept of impressed force. According to this view, there is an incorporeal motive force that is imparted to the projectile, causing it to continue moving. Such views were espoused by John Philoponus of Alexandria (flourished 6th century), Avicenna, the Persian philosopher (died 1037), and the Arab Abū al Barakāt al-Baghdādi (died 1164). In the 14th century the French philosopher Jean Buridan developed a new version of the impressed-force theory, calling the quality impressed on the projectile "impetus." Impetus, a permanent quality for Buridan, is measurable by the initial velocity of the projectile and by the quantity of matter contained in it. Buridan employed this concept to suggest an explanation of the everlasting motions of the heavens.

A group of Oxford scholars in the 14th century, considering the philosophical problem of how to describe the change that occurs when qualities increase or decrease in intensity, were led to consider the kinematic aspects of motion. Dealing with these problems in a purely hypothetical manner without any attempt to describe actual motions in nature or to test their formulas experimentally, they were able to derive the result that in a uniformly accelerated motion, distance increases as the square of the time.

Criticism of Aristotle

Although medieval science was deeply influenced by Aristotle's philosophy, adherence to his point of view was by no means dogmatic. During the 13th century, theologians at the University of Paris were disturbed by certain statements in Aristotle that seemed to imply limitations of God's powers as well as other statements, such as the eternity of the world, that stood in apparent contradiction to scripture. In 1277, Pope John XXI condemned 219 propositions, many from Aristotle and St. Thomas Aquinas, which had clearly theological consequences. Many of the condemned propositions had scientific implications as well. For example, one of the condemned propositions states, "That the first cause (*i.e.*, God) could not make several worlds." Although it is unlikely that anyone in the Middle Ages actually asserted the existence of many worlds, the condemnation led to the discussion of that possibility, as well as other important problems such as the possibility of the Earth's moving. In a backhanded way, the condemnations of 1277 may have stimulated a wider discussion of scientific and cosmological questions in the late Middle Ages.

THE SCIENTIFIC REVOLUTION

During the 15th, 16th, and 17th centuries scientific thought underwent a revolution. A new view of nature emerged, replacing the Greek view that had dominated science for almost 2,000 years. Science became an autonomous discipline, distinct from both philosophy and technology, and it came to be regarded as having utilitarian goals. By the end of this period, it may not be too much to say that science had replaced Christianity as the focal point of European civilization. Out of the ferment of the Renaissance and Reformation there arose a new view of science, bringing about the following transformations: the re-education of common sense in favour of abstract reasoning; the substitution of a quantitative for a qualitative view of nature; the view of nature as a machine rather than an organism; the development of an experimental method that sought definite answers to certain limited questions couched in the framework of specific theories; the acceptance of new criteria for explanation, stressing the "how" rather than the "why" that had characterized the Aristotelian search for final causes.

Astronomy. The scientific revolution began in astronomy. Although there had been earlier discussions of the possibility of the Earth's motion, Nicolaus Copernicus was the first to propound a comprehensive heliocentric theory, equal in scope and predictive capability to Ptolemy's geocentric system. Motivated by the desire to satisfy Plato's dictum, Copernicus was led to overthrow traditional astronomy because of its alleged violation of the principal of uniform circular motion and its lack of unity and harmony as a system of the world. Relying on virtually the same data as Ptolemy had possessed, Copernicus turned the world inside out, putting the Sun at the centre and setting the Earth into motion around it. Copernicus' theory, published in 1543, possessed a qualitative simplicity that Ptolemaic astronomy appeared to lack. To achieve comparable levels of quantitative precision, however, the new system became just as complex as the old. Perhaps the most revolutionary aspect of Copernican astronomy lay in Copernicus' attitude toward the reality of his theory. In contrast to Platonic instrumentalism, Copernicus asserted that to be satisfactory astronomy must describe the real, physical system of the world.

Reception of Copernican view

The reception of Copernican astronomy amounted to victory by infiltration. By the time large-scale opposition to the theory had developed in the Church and elsewhere, most of the best professional astronomers had found some aspect or other of the new system indispensable. Copernicus' book *De revolutionibus orbium coelestium, libri VI* ("Six Books on the Revolutions of the Celestial Orbs") became a standard reference for advanced problems in astronomical research, particularly for its mathematical techniques. Thus, it was widely read by mathematical astronomers, in spite of its central cosmological hypothesis, which was widely ignored. In 1551 the German astronomer Erasmus Reinhold published the *Tabulae prutenicae* ("Prutenic Tables"), computed by Copernican methods. The tables were more accurate and more up-to-date than their 13th-century predecessor and became indispensable to both astronomers and astrologers.

In the 16th century the Danish astronomer Tycho Brahe, rejecting both the Ptolemaic and Copernican systems, was responsible for immense changes in observation, unwittingly providing the data that ultimately decided the argument in favour of the new astronomy. Using larger, more stable, more rigid, and better calibrated instruments, he observed regularly over extended periods of time, thereby obtaining a continuity of observations that were accurate for planets to within about one minute of arc—better by several factors than any previous observations. Several of Tycho's observations contradicted Aristotle's system: a nova, which appeared in 1572, exhibited no parallax and was thus contrary to the Aristotelian assertion of the immutability of the heavens; similarly, a succession of comets appeared to be moving freely through a region that was supposed to be filled with solid, crystalline spheres. Tycho devised his own world system, which was a modification of Heracleides', to avoid various undesirable implications of both the Ptolemaic and Copernican systems.

Johannes Kepler, at the end of the 16th century, placed the Copernican hypothesis on firm astronomical footing. Converted to the new astronomy as a student and deeply motivated by a neo-Pythagorean desire for finding the mathematical principles of order and harmony according to which God had constructed the world, Kepler spent his life looking for simple mathematical relationships that described planetary motions. His painstaking search for the real order of the universe forced him finally to abandon the Platonic ideal of uniform circular motion in his search for a physical basis for the motions of the heavens.

Kepler's laws

In 1609 he announced two new planetary laws derived from Tycho's data: (1) the planets travel around the Sun in elliptical orbits, one focus of the ellipse being occupied by the Sun; (2) a planet moves in its orbit in such a manner that a line drawn from the planet to the Sun always sweeps out equal areas in equal times. With these two laws, Kepler abandoned uniform circular motion of the planets on their spheres, thus raising the fundamental physical question of what holds the planets in their orbits. He attempted to provide a physical basis for the planetary motions by means of a force analogous to the magnetic force, the qualitative properties of which had been recently described in England by William Gilbert in his influential treatise, *De Magnete, magneticisque corporibus et de magno magnete tellure*. The impending marriage between astronomy and physics had been announced. In 1618, Kepler stated his third law, which was one of many laws concerned with the harmonies of the planetary motions: (3) the square of the period in which a planet orbits the Sun is proportional to the cube of its mean distance from the Sun.

Galileo's observation

A powerful blow was dealt to the traditional cosmology by Galileo Galilei, who early in the 17th century used the telescope, a recent invention of Dutch lens grinders, to look toward the heavens. In 1610 Galileo announced observations that contradicted many traditional cosmological assumptions. He observed that the Moon was not a smooth, polished surface, as Aristotle had claimed, but that it was jagged and mountainous. Earthshine on the Moon revealed that the Earth, like the other planets, shone by reflected light. Like the Earth, Jupiter was observed to have satellites; hence, the Earth had been demoted from its unique position. The phases of Venus proved that that planet orbits the Sun, not the Earth.

Physics. *Mechanics.* The battle for Copernicanism was fought in the realm of mechanics as well as astronomy. The Ptolemaic–Aristotelian system stood or fell as a monolith, and it rested on the idea of Earth's fixity at the centre of the cosmos. Removing the Earth from the centre destroyed the doctrine of natural motion and place, and circular motion of the Earth was incompatible with Aristotelian physics.

Galileo's contributions to the science of mechanics were related directly to his defense of Copernicanism. Al-

though in his youth he adhered to the traditional impetus physics, his desire to mathematize in the manner of Archimedes led him to abandon the traditional approach and develop the foundations for a new physics that was both highly mathematizable and directly related to the problems facing the new cosmology. Interested in finding the natural acceleration of falling bodies, he was able to derive the law of free fall (the distance, s, varies as the square of the time, t^2). Combining this result with his rudimentary form of the principle of inertia, he was able to derive the parabolic path of projectile motion. Furthermore, his principle of inertia enabled him to meet the traditional physical objections to the Earth's motion: since a body in motion tends to remain in motion, projectiles and other objects on the surface of the Earth will tend to share the motions of the Earth, which will thus be imperceptible to someone standing on the Earth.

The 17th-century contributions to mechanics of the French philosopher René Descartes, like his contributions to the scientific endeavour as a whole, were more concerned with problems in the foundations of science than with the solution of specific technical problems. He was principally concerned with the conceptions of matter and motion as part of his general program for science—that is, to explain all the phenomena of nature in terms of matter and motion. This program, known as the mechanical philosophy, came to be the dominant theme of 17th-century science.

Descartes's principle of inertia

Descartes's interest in foundations led him to express the principle of inertia in much the form it retains today. He addressed himself to several important problems in mechanics, in particular the problems of impact and circular motion. Taking his lead from Descartes's formulations of these problems, the Dutch physicist Christiaan Huygens later in the 17th century solved them mathematically.

The work of Sir Isaac Newton represents the culmination of the scientific revolution at the end of the 17th century. His monumental *Principia* (1687) solved the major problems posed by the scientific revolution in mechanics and in cosmology. It provided a physical basis for Kepler's laws, unified celestial and terrestrial physics under one set of laws, and established the problems and methods that dominated much of astronomy and physics for well over a century. By means of the concept of force, Newton was able to synthesize two important components of the scientific revolution, the mechanical philosophy and the mathematization of nature.

The impact of Newton

Newton was able to derive all these striking results from his three laws of motion:

1. Every body continues in its state of rest or of motion in a straight line unless it is compelled to change that state by force impressed on it;
2. The change of motion is proportional to the motive force impressed and is made in the direction of the straight line in which that force is impressed;
3. To every action there is always opposed an equal reaction: or, the mutual actions of two bodies upon each other are always equal.

From these laws, Newton was able to derive the law of universal gravitation. He demonstrated that the fact that the planets obey Kepler's laws entails and is entailed by the existence of a force of attraction that is measured by the inverse of the square of the distance between two attracting bodies. Galileo's law of free fall also is derivable from Newton's laws of motion. The same physics that accounts for the motions of bodies on the surface of the Earth answers the question of what holds the planets in their orbits.

Optics. The science of optics in the 17th century expressed the fundamental outlook of the scientific revolution by combining an experimental approach with a quantitative analysis of phenomena. Optics had its origins in Greece, especially in the works of Euclid (*c.* 300 BC), who stated many of the results in geometrical optics that the Greeks had discovered, including the law of reflection: the angle of incidence is equal to the angle of reflection. In the 13th century, such men as Roger Bacon, Robert Grosseteste, and John Pecham, relying on the work of the Arab Alhazen (died 1039), considered numerous optical problems, including the optics of the rainbow. It was Kepler, taking his lead from the writings of these 13th-century opticians, who set the tone for the science in the 17th century. Kepler introduced the point by point analysis of optical problems, tracing rays from each point on the object to a point on the image. Just as the mechanical philosophy was breaking the world into atomic parts, so Kepler approached optics by breaking organic reality into what he considered to be ultimately real units. He developed a geometrical theory of lenses, providing the first mathematical account of Galileo's telescope.

Descartes sought to incorporate the phenomena of light into the mechanical philosophy by demonstrating that they can be explained entirely in terms of matter and motion. Using mechanical analogies, he was able to derive mathematically many of the known properties of light, including the law of reflection and the newly discovered law of refraction.

Newton's theory of colours

Many of the most important contributions to optics in the 17th century were the work of Newton, especially the theory of colours. Traditional theory considered colours to be the result of the modification of white light. Descartes, for example, thought that colours were the result of the spin of the particles that constitute light. Newton upset the traditional theory of colours by demonstrating in an impressive set of experiments that white light is a mixture out of which separate beams of coloured light can be separated. He associated different degrees of refrangibility with rays of different colours, and in this manner he was able to explain the way prisms produce spectra of colours from white light.

His experimental method was characterized by a quantitative approach, since he always sought measurable variables and a clear distinction between experimental findings and mechanical explanations of those findings. His second important contribution to optics dealt with the interference phenomena that came to be called "Newton's rings." Although the colours of thin films (*e.g.*, oil on water) had been previously observed no one had attempted to quantify the phenomena in any way. Newton observed quantitative relations between the thickness of the film and the diameters of the rings of colour, a regularity he attempted to explain by his theory of fits of easy transmission and fits of easy reflection. Although he generally conceived of light as being particulate, his theory of fits involves periodicity and vibrations of the ether.

Huygens was the second great optical thinker of the 17th century. Although he was critical of many of the details of Descartes's system, he wrote in the Cartesian tradition, seeking purely mechanical explanations of phenomena. Huygens regarded light as something of a pulse phenomenon, though he explicitly denied the periodicity of light pulses. He developed the concept of wave front, by means of which he was able to derive the laws of reflection and refraction from his pulse theory and to explain the recently discovered phenomenon of double refraction.

Chemistry. Chemistry had manifold origins, coming from such diverse sources as philosophy, alchemy, metallurgy, and medicine. It emerged as a separate science only with the rise of the mechanical philosophy in the 17th century. Aristotle had regarded the four elements earth, water, air, and fire as the ultimate constituents of all things. Transmutable each into the other, all four elements were believed to exist in every substance. Alchemy, having its origins in Egypt and the Near East, had a double aspect: on the one hand it was a practical endeavour aimed to make gold from baser substances, while on the other it was a cosmological theory based on the correspondence between man and the universe at large. Alchemy contributed to chemistry a long tradition of experience with a wide variety of substances. Paracelsus, a 16th-century Swiss natural philosopher, was a seminal figure in the history of chemistry, putting together in an almost impenetrable combination the Aristotelian theory of matter, alchemical correspondences, mystical forms of

knowledge, and chemical therapy in medicine. His influence was widely felt in succeeding generations.

During the first half of the 17th century, there were few established doctrines that chemists generally accepted as a framework. As a result there was little cumulative growth of chemical knowledge. Chemists tended to build detailed systems, "chemical philosophies," attempting to explain the entire universe in chemical terms. Most chemists accepted the traditional four elements (air, earth, water, fire), or the Paracelsian principles (salt, sulfur, mercury), or both, as the bearers of real qualities in substances; they also exhibited a marked tendency toward the occult.

Effect of mechanical theory on chemistry

The interaction between chemistry and the mechanical philosophy altered this situation by providing chemists with a shared language. The mechanical philosophy had been successfully employed in other areas; it seemed consistent with an experimental empiricism, and it seemed to provide a way to render chemistry respectable by translating it into the terms of the new science. Perhaps the best example of the influence of the mechanical philosophy is the work of Robert Boyle. The thrust of his work was to understand the chemical properties of matter, to provide experimental evidence for the mechanical philosophy, and to demonstrate that all chemical properties can be explained in mechanical terms. He was an excellent laboratory chemist and developed a number of important techniques, especially colour-identification tests.

EIGHTEENTH CENTURY TO THE PRESENT

Seminal contributions to science are those that change the tenor of the questions asked by succeeding generations. The works of Sir Isaac Newton formed such a contribution. The mathematical rigour of the *Principia* and the experimental approach of the *Opticks* became models for scientists of the 18th and 19th centuries. Celestial mechanics developed in the wake of the *Principia*, extending its scope and refining its mathematical methods. The more qualitative, experimental, and hypothetical approach of the *Opticks* deeply influenced the sciences of optics, electricity and magnetism, and chemistry.

Celestial mechanics and astronomy. Eighteenth-century theoretical astronomy in large measure derived both its point of view and its problems from the *Principia*. In this work Newton had provided a physics for the Copernican world view by, among other things, demonstrating the implications of his gravitational theory for a two-body system consisting of the Sun and a planet. While Newton himself had grave reservations as to the wider scope of his theory, the 18th century witnessed various attempts to extend it to the solution of problems involving three gravitating bodies.

An English astronomer, Edmond Halley, early in the 18th century, having noted striking similarities in the comets that had been observed in 1378, 1456, and especially in those of 1531, 1607, and 1682, argued that they were the periodic appearances every 75 years or so of but a single comet that he predicted would return in 1758. Months before its expected return, the French mathematician Alexis Clairaut employed rather tedious and brute-force mathematics to calculate the effects of the gravitational attraction of Jupiter and Saturn upon the otherwise elliptical orbit of Halley's Comet. Clairaut was finally able to predict in the fall of 1758 that Halley's Comet would reach perihelion, or the point in its orbit nearest the Sun, in April of 1759, with a leeway of one month. Its actual return, in March, was an early confirmation of the wider scope and power of the Newtonian theory.

It was, however, the three-body problem of either two planets and the Sun or the Sun–Earth–Moon system that provided the most persisting and profound tests of Newton's theory. These problems, involving more regular members of the solar system (*i.e.*, those describing nearly circular orbits having the same sense of revolution and in nearly the same plane), permitted certain simplifying assumptions, thereby inviting more general and elegant mathematical approaches than the cometary problem. An illustrious group of 18th-century continental mathematicians (including Clairaut, the Swiss Bernoullis and Leonhard Euler, and the Frenchmen Jean Le Rond d'Alembert, Joseph-Louis Lagrange and Pierre-Simon Laplace) attacked these astronomical problems, as well as related ones in Newtonian mechanics, by developing and applying the calculus of variations as it had been formulated by Leibniz. It is a lovely irony that this continental exploitation of Leibniz' mathematics—which was itself closely akin to Newton's version of calculus, which he called fluxions—was fundamental for the deepening establishment of the Newtonian theory to which Leibniz had objected because it reintroduced, according to Leibniz, occult forces into physics.

In order to attack the lunar theory, which also commanded attention as the most likely astronomical approach to the navigational problem of determining longitude at sea (for the solution of which was offered a £20,000 prize set by the Board of Longitude), Clairaut, having derived general equations that neither he nor anyone else could integrate, was forced to adopt methods of approximation. Even so, Clairaut was unable to calculate from gravitational theory a value for the progression of the lunar apogee greater than 50 percent of the observed value; therefore, he supposed in 1747 (with Euler) that Newton's inverse-square law was but the first term of a series and, hence, an approximation not valid for distances as small as that between the Earth and Moon. This attempted refinement of Newtonian theory proved to be fruitless, however, and two years later Clairaut was able to obtain, by more detailed and elaborate calculations, the observed value from the simple inverse-square relation.

Certain of the three-body problems, most notably that of the secular acceleration of the Moon, defied early attempts at solution but finally yielded to the increasing power of the calculus of variations in the service of Newtonian theory. Thus it was that Laplace—in his five-volume *Traité de mécanique céleste* (1798–1827)—was able to comprehend the whole solar system as a dynamically stable, Newtonian gravitational system. The secular acceleration of the Moon reappeared as a theoretical problem in the middle of the 19th century, persisting into the 20th century, and ultimately requiring tidal action as part of its solution.

Newtonian theory was also employed in much more dramatic discoveries that captivated the imagination of a tremendously broad and varied audience. Within 40 years of the discovery of Uranus in 1781 by the English astronomer William Herschel, it was recognized that the planet's motion was somewhat anomalous. In the next 20 years the gravitational attraction of an unobserved planet was suspected to be the cause of Uranus' persisting deviations. In 1845–46 the English astronomer J.C. Adams and the Frenchman U.-J.-J. Le Verrier independently calculated the position of this unseen body. The visual discovery (at the Berlin Observatory in 1846) of Neptune in just the position predicted constituted an immediately engaging and widely understood confirmation of Newtonian theory. In 1905 the English astronomer Percival Powell predicted the position of yet another outer planet to account for further perturbations of Uranus not caused by Neptune. Pluto was finally discovered by sophisticated photographic techniques in 1930 and observed visually in 1950.

In the second half of the 19th century, the innermost region of the solar system also received attention. In 1859 Le Verrier calculated the specifications of an intramercurial planet to account for a residual advance in the perihelion of Mercury's orbit (38 seconds of arc per century), an effect that was not gravitationally explicable in terms of known bodies. While a number of sightings of this predicted planet were reported between 1859 and 1878, the first of these resulting in Le Verrier's naming the new planet Vulcan, they were not confirmed by observations made either during subsequent solar eclipses or at the times of predicted transits of Vulcan across the Sun.

The theoretical comprehension of Mercury's residual motion involved the first successful departure from Newtonian gravitational theory. This came in the form of Einstein's formulation of the principle of the general the-

ory of relativity, one consequence of which was a correction term for a body in motion; the correction involves the square of the ratio of the body's velocity to the enormous velocity of light and hence, among the planets, is appreciable only for the fastest: Mercury. Einstein's theory accounted for the residual effect, which by 1915 was calculated to be 43 seconds, and this achievement combined with the 1919 observation of the bending of a ray of light passing near a massive body (another consequence of general relativity theory) constitute the main experimental verifications of that theory. In 1967, however, the U.S. scientist Robert Dicke argued that an equatorial bulge that he had observed on the Sun sufficed to account, in good Newtonian fashion, for 4 seconds of arc which, if confirmed, leaves a residual effect of only 39 seconds to be accounted for relativistically.

Mercury's motion and the theory of relativity

The 18th-, 19th-, and early-20th-century astronomy was not quite so completely Newtonian, however. Herschel's discovery of Uranus, for example, was not directly motivated by gravitational considerations. Nine years earlier, a German astronomer, J.D. Titius, had announced a purely numerical sequence, subsequently refined by another German astronomer, Johann E. Bode, that related the mean radii of the planetary orbits—a relation entirely outside gravitational theory. The sequence is given by $0 + 4 = 4$, $3 + 4 = 7$, $3 \times 2 + 4 = 10$, $3 \times 4 + 4 = 16$, and so on, yielding additional values of 28, 52 and 100. If the measured radius of the Earth's orbit is defined as being 10, then to a very good approximation that of Mercury is 4, Venus is 7, Mars is 15 plus, Jupiter is 52, and Saturn is 95 plus. The fit where it can be made is good and continues since the next number in the sequence is 196 and the measured radius for Uranus' orbit is 191, but no planet had been observed to correspond to the Bode–Titius law value of 28. Powerful computational methods finally led to the rediscovery of the asteroid Ceres that confirmed the law. The Bode–Titius law subsequently provided first simplifying assumptions for the calculations of the predicted positions, which led to the observations of Neptune and Pluto, while the novel properties of the asteroids (nearly 500 of which had been discovered by the end of the century) stimulated star charts of the zodiacal regions, provided the means for improved measurements of solar-system distances, and forced astronomers, by their very number and variety, to face the question of the allocation of resources.

Attempts to detect the motion of the Earth provided the 18th and 19th centuries with observational problems directly motivated by the Copernican theory. In 1728 the English astronomer James Bradley attributed annual changes that he observed in stellar positions to a slight tilting of the telescope with respect to the true direction of the star's light, a tilting that compensated for the Earth's motion. This effect, which depends also upon the ratio of the Earth's velocity to the velocity of light, is the so-called aberration of light.

While attention has been focussed on the more positional aspects of astronomy, mention should be made of two other broad areas of investigation that in their 19th-century form derived largely from the work of William Herschel. These areas, dealing with more structural features of the heavens and with the physical character of the stars, developed in large measure with advancements in physics.

Optics. Since they provided the principal basis for subsequent investigations, Newton's optical views were subject to close consideration until well into the 19th century. From his researches into the phenomena of colour, Newton became convinced that dispersion necessarily accompanies refraction and that chromatic aberration (colour distortion) could therefore be eliminated by employing reflectors, rather than refractors, as telescopes. By the mid-18th century Euler and others had theoretical arguments against Newton, and Euler offered the human eye as an example of an achromatic lens system. Although he was virtually alone in this, Euler also rejected Newton's essentially corpuscular theory of the nature of light by explaining optical phenomena in terms of vibrations in a fluid ether. The dominance of Newton's theory throughout the 18th century was due partly to its successful direct application by Newton and his followers and partly to the comprehensiveness of Newton's thought. For example, Bradley's observations found an immediate and natural explanation in terms of the corpuscular theory that also was supported by the accelerating success of Newton's gravitational theory involving discrete particles of matter.

Criticism of Newton's optics

At the turn of the century, Thomas Young, an English physician studying the power of accommodation of the eye (*i.e.*, its focussing power), was led gradually to extensive investigations and discoveries in optics, including the effect of interference. By means of a wave theory of light, Young was able to explain both this effect, which in its most dramatic manifestation results in two rays of light cancelling each other to produce darkness, and also the various colour phenomena observed by Newton. The wave theory of light was developed from 1815 onward in a series of brilliant mathematical and experimental memoirs of Augustin Fresnel but was countered by adherents of the corpuscular theory, most notably by a group of French scientists, Pierre-Simon Laplace, Siméon-Denis Poisson, Étienne Malus, and Jean-Baptiste Biot, and most strikingly in connection with Malus' discovery (1808) of the polarization of light by reflection. Following Young's suggestion in 1817, Fresnel was able to render polarization effects comprehensible by means of a wave theory by considering light to be a transverse rather than a longitudinal wave, as the analogy with sound had suggested.

The propagation of a transverse wave, the velocity of which through various media and under a variety of conditions was measured terrestrially with increasing accuracy from midcentury onward, seemed to require an ether having the properties of a highly elastic solid (*e.g.*, such as steel), which, however, offered no resistance to the planetary motions. These bizarre properties stimulated a number of mechanical models of the ether, most notably those of the English physicist William Thomson, Lord Kelvin. In order to encompass the aberration of light by means of his wave theory, Fresnel had assumed that the motionless ether freely permeated the opaque Earth and thus remained unaffected by its motions. Furthermore, he derived as a theoretical consequence (verified experimentally in midcentury by Armand-Hippolyte-Louis Fizeau) that the ether was partially, and only partially, dragged along by a moving transparent substance depending upon the index of refraction of the substance. However, all subsequent investigators (most notably the U.S. physicists A.A. Michelson and Edward W. Morley, in 1887) failed in their attempts to measure the required ether drift. It was just to escape this difficulty of a necessary but undetected ether drift that George Francis FitzGerald of England and the Dutch theorist Hendrik Antoon Lorentz independently, at the close of the century, postulated the contraction of moving bodies in the direction of their motion through the ether. The Lorentz-FitzGerald contraction involves the square of the ratio of the velocity of the body to the velocity of light to insure theoretically the experimental undetectability of the ether drift. It was supported by late-19th-century considerations of the electrical nature of matter. It was partly the seeming necessity of theoretical postulations of this kind that led Einstein to his formulation of relativity theory.

Electricity and magnetism. Until the end of the 18th century, investigations in electricity and magnetism exhibited more of the hypothetical and spontaneous character of Newton's *Opticks* than the axiomatic and somewhat forbidding tone of his *Principia*. Early in the century, in England Stephen Gray and in France Charles François de Cisternay DuFay studied the direct and induced electrification of various substances by the two kinds of electricity (then called vitreous and resinous and now known as positive and negative), as well as the capability of these substances to conduct the "effluvium" of electricity. By about midcentury, the use of Leyden jars (to collect charges) and the development of

Theories of electricity

large static electricity machines brought the experimental science into the drawing room, while the theoretical aspects were being cast in various forms of the single-fluid theory (by Benjamin Franklin of the U.S. and F.M.U.T.H. Aepinus among others) and the two-fluid theory.

By the end of the 18th century, in England, Joseph Priestley had noted that no electric effect was exhibited inside an electrified hollow metal container and had brilliantly inferred from this similarity that the inverse-square law (of gravity) must hold for electricity as well. In a series of painstaking memoirs, Charles-Augustin Coulomb, using a torsion balance Henry Cavendish had used in England to measure the gravitational force, demonstrated the inverse-square relation for electrical and magnetic attractions and repulsions. Coulomb went on to apply this law to calculate the surface distribution of the electrical fluid in such a fundamental manner as to provide the basis for the 19th-century extensions by Poisson and Lord Kelvin.

The discovery of galvanic electricity and the development of voltaic electricity opened whole new areas of investigation for the 19th century by providing convenient sources of sustained electrical current. Danish physicist Hans Christian Ørsted's discovery, in 1819, of the magnetic effect accompanying an electric current led almost immediately to quantitative laws of electromagnetism and electrodynamics. By 1827, André-Marie Ampère had published a series of mathematical and experimental memoirs on his electrodynamical theory that not only rendered electromagnetism comprehensible but also ordinary magnetism, identifying both as the result of molecular electrical currents. Ampère solidly established his electrodynamics by basing it on inverse-square forces (which, however, are directed at right angles to, rather than in the line connecting the two interacting elements) and by demonstrating that the effects do not violate Newton's third law of motion, notwithstanding their transverse direction.

English scientist Michael Faraday's discovery in 1831 of electromagnetic induction (the inverse of the effect discovered by Ørsted), his experimental determination of the identity of the various forms of electricity (1833), his discovery of the rotation of the plane of polarization of light by magnetism (1845), in addition to certain findings of other investigators—*e.g.*, the discovery by James Prescott Joule in 1843 (and others) of the mechanical equivalent of heat (the conservation of energy)—all served to emphasize and re-emphasize the essential unity of the forces of nature. Within electricity and magnetism attempts at theoretical unification were conceived in terms of either gravitational-type forces acting at a distance, as with Ampère, or, with Faraday, in terms of lines of force and the ambient medium in which they were thought to travel. The German physicist Wilhelm Eduard Weber, in order to determine the coefficients in his theory of the former kind, measured the ratio of the electromagnetic and electrostatic units of electrical charge to be equal to the velocity of light. Physicist James Clerk Maxwell, drawing his conceptions from Faraday and, thus, relying fundamentally on the ether required by optical theory, while using ingenious mechanical models, developed his profound mathematical electromagnetic theory from 1855 onward. One consequence of Maxwell's mature theory was that an electromagnetic wave must be propagated through the ether with a velocity equal to the ratio of the electromagnetic to electrostatic units. Combined with the earlier results of Weber and Rudolph Kohlrausch, this result implied that light is an electromagnetic phenomenon. Maxwell's inclusion of light within his electromagnetic theory received direct experimental verification in 1886, when Heinrich Hertz detected the predicted electromagnetic waves.

The nature of light

Impressive as this synthesis was, there remained, of course, a number of outstanding problems to which fresh difficulties were immediately added. On the experimental side, for example, Hertz' discovery in 1887 of the photoelectric effect led very quickly to a number of experimental results that were fundamentally inconsistent with any conceivable wave theory of light. Theoretical problems such as those involving the nature of the electric charge, the interaction between matter and ether, the optics of moving bodies, and the dispersion of light led Lorentz toward the end of the 19th century to refine Maxwell's theory in terms of charged particles, a conception that he developed into his 20th-century theory of electrons.

Chemistry. Contemporary ideas of the nature and scope of chemistry are much less suitable for the 18th century than are modern notions of the domain of astronomy and physics. Eighteenth-century chemistry derived from and remained involved with questions of mechanics, light, and heat as well as iatrochemistry and notions of medical therapy and the interaction between substances and the formation of new substances. Like the other sciences, chemistry also took many of its problems and much of its viewpoint from the *Opticks* and especially the "Queries" with which that work ends. Newton's suggestion of a hierarchy of clusters of unalterable particles formed by virtue of the specific attractions of its component particles led directly to comparative studies of interactions and thus to the tables of affinities of the physician Herman Boerhaave and others early in the century. This work culminated at the end of the century in the Swede Torbern Bergman's table that gave quantitative values of the affinity of substances both for reactions when "dry" and when in solution and that considered double as well as simple affinities.

Seventeenth-century investigations of "airs" or gases, combustion and calcination, and the nature and role of fire were incorporated by the chemists Johann Joachim Becher and Georg Ernst Stahl of Sweden into a theory of phlogiston. According to this theory, which was most influential after the middle of the 18th century, the fiery principle, phlogiston, was released into the air in the processes of combustion, calcination, and respiration. According to the theory, air was simply the receptacle for phlogiston, and any combustible or calcinable substance contained phlogiston as a principle or element and thus could not itself be elemental. Iron, in rusting, was considered to lose its compound nature and to assume its elemental state as the calx of iron by yielding its phlogiston into the ambient air.

Investigations that isolated and identified various gases in the second half of the 18th century, most notably the English chemist Joseph Black's quantitative manipulations of "fixed air" (carbon dioxide) and Joseph Priestley's discovery of "dephlogisticated air" (oxygen), were instrumental for the French chemist Antoine Lavoisier's formulation of his own oxygen theory of combustion and rejection of the phlogiston theory. This transformation coupled with the reform in nomenclature at the end of the century (due to Lavoisier and others)—a reform that reflected the new conceptions of chemical elements, compounds, and processes—comprised the revolution in chemistry.

Very early in the 19th century, another study of gases, this time in the form of a persisting Newtonian approach to certain meteorological problems by the chemist John Dalton, led to the enunciation of a chemical atomic theory. From this theory, which was demonstrated to agree with the law of definite proportions and from which the law of multiple proportions was derived, Dalton was able to calculate definite atomic weights by assuming the most simple possible ratio for the numbers of combining atoms. For example, knowing from experiment that the ratio of the combining weights of hydrogen to oxygen in the formation of water is 1 to 8 and by assuming that one atom of hydrogen combined with one atom of oxygen, Dalton affirmed that the atomic weight of oxygen was eight, based on hydrogen as one. At the same time, however, in France, Joseph-Louis Gay-Lussac, from his volumetric investigations of combining gases, determined that two volumes of hydrogen combined with one of oxygen to produce water. While this suggested H_2O rather than Dalton's HO as the formula for water, with the result that the atomic weight of oxygen becomes 16, it did involve certain inconsistencies with Dalton's theory.

As early as 1811 the chemist Amedeo Avogadro was able to reconcile Dalton's atomic theory with Gay-Lussac's volumetric law by postulating that Dalton's atoms were indeed compound atoms, or polyatomic. For a number of reasons, one of which involved the recent successes of electrochemistry, Avogadro's hypothesis was not accepted until it was reintroduced by another chemist, Stanislao Cannizzaro, half a century later. From the turn of the century, the English scientist Humphry Davy and many others had employed the strong electric currents of voltaic piles for the analysis of compound substances and the discovery of new elements. From these results, it appeared obvious that chemical forces were essentially electrical in nature and that two hydrogen atoms, for example, having the same electrical charge, would repel each other and could not join to form the polyatomic molecule required by Avogadro's hypothesis.

Nature of chemical forces

Between the presentation of Avogadro's hypothesis in 1811 and its general acceptance soon after 1860, several experimental techniques and theoretical laws were used by various investigators to yield different but self-consistent schemes of chemical formulas and atomic weights. After its acceptance, these schemes became unified. Within a few years of the development of another powerful technique, spectrum analysis by the German physicists Gustav Kirchhoff and Robert Bunsen in 1859, the number of chemical elements whose atomic weights and other properties were known had approximately doubled since the time of Avogadro's announcement. By relying fundamentally but not slavishly upon the determined atomic weight values and by using his chemical insight and intuition, the Russian chemist Dmitry Ivanovich Mendeleyev provided a classification scheme that ordered much of this burgeoning information and was a culmination of earlier attempts to represent the periodic repetition of certain chemical and physical properties of the elements.

20th-century trends. Development after the early 20th century present special problems to the historian, placing them beyond the scope of the present article. Recent science is relatively uncharted in comparison with early periods that have been ordered and reordered by successive generations of historians. It is consequently more difficult to discern the patterns of development of recent science, as well as the connections that relate it to philosophy, religion, and other areas of intellectual and social development. Furthermore, contemporary developments are still too close to read the patterns and trends that will undoubtedly emerge in the studies of historians of a later generation. All that can now be attempted are a few observations of the breakdown of the classical picture and the emergence of what appears to be a new view of the nature of the universe. In the middle of the 1890s a few totally unexpected phenomena were observed, which, together with the more conventional accumulations of knowledge, provided the impetus for a 20th-century transformation of vast realms of the physical sciences. William Röntgen's investigations in Germany of cathode rays, a subject thoroughly studied during the preceding half century, led to his discovery of X-rays, a completely unanticipated phenomenon; and, within a few weeks, in France, Henri Becquerel, following this line, discovered radioactivity in uranium, an element that also had been known for 50 years. In England, J.J. Thomson's much more traditional investigations of the nature of cathode rays led him in 1897 to a convincing demonstration (one of several at this time) of the existence of a subatomic particle, the electron.

In 1900 physicist Max Planck employed the novel idea of discrete energy levels for the vibrations of radiating substances in order to represent mathematically the observed distribution of radiant energy, a problem of some standing in thermodynamics and statistical mechanics. Quantum theory, thus introduced into physics, was accepted and exploited by Einstein, as in his photoelectric-effect equation of 1905. Einstein's theory, coupled with the electromagnetic wave theory of light, led to the view of light in terms of fundamental wave–particle duality. Niels Bohr, in Denmark, employed the quantum theory in 1913 in his version of the Rutherford model of the atom, and its subsequent application to the realm of the atom and the nucleus has transformed man's understanding of radiation and energy processes.

The acceptance of quantum mechanics and relativity theory marked the decline of the classical Newtonian picture of the world that dominated the physical sciences since the early 18th century. Modern physics replaced the absolute space–time reference frame and determinism of classical physics with relativity and uncertainty. It is no longer theoretically possible, as Laplace asserted, to predict all future states of the universe on the basis of thorough knowledge of the present position and momentum of each particle in the universe. Physical scientists now deny the very possibility of determining both the position and momentum of any single atomic particle with arbitrary precision. The breakdown of the classical framework, following from the dramatic discoveries and reformulations of the early 20th century, leaves science re-examining in depth the breadth and implications of its foundations.

BIBLIOGRAPHY. GEORGE SARTON, *History of Science*, 2 vol. (1952–59, reprinted 1970), two comprehensive volumes covering ancient science through the golden age of Greece and Hellenistic science and culture in the last three centuries BC; CHARLES SINGER, *A Short History of Scientific Ideas to 1900* (1959), a simple but comprehensive work outlining the development of the concepts of the physical worlds, all parts of which are rationally interrelated—treats not only physical and chemical but also biological disciplines; MAX JAMMER, *Concepts of Space*, 2nd ed. (1969), a scholarly yet reasonably popular work showing that the study of scientific thought within the framework of cultural setting produces a decisive influence on the modern mind; JAMES R. NEWMAN (ed.), *The World of Mathematics*, 4 vol. (1956), an indispensable manual for the student in history of mathematics; WILLIAM DAMPIER, *A History of Science and Its Relations with Philosophy and Religion*, 4th ed. rev. (1948, reprinted 1961), one of the earliest standard classics in the history of science; PIERRE DUHEM, *Le Système du monde: histoire des doctrines cosmologiques de Platon à Copernic*, 10 vol. (1913–59), famous classic work that presents in remarkable detail the history of the evolution of cosmological doctrines from Plato to Copernicus, and their bearing on man's ideas on space and force; THOMAS HEATH, *Aristarchus of Samos* (1913), a fascinating analysis of Aristarchus of Samos, a Copernicus of antiquity, who taught the earth's rotation on its axis and its anual revolution about the sun in the 3rd century BC; OTTO NEUGEBAUER, *The Exact Sciences in Antiquity*, 2nd ed. (1957, reprinted 1969), an account of mathematics and astronomy in antiquity, especially among the Egyptians and Babylonians; J.L.E. DREYER, *A History of Astronomy from Thales to Kepler*, 2nd ed. (1953), and ANTONIE PANNEKOEK, *De groei van ons wereld beeld* (1951; Eng. trans., *A History of Astronomy*, 1961), survey, between them, the history of astronomy from antiquity to the present; EDWARD GRANT, *Physical Sciences in the Middle Ages* (1971); and RICHARD S. WESTFALL, *The Construction of Modern Science: Mechanisms and Mechanics* (1971), two introductory surveys of the sciences in the Middle Ages and Scientific Revolution, which also contain useful and fairly comprehensive bibliographies of both primary and secondary sources; MARSHALL CLAGETT, *The Science of Mechanics in the Middle Ages* (1959), contains a good description of this subject supplemented by many important medieval texts; RICHARD S. WESTFALL, *Force in Newton's Physics* (1971), discusses the history of mechanics in the 17th century as well as Newton's thought; I. BERNARD COHEN, *Franklin and Newton* (1956), a good introduction to the Newtonian aspects of electricity and magnetism in the 18th century; RENE TATON (ed.), *La Science moderne de 1450 à 1800* (1958; Eng. trans., *The Beginnings of Modern Science, from 1450 to 1800*, 1964), and *La Science contemporaine*, pt. 1, *Le XIX^e Siècle* (1961; Eng. trans., *Science in the Nineteenth Century*, 1965), broad surveys of all aspects of the sciences of these periods that provide a useful introduction to the subject and contain extensive bibliographies for further reading; GALILEO GALILEI, *Discorsi e dimonstrazioni matematiche* (1638; Eng. trans., *Dialogues Concerning Two New Sciences*, 1914, reprinted 1963), perhaps greatest of the Renaissance scientific papers, these Dialogues summarized Galileo's early experiments and mature thoughts on the principles of mechanics; ISAAC NEWTON, *Principia* (1687; Eng. trans., *Sir Isaac Newton's Mathematical Principles of Natural Philosophy*, 1729; reprinted in 2 vol., 1962), Newton's famous classical work represents a milestone in the evolution

of physics and the foundation of the Newtonian world view; E.J. DIJKSTERHUIS, *Die Mechanisierung des Weltbildes* (1956; Eng. trans., *The Mechanization of the World Picture*, 1961), makes clear the relationship of ancient and medieval science and philosophy to the development of modern physical science; PAUL A. SCHILPP (ed.), *Albert Einstein, Philosopher Scientist*, 2nd ed. (1951), contains Einstein's autobiography and contributions from 25 scholars and scientists who discuss various aspects of Einstein's epoch-making achievements; ALBERT EINSTEIN and LEOPOLD INFELD, *The Evolution of Physics from Early Concepts to Relativity and Quanta* (1938, reprinted 1967), a well-written, clear, and penetrating account of the main stages in the evolution of modern physics; PHILIPP FRANK, *Einstein: His Life and Times*, rev. ed. (1953), a warm, human presentation of Einstein, and an explanation of the principle of relativity; WERNER HEISENBERG, *Physics and Philosophy* (1958), a thoughtful and excellent discussion of the problems arising from quantum mechanics by the originator of the uncertainty principle; ISAAC NEWTON, *Opticks*, 4th ed. (1730, reprinted 1952), a great classic of science that can be read with ease and enjoyment by the non-scientist; LOUIS DE BROGLIE, *La Physique nouvelle et les quanta* (1937; Eng. trans., *The Revolution in Physics*, 1953), story of the second scientific revolution, and *Physique et microphysique* (1947; Eng. trans., *Physics and Microphysics*, 1955), a discussion of the phenomena and their interpretation in the realm of the microcosmos.

(M.J.O./J.B.Sp.)

Physical Theories, Mathematical Aspects of

Practically all physical theories, from their basic concepts to their final formulations and predictive results, are essentially mathematical. Indeed, it may be said that the very language of theoretical physics is mathematics.

This article deals with some fundamental mathematical aspects of nine subdivisions of theoretical physics. An outline of the topics treated is as follows:

Mechanics of particles and systems
- Newton's laws
- Conservation of linear momentum
- Conservation of angular momentum
- Work and energy relations
- A system of particles

Fluid mechanics
- Three-dimensional flow
- Velocity and acceleration
- Equation of continuity
- Dynamic equation for an inviscid fluid
- Steady flow of an inviscid fluid
- Velocity potential and stream function

Mechanics of solids
- Elasticity; concept of stress
- Equilibrium conditions
- Tensile and compressive stress; triaxial and plane stress
- Concept of strain; Poisson's ratio
- Shear strain; general deformation; compatibility
- Relation between stress and strain
- Principal stresses and strains; invariants and strain energy

Statistical mechanics
- Basic definitions
- Probability
- Perfect gas pressure and energy density
- Maxwell-Boltzmann distribution
- Gibbs's ensembles
- The Hamiltonian
- Liouville's theorem
- Equations for ensemble probability densities
- Justification of the equations
- Quantum statistics

Electromagnetic theory
- Origin of the electromagnetic wave concept
- Maxwell's field theory
- Velocity of electromagnetic waves

Relativity theory
- Space
- Time

Riemannian geometry
- Basic properties
- Lorentz transformation and Lorentz invariance
- Lorentz contraction and time dilation

Quantum mechanics
- Transition from classical to quantum mechanics
- Wave mechanics
- Matrix form of quantum mechanics
- Transformation theory
- Statistical significance of the wave function
- Probability distribution in momentum space
- Relativistic quantum fields

Dimensional analysis
- General considerations
- The pi theorem
- Applications

MECHANICS OF PARTICLES AND SYSTEMS

Mechanics is the science concerned with the motion of bodies under the action of forces. This section is confined to basic, introductory considerations of classical, or Newtonian, mechanics and is distinct from relativistic mechanics (see RELATIVITY and the sections below *Relativity theory* and *Riemannian geometry*) and from quantum mechanics (see MECHANICS, QUANTUM and the section below *Quantum mechanics*).

A logical, mathematical development of mechanics requires that terminology must first be specified. Some terms (*e.g.*, position, time, and mass) are taken as primitive or undefined. Others may be defined by using primitive or previously defined terms: for example, velocity may be defined as time rate of change of position; acceleration as time rate of change of velocity; (linear) momentum as the product of mass times velocity. Once the terminology has been specified, a suitable set of initial assumptions or postulates is chosen from which the theorems of the theory can be deduced. Newton's three laws (see below) may be taken as such postulates; other initial assumptions are also possible.

Newton's laws. Historically, although some of the principles of mechanics were known to the ancient Greeks, the beginning of this discipline as an exact science is generally associated with the work of Galileo, who was born in 1564 and died in 1642. By brilliantly clarifying, summarizing, and extending this work, Isaac Newton, born the year that Galileo died, succeeded in formulating his three laws of motion. In modern terms, the first law states that in the absence of an external force, a particle initially at rest will remain at rest; or, if initially moving with velocity v, the particle will continue to do so with constant speed along a straight line.

The second law asserts that if an external force does act on a particle, then the time rate of change of its momentum (product of mass times velocity) is equal to (and in the same direction as) the external force. In the special but frequently occurring case of constant (unvarying) mass, the second law reduces (as shown below *Conservation of linear momentum*) to the assertion that force equals mass times acceleration. Newton, however, enunciated this second law in the first of these two forms, thus allowing for the variability of mass, well over two centuries before this property became an important physical feature in the study of rocket motion and, in a more profound manner, in Einstein's theory of relativity.

Newton's third law states that to every action, there is an equal and opposite reaction: If particle A exerts a force (action) on particle B, then particle B in turn will exert on particle A a force (reaction) such that this latter force is equal in magnitude but opposite in direction to the former. Action and reaction

From the definitions of velocity and acceleration given above, it is apparent that formulating and solving problems by means of Newton's second law require mathematical techniques for dealing with rates of change. Such techniques, along with appropriate notation, the calculus, were developed by Newton and by G.W. Leibniz (see ANALYSIS, REAL).

$$(1) \quad f = ma$$

In the usual mathematical formulation of Newton's second law for a particle of constant mass m, f is the force acting on it and a is its acceleration; f is equal to the product m times a (see Box, equation 1). Acceleration is a vector, a quantity specified by both mag-

nitude and direction; mass is a scalar, a quantity without an associated direction; and the product ma is interpreted as a vector quantity in the same direction as the vector a and of magnitude m times the magnitude of the vector a. Newton's second law then implies that the force is a vector f in the same direction and of the same magnitude as ma. This law may also be expressed in a scalar form by writing the equation in terms of components; *i.e.*, the magnitudes of the quantities along the axes OX, OY, OZ in an orthogonal (axes mutually at right angles) coordinate system. In this form, three equalities relate components of f, components of a, and m (see 2). In the case of a particle confined to a line (or plane), only one or two of these component equations are relevant.

For a particle not acted on by any force, the force is set equal to zero in Newton's second law; this implies that acceleration is zero. Recalling that acceleration is defined as the time rate of change of velocity, zero acceleration implies constant velocity, unvarying in magnitude and direction. Hence, Newton's first law is but a special case of the second law.

If the force components are known as functions of the time t, and perhaps also of the coordinates x, y, z, and of the velocity components v_x, v_y, and v_z, then Newton's second law expressed in component form becomes three second-order differential equations (so called because the acceleration components are the second derivatives of the coordinates) to be solved for $x(t)$, $y(t)$, $z(t)$. Given x, y, z and v_x, v_y, v_z at an instant t, the equations in terms of components determine the acceleration. If position, velocity, and acceleration are known at t, the change in position during a brief interval, denoted as δt, can be computed approximately by multiplying δt by the rate of change of position (*i.e.*, the velocity); similarly, the change in velocity during the interval is the product of δt times the rate of change of velocity (*i.e.*, the acceleration). It is thus plausible that the component equations (see 2) permit the motion to be traced from a position and velocity specified at some instant.

If a particle interacts with (*i.e.*, is attracted or repelled by) one or more bodies, perhaps in several ways (*e.g.*, electrically and gravitationally), it may be subject to several forces at the same time. If any one force f were acting alone, it would produce an acceleration given by Newton's second law (see 1). Because accelerations are added vectorially (in which directions as well as magnitudes are combined), the equation implies that the forces are to be combined in the same way to specify the total force f related to the total acceleration a.

For two or more interacting particles, Newton's second law is to be written separately for each particle. If any force f exerted by a particle A on a particle B appears in the equation of motion for B, then Newton's third law implies that a corresponding force $-f$ exerted by B on A must appear in A's equation of motion. Here f is a vector; $-f$ means a vector of equal magnitude, but oppositely directed.

When dealing with systems containing more than one particle, it becomes of paramount importance to be quite clear as to which particle or particles belong to any one specific system. For example, the system may be defined by means of a suitable bounding surface. It will be necessary to distinguish between internal forces, those that arise from interactions among the system's particles, and external forces that stem from interactions with particles or bodies not included in the system. The particles with masses m_1, m_2, $\cdots$ have positions located by coordinates (x_1, y_1, z_1), (x_2, y_2, z_2), $\cdots$, respectively. If f_1 and f'_1, respectively, denote the total external and internal forces on particle 1, then its equations of motion, relative to a chosen set of axes $OXYZ$, are expressed in a system of equations involving components of f, components of f', and components of a, each involving m (see 3). Similar sets of equations may be written for the other particles, 2, 3, $\cdots$, comprising the system.

Internal and external forces

Conservation of linear momentum. In the general form of Newton's second law, force is equal to the time rate of change of momentum (mv), which in differential form is written $f = \frac{d}{dt}(mv)$. For constant mass this may be rewritten as $f = mdv/dt$. Because acceleration a is the time rate of change of velocity, the x-component of the total force is expressible as a differential equation (see 4). If all such equations for the x-motion of all the particles are added, and similarly for the y- and z-motions, then a system of equations is obtained (see 5).

$$(2)\qquad f_x = ma_x, \qquad f_y = ma_y, \qquad f_z = ma_z$$

$$(3)\qquad \begin{cases} (f_x)_1 + (f'_x)_1 = m_1(a_x)_1 \\ (f_y)_1 + (f'_y)_1 = m_1(a_y)_1 \\ (f_z)_1 + (f'_z)_1 = m_1(a_z)_1 \end{cases}$$

$$(4)\qquad f_x + f'_x = ma_x = \frac{d}{dt}(mv_x)$$

$$(5)\qquad \begin{cases} \sum \frac{d}{dt} mv_x = \Sigma f_x + \Sigma f'_x \\ \sum \frac{d}{dt} mv_y = \Sigma f_y + \Sigma f'_y \\ \sum \frac{d}{dt} mv_z = \Sigma f_z + \Sigma f'_z \end{cases}$$

$$(6)\qquad \Sigma f'_x = 0; \qquad \Sigma f'_y = 0; \qquad \Sigma f'_z = 0$$

$$(7)\qquad \begin{cases} \frac{d}{dt} \Sigma mv_x = \Sigma f_x \\ \frac{d}{dt} \Sigma mv_y = \Sigma f_y \\ \frac{d}{dt} \Sigma mv_z = \Sigma f_z \end{cases}$$

Newton's third law asserts that for every action there is an equal and opposite reaction. This means that internal forces f' between particles exist in equal and opposite pairs, so that the sum of their components in any direction must vanish (see 6). This result simplifies the system of equations obtained (see 5) so that they state that the time rate of change of each component of the total linear momentum is equal to the corresponding component of the total external force (see 7).

If the total external force (or its component in any direction) is zero, then the final system obtained implies that the total linear momentum (or its component in that direction) will remain constant. This is known as the principle of conservation of linear momentum. One important consequence of this principle may be noted immediately: If two bodies collide, then the sum of their linear momenta remains constant. This is so because their interaction can only cause equal and opposite forces to act between pairs of particles. These forces, being internal to the system comprising the two bodies, can therefore produce no increment of total momentum.

Conservation of angular momentum. Two new concepts are needed at this point: torque (also called moment of force) and angular momentum (also called moment of momentum). Both are vector quantities associated with rotational motion of a mass and are related through Newton's second law as follows. For a system having a single particle, the first and second of equations (5) when multiplied by y and x, respectively, give two differential equations, in one of which y is a multiplicative factor common to all terms, in the other of which x is a factor common to all terms (see 8).

Rotational motion

Subtracting the first of these equations from the second yields an equality in which there occurs a difference between two terms, each involving derivatives (see 9).

There is an equation in which the derivative with respect to t operates upon a difference of two terms and upon factors in two other terms (see 10). The first and third terms on the right-hand side here cancel, so that there occurs an equality between terms that involve the derivative with respect to t (see 11) and hence equation (9) can be rewritten as an equation in which the derivative with respect to t operates upon a difference of two terms and a symmetry occurs in the right-hand side (see 12). The expression inside the left-hand parentheses is called the z-component of angular momentum L, and the entire expression on the right-hand side is called the z-component of the torque N. Two more equations of this type can be written, because the same procedure may be applied to obtain also the x- and y-components of L and N.

If the system is comprised of more than one particle, equation (12), after rearrangement, may by written for each particle and added to give relations in the three planes xy, xz, and yz about the respective axes z, y, and x (see 13). A sum involving differences of components as well as x and y (see 14) represents the sum of the torques about the z-axis caused by the internal forces of the particle aggregate. These forces, however, occur in pairs of which the members are equal in magnitude, collinear (lying on or passing through the same straight line), and opposite in direction. The torque caused by one member of such a pair is thus equal and opposite to that of the other, so that the total effect of the sum is zero. The above equations can therefore be reduced to the simpler form in which all internal forces are absent: three equations in components, each involving two major terms of differences (see 15). Equations (15) state that the rate of change of the total angular momentum about any fixed axis is equal to the sum of the torques due to external forces about that axis.

The principle of conservation of angular momentum can now be stated as follows. If the sum of the torques resulting from external forces about a fixed axis is zero, then the total angular momentum of the aggregate of particles about that axis is constant.

Work and energy relations. The work performed by a force f_x, acting on a particle of mass m, moving from an initial position x_A to a final position x_B, is defined by the integral from x_A to x_B of f_x (see 16). From the definition of acceleration as the time rate of change of velocity, it follows that a_x is shown through calculation to be one-half the x-derivative of the square of v_x (see 17), so that by equations (2) f_x is one-half the product of m and the derivative with respect to x of the square of v_x (see 18). Because m is constant, this may also be written so that f_x is equal to the derivative with respect to x of an expression involving m and v_x (see 19), and hence the integral becomes the integral from x_A to x_B of f_x evaluated as a difference of two terms (see 20).

For motion along the y- and z-axis, respectively, one finds by a precisely analogous procedure two integrals of the above type, one involving f_y, the other involving f_z (see 21). Because v_x, v_y, v_z are mutually perpendicular components of the velocity vector v, the magnitude of which will be denoted by s, it follows from the Pythagorean theorem of the right triangle that the square of s is equal to the sum of the squares of components of v, and the integral from A to B of products of the type $f_y dy$ is shown by calculation to equal the increment (denoted Δ) of the quantity one-half the product of m and the square of s (see 22).

Kinetic energy

The quantity $ms^2/2$ is called the kinetic energy of the particle. Equation (22) therefore asserts that the work done as the particle moves from the point A having position coordinates x_A, y_A, and z_A, or $A(x_A, y_A, z_A)$, to the point $B(x_B, y_B, z_B)$ is equal to the change in the kinetic energy of the particle, namely, $\Delta(ms^2/2)$. If that particle is one of an aggregate, the forces acting on it will in general be both external and internal. It follows from equation (22) that the change in the particle's kinetic energy is equal to the total work of the external and internal forces during the displacements in the x, y, and z directions so that appropriate limits are implied on the integrals; or, Δ of one-half the product of m and the square of s is equal to a sum of three integrals, integrated with respect to the differentials of the three coordinates (see 23).

(8) $$\begin{cases} y\dfrac{d}{dt}mv_x = yf_x + yf'_x \\ x\dfrac{d}{dt}mv_y = xf_y + xf'_y \end{cases}$$

(9) $$x\frac{d}{dt}mv_y - y\frac{d}{dt}mv_x = xf_y + xf'_y - yf_x - yf'_x$$

(10) $$\frac{d}{dt}(mxv_y - ymv_x) = mv_xv_y + x\frac{d}{dt}mv_y - mv_yv_x - y\frac{d}{dt}mv_x$$

(11) $$x\frac{d}{dt}mv_y - y\frac{d}{dt}mv_x = \frac{d}{dt}(mxv_y - ymv_x)$$

(12) $$\frac{d}{dt}(xmv_y - ymv_x) = x(f_y + f'_y) - y(f_x + f'_x)$$

(13) $$\begin{cases} \sum\dfrac{d}{dt}(xmv_y - ymv_x) = \Sigma(xf_y - yf_x) + \Sigma(xf'_y - yf'_x) \\ \sum\dfrac{d}{dt}(ymv_z - zmv_y) = \Sigma(yf_z - zf_y) + \Sigma(yf'_z - zf'_y) \\ \sum\dfrac{d}{dt}(zmv_x - xmv_z) = \Sigma(zf_x - xf_z) + \Sigma(zf'_x - xf'_z) \end{cases}$$

(14) $$\Sigma(xf'_y - yf'_x)$$

(15) $$\begin{cases} \dfrac{d}{dt}\sum m(xv_y - yv_x) = \Sigma(xf_y - yf_x) \\ \dfrac{d}{dt}\sum m(yv_z - zv_y) = \Sigma(yf_z - zf_y) \\ \dfrac{d}{dt}\sum m(zv_x - xv_z) = \Sigma(zf_x - xf_z) \end{cases}$$

(16) $$\int_{x_A}^{x_B} f_x\,dx$$

(17) $$a_x = \frac{dv_x}{dt} = \frac{dv_x}{dx}\frac{dx}{dt} = v_x\frac{dv_x}{dx} = \tfrac{1}{2}\frac{d}{dx}(v_x^2)$$

(18) $$f_x = \tfrac{1}{2}m\frac{d}{dx}(v_x^2)$$

(19) $$f_x = \frac{d}{dx}(\tfrac{1}{2}mv_x^2)$$

(20) $$\int_{x_A}^{x_B} f_x\,dx = (\tfrac{1}{2}mv_x^2)_B - (\tfrac{1}{2}mv_x^2)_A$$

(21) $$\begin{cases} \displaystyle\int_{y_A}^{y_B} f_y\,dy = (\tfrac{1}{2}mv_y^2)_B - (\tfrac{1}{2}mv_y^2)_A \\ \displaystyle\int_{z_A}^{z_B} f_z\,dz = (\tfrac{1}{2}mv_z^2)_B - (\tfrac{1}{2}mv_z^2)_A \end{cases}$$

(22) $$\begin{cases} s^2 = v_x^2 + v_y^2 + v_z^2 \\ \displaystyle\int_A^B (f_x\,dx + f_y\,dy + f_z\,dz) = (\tfrac{1}{2}ms^2)_B - (\tfrac{1}{2}ms^2)_A \\ \qquad\qquad \equiv \Delta(\tfrac{1}{2}ms^2) \end{cases}$$

Writing one such equation for each of the particles comprising the system, and then adding these equations, gives Δ of a sum extended over one-half the product of m and the square of s equal to a sum of the integrals of the above type (see 24). That is, the change in the total kinetic energy of all the particles is equal to the work done by all the forces as the system moves from one configuration to another. The concept of work done by internal and external forces is of fundamental importance in all branches of applied mechanics. An important extension of this concept follows.

A system of particles. If the total mass of a system of particles $\Sigma m = m_1 + m_2 + \cdots$ is denoted by M; then a quantity $\bar{x}$ may be defined such that M times $\bar{x}$ is equal to a sum of products of the mass of each particle times its x-coordinate (see 25). Then $\bar{x}$ can be expressed as the sum over mx divided by M; $\bar{y}$ and $\bar{z}$ can be expressed similarly (see 26). Differentiating equations (26) with respect to time gives three equalities relating M, derivatives of x, y, and z, as well as m and components of v (see 27); that is, the total momentum of the system in any of the directions OX, OY, OZ is precisely that of a hypothetical particle of mass M (the total mass) located at the position defined by the coordinates $(\bar{x}, \bar{y}, \bar{z})$. This position is called the centre of mass of the system, a concept of great utility in a number of applications of mechanics.

Centre of mass

Differentiating the preceding equations again with respect to time gives three equations of second order involving M, x, y and z, m, t and components of a (see 28), which in turn permit previous equations to be rewritten more compactly. To do this, the sums of all the external forces acting on the system are denoted by three components of F that are equal to three sums of components of f (see 29), so that, by equations (28), three equations relating components F to M and second derivatives with respect to t of x, y, and z are obtained (see 30). The idea of the total mass of the system as being located at its centre of mass can be taken further: equations (30) are precisely those of the hypothetical particle when it is imagined as being acted on by the resultant of all the external forces.

One important deduction follows immediately: if a system is acted on by external forces such that the sums of their components in the directions OX, OY, and OZ are all zero, then the centre of mass of that system must either be at rest or moving with uniform velocity in accordance with Newton's first law. A special case of this arises when a system of particles is entirely free from external forces.

Because the centre of mass is defined by the coordinates x, y, z, the coordinates of any particle within the system can be measured relative to that point. For instance, the position of particle 1 may be specified by the coordinates either x_1, y_1, z_1, or $\bar{x} + \xi_1$, $\bar{y} + \eta_1$, $\bar{z} + \zeta_1$. That is, if a set of axes, parallel to OX, OY, OZ, were erected with the centre of mass as origin, then the corresponding coordinates of 1 would be ξ_1, η_1, ζ_1. From the definitions of x, y, z, it follows that each of the quantities $\Sigma m\xi$, $\Sigma m\eta$, $\Sigma m\zeta$, together with their derivatives, must be zero.

If $\bar{x} + \xi$, $\bar{y} + \eta$ and $\bar{z} + \zeta$ are substituted for x, y, and z in the expressions for the angular momenta about the fixed axes, three equations result involving differences of terms that include derivatives as factors (see 31) for the axes OX, OY, and OZ, respectively.

Consideration of the first of these results shows that the quantity M times a difference between products of the type x times the derivative of y with respect to t (see 32) is the angular momentum about OZ of a single particle of mass equal to the total mass and located at the centre of mass. Again, by analogy with the expression on the left-hand side of equations (31), a summation of terms involving m and differences of products, each product containing a derivative with respect to t (see 33), and the corresponding expressions in the latter two equations (31) are the angular momenta calculated relative to a coordinate system, designated arbitrarily as $O\Xi HZ$, having origin O at the centre of mass (and moving with it) and axes $O\Xi$, OH, and OZ that remain parallel to the corresponding axes of the system $OXYZ$. If the system of particles comprises a single body (not necessarily rigid or even solid), then the motion of the centre of mass, given by $\bar{x}(t), \bar{y}(t), \bar{z}(t)$, may be called the orbital motion; and the motion represented by $\xi_1, \eta_1, \zeta_1; \xi_2, \cdots$ may be referred to as the internal motion. The above argument thus shows that the angular momentum can be resolved into two parts: an orbital part, and a part associated with the internal motion. The latter part is often called the spin.

Orbital motion and spin

If the new expressions $\bar{x} + \xi$, etc., are substituted for the x, y, and z coordinates in a previous equation (see 15), then, in view of the foregoing remarks, these equations reduce to three differential equations in which summation extended over differences occurs (see 34). These

(23) $$\Delta(\tfrac{1}{2}ms^2) = \int(f_x + f'_x)\,dx + \int(f_y + f'_y)\,dy + \int(f_z + f'_z)\,dz$$

(24) $$\Delta\Sigma(\tfrac{1}{2}ms^2) = \Sigma[\int(f_x + f'_x)\,dx + \int(f_y + f'_y)\,dy + \int(f_z + f'_z)\,dz]$$

(25) $$M\bar{x} = m_1x_1 + m_2x_2 + \cdots = \Sigma mx$$

(26) $$\bar{x} = \frac{\Sigma mx}{M}, \qquad \bar{y} = \frac{\Sigma my}{M}, \qquad \bar{z} = \frac{\Sigma mz}{M}$$

(27) $$\begin{cases} M\dfrac{d\bar{x}}{dt} = \sum m\dfrac{dx}{dt} = \Sigma mv_x \\ M\dfrac{d\bar{y}}{dt} = \sum m\dfrac{dy}{dt} = \Sigma mv_y \\ M\dfrac{d\bar{z}}{dt} = \sum m\dfrac{dz}{dt} = \Sigma mv_z \end{cases}$$

(28) $$\begin{cases} M\dfrac{d^2\bar{x}}{dt^2} = \sum m\dfrac{d^2x}{dt^2} = \Sigma ma_x \\ M\dfrac{d^2\bar{y}}{dt^2} = \sum m\dfrac{d^2y}{dt^2} = \Sigma ma_y \\ M\dfrac{d^2\bar{z}}{dt^2} = \sum m\dfrac{d^2z}{dt^2} = \Sigma ma_z \end{cases}$$

(29) $$F_x = \Sigma f_x, \qquad F_y = \Sigma f_y, \qquad F_z = \Sigma f_z$$

(30) $$F_x = M\frac{d^2\bar{x}}{dt^2}, \qquad F_y = M\frac{d^2\bar{y}}{dt^2}, \qquad F_z = M\frac{d^2\bar{z}}{dt^2}$$

(31) $$\begin{cases} \Sigma m(xv_y - yv_x) = M\left(\bar{x}\dfrac{d\bar{y}}{dt} - \bar{y}\dfrac{d\bar{x}}{dt}\right) + \Sigma m\left(\xi\dfrac{d\eta}{dt} - \eta\dfrac{d\xi}{dt}\right) \\ \Sigma m(yv_z - xv_y) = M\left(\bar{y}\dfrac{d\bar{z}}{dt} - \bar{z}\dfrac{d\bar{y}}{dt}\right) + \Sigma m\left(\eta\dfrac{d\zeta}{dt} - \zeta\dfrac{d\eta}{dt}\right) \\ \Sigma m(zv_x - xv_z) = M\left(\bar{z}\dfrac{d\bar{x}}{dt} - \bar{x}\dfrac{d\bar{z}}{dt}\right) + \Sigma m\left(\zeta\dfrac{d\xi}{dt} - \xi\dfrac{d\zeta}{dt}\right) \end{cases}$$

(32) $$M\left(\bar{x}\frac{d\bar{y}}{dt} - \bar{y}\frac{d\bar{x}}{dt}\right)$$

(33) $$\sum m\left(\xi\frac{d\eta}{dt} - \eta\frac{d\xi}{dt}\right)$$

(34) $$\begin{cases} \dfrac{d}{dt}\sum m\left(\xi\dfrac{d\eta}{dt} - \eta\dfrac{d\xi}{dt}\right) = \Sigma(\xi f_y - \eta f_x) \\ \dfrac{d}{dt}\sum m\left(\eta\dfrac{d\zeta}{dt} - \zeta\dfrac{d\eta}{dt}\right) = \Sigma(\eta f_x - \zeta f_y) \\ \dfrac{d}{dt}\sum m\left(\zeta\dfrac{d\xi}{dt} - \xi\dfrac{d\zeta}{dt}\right) = \Sigma(\zeta f_x - \xi f_z) \end{cases}$$

equations state that the rate of change of any component of the spin angular momentum is equal to the total torque produced by the external forces about the corresponding axis (through the centre of mass).

In a similar manner, the kinetic energy of a system of particles can be resolved into an orbital and an internal part. The kinetic energy T of the aggregate is equal to a summation, there being three squared derivatives with respect to t in the summation (see 35). If this expression is now simplified by means of the method described, then it is found that the expression for the kinetic energy reduces to a sum of two terms, each involving squares of derivatives with respect to t (see 36). The kinetic energy of a system of particles is therefore seen to be equal to that of the hypothetical particle of mass M located at the centre of mass, plus the kinetic energy of all the particles as it would be calculated by an observer moving with the centre of mass.

(Ed.)

FLUID MECHANICS

Fluid mechanics is the science concerned specifically with the motion of fluids under the action of forces. A fluid is any substance that flows, or more precisely, one that cannot sustain shear forces (see below) when in static equilibrium. Thus, two parallel but oppositely directed forces acting on a pair of plates joined by a solid pin are said to be shearing forces, because they tend to produce a sliding motion of one plate with respect to the other. If these forces are not excessive, however, the (solid) pin will be able to sustain these stresses, and no motion on the part of the plates will occur. If a fluid, on the other hand, were to replace the pin, then, regardless how small the shearing forces were, the fluid could not sustain them: static equilibrium would be impossible, and the plates would move.

Gases and liquids are two forms of the fluid state; but whereas gases are generally highly compressible, liquids are often nearly incompressible. Furthermore, although gases and liquids are composed of molecules, it is customary in this discipline to adopt a macroscopic (large-scale) point of view and thus to treat fluids as continuous media. (The microscopic point of view, taking the discrete structure of matter into account, is adopted in the section below *Statistical mechanics*.)

Three-dimensional flow. Although one- or two-dimensional descriptions are adequate for certain types of flow, a full three-dimensional formulation is often necessary. In one method, developed in the 18th century by the Swiss-born mathematician Leonhard Euler, such motion is described in terms of velocity- and pressure-distributions and other similar characteristics in the fluid. To facilitate the application of Euler's method, at any point in the fluid having coordinates x, y, z, the velocity V of the fluid may be resolved into three components denoted u, v, and w in the directions parallel to the x, y, and z axes, respectively, and each of these velocity components may be expressed in terms of the three coordinates and the time t (see 37).

The first of equations (37) asserts that the velocity component u is some function f_1 of the space coordinates x, y, z, and of the time t. Specification of any set of values for these four quantities determines the corresponding value of u; similarly, the functions f_2 and f_3 determine values of v and w. Expressions analogous to equations (37) may be written for acceleration, pressure, temperature, and other fluid characteristics. If, at each point in a given region, the values of all of these characteristics do not change with time, then the flow is said to be steady. Hence in the special but frequently occurring case of steady flow, the time t will not appear in f_1, f_2, f_3, thus rendering the velocity V time-independent, as required.

Steady flow

Velocity and acceleration. In the most general case of fluid motion, however, the velocity V may be a function of both position, such as the distance s along a streamline, and of the time t. That is, V is a function of s and t (see 38). (A streamline may be defined operationally as a curve C constructed in a manner such that the tangent at any point along C is in the direction of the fluid velocity at that point.)

The velocity V may thus change from point to point in space at any fixed instant in time and also from instant to instant in time at any fixed point in space. The total differential of V, therefore, is composed of two terms, each dependent on space or time (see 39). It is to be noted that the derivative in each term appearing on the right-hand side of equation (39) is partial; *i.e.*, is a derivative with respect to one variable only, all others held constant for that particular differentiation.

From the definition of acceleration, a, it follows that the derivative of V with respect to t is equal to a sum of terms involving partial derivatives of V with respect to s and t (see 40). The quantity $\partial V/\partial t$ is often called the local acceleration, and the term $V\partial V/\partial s$ is called the convective acceleration. The former represents the time-rate of change in the local velocity at a fixed point. For steady flow, $\partial V/\partial t = 0$.

The components of the total acceleration $a = dV/dt$ in the x-, y-, and z-directions are, respectively, du/dt, dv/dt, and dw/dt; they may be obtained by differentiating equations (37), yielding three equations identifying total derivatives for u, v, and w with expressions that involve partial derivatives (see 41).

Equation of continuity. An infinitesimal parallelepiped may be imagined embedded in a fluid; the sides have the lengths dx, dy, dz. Considering first the mass flow with velocity u in the x-direction, it is deduced that, per unit time, the mass of fluid (having a density denoted by ρ) entering the parallelepiped in the x-direction is $\rho u dydz$. In traversing the distance dx, the quantity ρu may change by an amount $(\partial/\partial x)(\rho u)$. The time-rate of mass flow leaving the opposite face is therefore a factor multiplied by a differential of y and the differential of z, the factor being composed of a sum of two terms involving u, ρ, and the differential of x (see 42). Hence,

(35) $$\sum \tfrac{1}{2}m\left\{\left[\frac{d}{dt}(\bar{x}+\xi)\right]^2+\left[\frac{d}{dt}(\bar{y}+\eta)\right]^2+\left[\frac{d}{dt}(\bar{z}+\zeta)\right]^2\right\}=T$$

(36) $$T=\tfrac{1}{2}M\left[\left(\frac{d\bar{x}}{dt}\right)^2+\left(\frac{d\bar{y}}{dt}\right)^2+\left(\frac{d\bar{z}}{dt}\right)^2\right]+ +\sum \tfrac{1}{2}m\left[\left(\frac{d\xi}{dt}\right)^2+\left(\frac{d\eta}{dt}\right)^2+\left(\frac{d\zeta}{dt}\right)^2\right]$$

(37) $$\begin{cases} u=f_1(x, y, z, t) \\ v=f_2(x, y, z, t) \\ w=f_3(x, y, z, t) \end{cases}$$

(38) $$V=f(s, t)$$

(39) $$dV=\frac{\partial V}{\partial s}ds+\frac{\partial V}{\partial t}dt$$

(40) $$\frac{dV}{dt}=\frac{\partial V}{\partial s}\frac{ds}{dt}+\frac{\partial V}{\partial t}=V\frac{\partial V}{\partial s}+\frac{\partial V}{\partial t}$$

(41) $$\begin{cases} \dfrac{du}{dt}=u\dfrac{\partial u}{\partial x}+v\dfrac{\partial u}{\partial y}+w\dfrac{\partial u}{\partial z}+\dfrac{\partial u}{\partial t} \\ \dfrac{dv}{dt}=u\dfrac{\partial v}{\partial x}+v\dfrac{\partial v}{\partial y}+w\dfrac{\partial v}{\partial z}+\dfrac{\partial v}{\partial t} \\ \dfrac{dw}{dt}=u\dfrac{\partial w}{\partial x}+v\dfrac{\partial w}{\partial y}+w\dfrac{\partial w}{\partial z}+\dfrac{\partial w}{\partial t} \end{cases}$$

(42) $$\left[\rho u+\frac{\partial}{\partial x}(\rho u)\,dx\right]dydz$$

(43) $$\frac{\partial}{\partial x}(\rho u)\,dxdydz$$

the net outflow in the x-direction, per unit time, is a product of the differentials of x, y, and z and of the partial derivative, with respect to x, of the product of ρ and u (see 43). Combining this result with those of analogous considerations of fluid transport in the y- and z-directions shows that the net outflow rate caused by variations in space alone is given by a term composed of the product of three differentials and a factor that is a sum of partial derivatives (see 44). To cover the most general case, however, allowance must also be made for a possible variation (of mass inside the parallelepiped) with time; *i.e.*, for a change by an amount that is the partial derivative of ρ with respect to t, weighted with the product of the three differentials (see 45). The principle of the conservation of mass asserts that matter can neither be created nor destroyed. Accordingly, the overall net gain—*i.e.*, the sum of the preceding two expressions—must vanish, giving an equation restricting a sum of four partial derivatives to equal zero (see 46) or, transposing the first term above to the right, a restriction that equates the sum of four partial derivatives with the negative partial derivative ρ with respect to t (see 47). The sum of three terms on the left-hand side of equation (47) is called the divergence (div) of the fluid current density, ρV, and is generally written as $\text{div}(\rho V)$. Hence, the divergence of the product of ρ and V equals the negative of the partial derivative of ρ with respect to t (see 48). This result, known as the equation of continuity, is of fundamental significance, not only in the present context but throughout many other branches of science. It states that if $\text{div}(\rho V)$ on the left of equation (48) is a positive quantity and thus signifying a net outflow (from the parallelepiped), then $\partial\rho/\partial t$ must be negative (so as to make the right-hand member also positive). That is, exact balance is maintained between the net outflow and the decrease in density, and hence mass, within the parallelepiped. The converse also holds: If the mass, and hence the density, within the parallelepiped is to increase ($\partial\rho/\partial t$ positive, right-hand side negative), then this gain must be accompanied by a net inflow, or negative $\text{div}(\rho V)$, again in exact balance, as prescribed by equation (48). If the flow is steady, the divergence of the products of ρ and V is zero (see 49) and asserts that if there is to be no change in the fluid density or mass within the parallelepiped, then the net outflow or inflow, that is, $\pm\text{div}(\rho V)$, must equal zero. Again, equation (49) also states the converse: If there is no net flow out of or into the region $\pm\text{div}(\rho V) = 0$, then there can be no change in the fluid density or mass there. Finally, in the special but frequently encountered case of an incompressible fluid, equation (49) simplifies to the divergence of V, which is zero (see 50).

Meaning of divergence

Dynamic equation for an inviscid fluid. The preceding considerations have dealt mainly with the kinematics (*i.e.*, with the geometrical aspects) of fluid motion. Now attention is focussed on fluid dynamics; that is, on the interrelation between a specified set of forces and the resultant fluid motion. For the sake of clarity, only the case of an inviscid fluid—*i.e.*, one free from viscosity (internal friction or resistance to flow)—will be treated here. An infinitesimal parallelepiped of sides dx, dy, dz can be thought to be embedded in a fluid. Examining first the pressures exerted on this mass of fluid along the x-direction only, it is reasoned that if the pressure on the left-hand face is p, then that on the right-hand face is $-(p + \frac{\partial p}{\partial x}dx)$, the minus sign indicating that this pressure is exerted along the negative x-direction. Inasmuch as pressure is defined as force f_p per unit area normal (perpendicular) to f_p ($p = f_p/A$, $A =$ area), it follows that $f_p = pA$. Using subscripts L and R to indicate quantities to be evaluated at the left- and right-hand faces, respectively, the resultant pressure force in the $+x$-direction is found to be the x component of f_p expressed in terms of the partial derivative of p with respect to x (see 51), and in an exactly analogous manner it is found that the y component of f with respect to p and the z component have similar representations (see 52). If, in addition to the pressure, the fluid is also subject to an external pressure f_b per unit volume of fluid, then the additional force exerted on the mass contained in the parallelepiped is $f_b dxdydz$. This is called a body force, and f_b is called the body-force density. (A common example of a body force is that caused by gravity, or $f_b = \rho g$, in which g is the acceleration of gravity.) If P, Q, R, respectively, denote the x, y, and z components of f_b, then Newton's second law of motion—which states that net force is equal to mass times acceleration—may be invoked to formulate an inequality between, on the one hand, the difference between P and the partial derivative of p with respect to x and, on the other hand, a product of ρ and the derivative of u with respect to t (see 53), in which the acceleration is the time-rate of change of velocity, du/dt. Upon cancellation of the infinitesimal volume $dxdydz$ appearing on both sides of this equation, and applying corresponding considerations to the y and z components, it is found that three equations relate P, Q, R with partial derivatives of p, u, v, and w, ρ being a common factor in each (see 54). These three relations, sometimes called Euler's hydrodynamic equations, can be expressed more concisely by the single vector equation that identifies the difference between f_b and the gradient of p with the product of ρ and the derivative of V with respect to t (see 55) in which the vector symbol "grad" stands for the gradient of, and has components $\partial/\partial x$, $\partial/\partial y$, $\partial/\partial z$, respectively. If desired, the components of acceleration (du/dt, etc.) appearing on the right-hand side of equation (54) may be expressed in terms of the corresponding quantities given in equations (41).

Euler's hydrodynamic equations

(44) $$\left[\frac{\partial}{\partial x}(\rho u)+\frac{\partial}{\partial y}(\rho v)+\frac{\partial}{\partial z}(\rho w)\right]dxdydz$$

(45) $$\frac{\partial}{\partial t}(\rho)\,dxdydz$$

(46) $$\frac{\partial\rho}{\partial t}+\frac{\partial}{\partial x}(\rho u)+\frac{\partial}{\partial y}(\rho v)+\frac{\partial}{\partial z}(\rho w)=0$$

(47) $$\frac{\partial}{\partial x}(\rho u)+\frac{\partial}{\partial y}(\rho v)+\frac{\partial}{\partial z}(\rho w)=-\frac{\partial\rho}{\partial t}$$

(48) $$\text{div}(\rho V)=-\partial\rho/\partial t$$

(49) $$\text{div}(\rho V)=0$$

(50) $$\text{div}\,V=\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}=0$$

(51) $$(f_p)_x=p_L dydz+p_R dydz=-\frac{\partial p}{\partial x}dxdydz$$

(52) $$(f_p)_y=-\frac{\partial p}{\partial y}dxdydz,\quad\text{and}\quad(f_p)_z=-\frac{\partial p}{\partial z}dxdydz$$

(53) $$\left(P-\frac{\partial p}{\partial x}\right)dxdydz=\rho\frac{du}{dt}dxdydz$$

(54) $$\begin{cases}P-\dfrac{\partial p}{\partial x}=\rho\dfrac{du}{dt}\\ Q-\dfrac{\partial p}{\partial y}=\rho\dfrac{dv}{dt}\\ R-\dfrac{\partial p}{\partial z}=\rho\dfrac{dw}{dt}\end{cases}$$

(55) $$f_b-\text{grad}\,p=\rho dV/dt$$

Steady flow of an inviscid fluid. In the study of energy transfer and of the forces exerted between the fluid and bodies interacting with it (*e.g.*, the blades of fans, pumps, compressors, turbines, and ship propellers), a knowledge of the pressure and velocity distribution is often essential. In principle, this knowledge may be obtained by experimentation: The actual body, a prototype or, alternately, a geometrically similar body equipped with suitable sensors and placed in a wind- or water-tunnel, may serve to furnish the desired information. In practice, however, such a procedure may not be feasible for technical, economical, or other reasons. A theoretical determination, instead, is therefore desirable. That this indeed is possible, will be shown here for a system comprising an incompressible and inviscid fluid, not subject to any body forces. For the sake of clarity, it is further assumed that the flow is essentially two-dimensional; *i.e.*, describable in terms of two coordinates, x and y, only. Physically, this means that the flow pattern does not vary with z: the flow in all planes parallel to the xy-plane is identical with that in the latter. Under these conditions, equations (46) and (54), respectively, reduce to the sum of partial derivatives of u and v, one with respect to x, one with respect to y, equalling zero (see 56), and two similar equations involving, as well, p, ρ, and t (see 57). Multiplying the two equations in (57) by dx and dy, respectively, and adding the results, gives an equality involving partial derivatives of p, derivatives of u and v, and a factor ρ (see 58), which equivalently may be written as the negative of the differential of p equal to ρ times a sum of terms involving differentials of u and v (see 59); whence, by integration, the negative of p equals a sum of two terms, one involving ρ and V, one involving K (see 60), with the result that the sum of p and a term composed of one-half ρ times the square of V is a constant (see 61).

Relation (61), which is sometimes called Bernoulli's equation, after the Swiss mathematician Daniel Bernoulli, shows that if the velocity V is known at each point of interest, then it is possible to calculate the corresponding pressure distribution, and vice versa. (To evaluate the constant in Bernoulli's equation, a knowledge of pressure p and velocity V at only one single point is needed.)

Velocity potential and stream function. Physical theories, when formulated in terms of vector quantities, have the advantage of strong physical directness or immediacy. On the other hand, such formulations tend to be unwieldy. Each vector relation referring to an n-dimensional event actually consists of n independent component equations. In addition, most operations with vectors are considerably more involved than operations with scalars. For these and related reasons, much effort has been expanded to devise alternative formulations of physical theories, not in terms of vectors, but in terms of scalar point functions $h(x, y, z)$, instead. With these, only one quantity (instead of n) is needed at each point, and only the simpler rules applicable to calculating with scalars need to be used.

In the present context, two such scalar functions will be considered, the treatment again being restricted to two dimensions (for greater clarity). The first of the functions, called the velocity potential, (x, y), is defined by certain relations (see 62), in which u and v are the velocity components in the x- and y-directions. It is to be noted that the single function, symbolized by the Greek letter phi, ϕ, implicitly contains both components of the velocity V. Furthermore, if there exist two velocity fields characterized by velocity potentials ϕ_1 and ϕ_2, respectively, then the scalar representation of the resultant velocity field is obtained by simply adding these potentials algebraically; *i.e.*, $\phi = \phi_1 + \phi_2$. The resultant velocities, if desired, are again obtained merely by differentiation of ϕ. Intricate forms of flow can thus be built up by simple superposition of two or more elementary velocity potentials.

The velocity potential

Using the relations (62) in the equation of continuity (50) or (56) gives a second-order partial differential equation for the function ϕ (see 63) known as Laplace's equation (first proposed by Pierre-Simon Laplace, an 18th–19th-century French scientist), of fundamental importance in mathematical physics: It must be satisfied for all incompressible flows.

Because $\phi = \phi(x, y)$—*i.e.*, ϕ is a function of position coordinates x and y—the total differential of ϕ is a sum of differential terms involving u, v and the differentials of x and y (see 64), for which use has been made of relations (62). Accordingly, an equipotential line is defined as a path along which $d\phi = 0$. That is, by equation (64), the derivative of y with respect to x restricted by ϕ is the negative ratio of u over v (see 65), which gives the slope of the equipotential line. (The subscript here means that ϕ is to be held constant.)

The other scalar, called the stream function and denoted by $\psi(x, y)$, is defined so that u and v are partial derivatives of ψ (see 66). With the aid of these relations, the total differential of ψ is expressible in terms of differentials for x and y, with v and u being multiplicative factors (see 67). Constancy of the stream function implies $d\psi = 0$, so that the derivative of y with respect to x restricted by ψ is the ratio of v to u (see 68). Recalling that u and v, respectively, denote the x- and y-components of the fluid velocity V, it is seen that equation (68) gives the slope of a line that is everywhere tangent to the fluid velocity. By definition, this line of constant ψ is therefore a streamline.

The stream function

It is further seen that the product of the slopes (65) and (68) is equal to -1. It follows by a theorem of elementary geometry that all streamlines are perpendicular to all equipotential lines. More succinctly, the two sets of lines are said to be mutually orthogonal. This orthogonality constitutes an important mathematical property, essential for the extensive and elegant treatment of fluid mechanics by methods developed in the so-called theory of functions of a complex variable.

Another property of the stream function is that it, too, satisfies Laplace's equation: the sum of second-order partial derivatives of ψ with respect to its two arguments is zero (see 69). This follows from differentiating equations (66) and using equations (62); *i.e.*, two partial differential equations of second order relate ψ and ϕ (see 70). Because the order of differentiating in the extreme right-hand members of equations (70) is immaterial, these members are equal in magnitude, but opposite in sign. Hence, adding these two equations yields the relation (69).

The preceding paragraphs illustrate some of the mathematical elegance (*e.g.*, formal symmetry, compact expression of orthogonality) achieved through the use of scalar functions in formulating the theory of fluid mechanics. In conclusion it may be noted, however, that this advantage is attained at some sacrifice in physical immediacy. Functions such as ϕ and ψ are essentially mathematical constructs and, as such, are more abstract than the corresponding vector fields. The prevalent and continuing use of such functions, nevertheless, signifies that this loss in physical directness is strongly outweighed, in a vast number of cases, by the advantages gained. (R.C.B.)

MECHANICS OF SOLIDS

Mechanics of solids is the science concerned with the behaviour of solid materials under the influence of forces that are in equilibrium. These forces may be mechanical in origin, such as gravity, inertia, or fluid pressure; or they may be electrical or magnetic.

Although solids essentially are systems of particles, the point of view adopted here differs from that in the first section of this article in which such systems are treated. There it is assumed that solids are rigid (*i.e.*, nondeformable) and hence undergo no changes under the action of forces except with regard to equilibrium and position. In mechanics of solids, the complementary point of view is taken: Attention is focussed on the deformation and disruption of solids under the influence of forces, without regard to changes of equilibrium or position.

The mathematical description of the relation between these forces and the concomitant deformations is the subject of the theory of elasticity. Accordingly, this will now be considered.

Elasticity; concept of stress. A solid of arbitrary shape may be considered subject to three forces, P_1, P_2, P_3, that are in equilibrium and therefore leave the solid at rest. If an imaginary cut is now made, destroying the cohesion of the solid, the two parts would move away from one another because the forces acting on each of the two parts are now no longer in equilibrium. Because, before the cut, the two parts did not move apart, it must of necessity be assumed that forces are acting within the solid on the cut face that prevent this movement. The forces must, therefore, be of such a character that their resultant balances the forces applied to each of the two parts. These internal forces, called into play by external forces acting on a solid, will be distributed in some manner over the surface of the imaginary cut, and each unit of area of the internal surface will carry an appropriate proportion of the force. This proportion of internal force carried by unit area is termed the stress.

Definition of stress

If the force that has to be balanced is P, and the area over which the stress can be assumed to be distributed is A, then the average magnitude of the stress at every point of the surface will be equal to the force divided by area, or P/A. In general, the distribution of stress over the area of an imaginary cut will not be uniform, and it is one of the general problems of the theory of elasticity to determine the exact distribution of stress in the interior of a solid subject to external forces.

If a solid, instead of being divided by one cut into two parts, is divided by a large number of cuts into many small cubes of unit volume, each of these cubes by the same reasoning will transmit a stress on each of its six cube faces. These stresses will not be acting at right angles to the cube faces, but each of them can be resolved into three components of which one, the normal stress, is acting at right angles to the cube face, and two components, called shear stresses, are acting in the cube face. Because there are six cube faces, there will be 18 such components. Those designated X_x, X_y, X_z, Y_x, Y_y, Y_z, Z_x, Z_y, Z_z are related to three faces.

Shear stresses

Equilibrium conditions. Considerations of equilibrium lead to certain relations between these 18 stress components which reduce their number to 6, of which 3 are normal stresses and 3 are shear stresses. Moreover, certain first-order partial differential equations containing the derivatives of the stress components with regard to a system of rectangular coordinates (Cartesian) can be deduced from considerations of equilibrium and continuity, and these equilibrium conditions form one set of fundamental equations that must be satisfied by the solutions obtained for any stress problem.

These three equilibrium conditions, in the absence of body forces—that is, forces proportional to mass and acting on each cubical element—are obtained in the form of three partial differential equations for components of X, Y, and Z (see 71), in which X_x, Y_y, Z_z are the three normal components, and $X_y = Y_x$, $X_z = Z_x$, $Y_z = Z_y$ are the three shear stress components.

Tensile and compressive stress; triaxial and plane stress. An example may be considered in the form of a solid composed of a fine wire, at the ends of which two forces of equal magnitude but opposite direction are acting, trying to tear the wire apart. Such forces produce stresses over any cross section of the wire, directed away from the area of the cross section. Such stresses are called tensile stresses.

A short pillar or block carrying a load at the top may also be considered; an equal and opposite reaction will be produced in the base of the pillar resting on the foundation. The pillar is under the influence of two forces trying to shorten it, and in every cross section of the pillar at right angles to the forces, stresses will be produced directed toward the surface of the cross section. Such stresses are termed compressive stresses. The wire under tension and the pillar under compression illustrate two of the simplest types of stress distributions in solids. The stress at every point of the solid will be the same, and will be in the same direction. If a small cubical element is considered with its edges parallel to the applied forces, only one of the six stress components is acting; all the others will be zero. Such a state of stress is called a uniaxial state of stress. The more general case in which stresses are acting on all six cube faces is a general triaxial state of stress.

If stresses are acting on only four of the six cube faces, the number of stress components will be reduced from six to three, and the problem is one of biaxial stress or plane stress.

In the wire under tension or the pillar under compression, the state of stress is the same at every point. Such states of stress, and they need not necessarily be uniaxial, are termed homogeneous states of stress. In general, the state of stress in a solid will be nonhomo-

$$(56)\quad \frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0$$

$$(57)\quad -\frac{\partial p}{\partial x}=\rho\frac{du}{dt}, \qquad -\frac{\partial p}{\partial y}=\rho\frac{dv}{dt}$$

$$(58)\quad -\frac{\partial p}{\partial x}dx-\frac{\partial p}{\partial y}dy=\rho\left(\frac{du}{dt}dx+\frac{dv}{dt}dy\right)$$

$$(59)\quad -dp=\rho\left(\frac{dx}{dt}du+\frac{dy}{dt}dv\right)$$

$$(60)\quad -p=\rho\left(\frac{u^2}{2}+\frac{v^2}{2}\right)=\tfrac{1}{2}\rho V^2+K$$

$$(61)\quad p+\tfrac{1}{2}\rho V^2=\text{constant}$$

$$(62)\quad u=\partial\phi/\partial x, \qquad v=\partial\phi/\partial y$$

$$(63)\quad \frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}=0$$

$$(64)\quad d\phi=\frac{\partial\phi}{\partial x}dx+\frac{\partial\phi}{\partial y}dy=udx+vdy$$

$$(65)\quad (dy/dx)_\phi=-u/v$$

$$(66)\quad u=\partial\psi/\partial y, \qquad v=-\partial\psi/\partial x$$

$$(67)\quad d\psi=\frac{\partial\psi}{\partial x}dx+\frac{\partial\psi}{\partial y}dy=-vdx+udy$$

$$(68)\quad (dy/dx)_\psi=v/u$$

$$(69)\quad \frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}=0$$

$$(70)\quad \begin{cases}\dfrac{\partial^2\psi}{\partial x^2}=-\dfrac{\partial v}{\partial x}=-\dfrac{\partial^2\phi}{\partial x\partial y}\\[2ex] \dfrac{\partial^2\psi}{\partial y^2}=\dfrac{\partial u}{\partial y}=\dfrac{\partial^2\phi}{\partial y\partial x}\end{cases}$$

$$(71)\quad \begin{cases}\dfrac{\partial X_x}{\partial x}+\dfrac{\partial X_y}{\partial y}+\dfrac{\partial X_z}{\partial z}=0\\[2ex] \dfrac{\partial Y_x}{\partial x}+\dfrac{\partial Y_y}{\partial y}+\dfrac{\partial Y_z}{\partial z}=0\\[2ex] \dfrac{\partial Z_x}{\partial x}+\dfrac{\partial Z_y}{\partial y}+\dfrac{\partial Z_z}{\partial z}=0\end{cases}$$

geneous; this means it will be different at every point, or mathematically speaking: the state of stress is a function of the coordinates.

Concept of strain; Poisson's ratio. Under the influence of the forces acting on the wire discussed above, the ends will be displaced from some initial point A, B to points A' and B', respectively.

If the original length of the wire AB was l, and the stretched length $A'B'$ is l', then the extension Δl of the wire by the force P is $\Delta l = l' - l$. If the material of the wire is perfectly homogeneous, each unit length of wire will make an equal contribution to the total extension and the extension per unit length will be $\Delta l/l$. This is called the longitudinal strain.

Longitudinal strain

Exact measurements of the diameter of the wire before and after application of the load would reveal that, as the wire lengthens, the diameter diminishes. The longitudinal dilatation of the wire under the influence of the tensile forces P is accompanied by a lateral contraction. This lateral contraction per unit length of diameter will be only a fraction of the longitudinal strain, so that the volume of the wire will slightly increase. The ratio of lateral contraction to longitudinal extension is a constant for a given material and is termed Poisson's ratio, μ (named after Siméon-Denis Poisson, a 19th-century French mathematician). Its value varies between 0.2 and 0.5 for different materials. If the value of Poisson's ratio approaches 0.5, as it does for cork, there will be no change in volume in consequence of stress.

$$\text{(72)}\qquad \begin{cases} \dfrac{\partial^2 e_{xx}}{\partial y^2}+\dfrac{\partial^2 e_{yy}}{\partial x^2}-\dfrac{\partial^2 e_{xy}}{\partial x\partial y}=0 \\ \dfrac{\partial^2 e_{yy}}{\partial z^2}+\dfrac{\partial^2 e_{zz}}{\partial y^2}-\dfrac{\partial^2 e_{yz}}{\partial y\partial z}=0 \\ \dfrac{\partial^2 e_{zz}}{\partial x^2}+\dfrac{\partial^2 e_{xx}}{\partial z^2}-\dfrac{\partial^2 e_{xz}}{\partial x\partial z}=0 \end{cases}$$

$$\text{(73)}\qquad \begin{cases} \dfrac{\partial}{\partial x}\left(\dfrac{\partial e_{xy}}{\partial z}+\dfrac{\partial e_{xz}}{\partial y}-\dfrac{\partial e_{yz}}{\partial x}\right)-2\dfrac{\partial^2 e_{xx}}{\partial y\partial z}=0 \\ \dfrac{\partial}{\partial y}\left(\dfrac{\partial e_{yz}}{\partial x}+\dfrac{\partial e_{xy}}{\partial z}-\dfrac{\partial e_{xz}}{\partial y}\right)-2\dfrac{\partial^2 e_{yy}}{\partial x\partial z}=0 \\ \dfrac{\partial}{\partial z}\left(\dfrac{\partial e_{xz}}{\partial y}+\dfrac{\partial e_{yz}}{\partial x}-\dfrac{\partial e_{xy}}{\partial z}\right)-2\dfrac{\partial^2 e_{zz}}{\partial x\partial y}=0 \end{cases}$$

$$\text{(74)}\qquad \begin{cases} e_{xx}=\dfrac{\partial u}{\partial x}, & e_{xy}=\dfrac{\partial u}{\partial y}+\dfrac{\partial v}{\partial x} \\ e_{yy}=\dfrac{\partial v}{\partial y}, & e_{yz}=\dfrac{\partial v}{\partial z}+\dfrac{\partial w}{\partial y} \\ e_{zz}=\dfrac{\partial w}{\partial z}, & e_{xz}=\dfrac{\partial u}{\partial z}+\dfrac{\partial w}{\partial x} \end{cases}$$

Shear strain; general deformation; compatibility. The deformation produced by a general state of strain cannot be described in terms of dilatations and contractions alone. A square $ABCD$, under the influence of shear stresses, will deform into a rhombus $ABC'D'$ much in the manner of cards sliding over each other in a pack. The deformation is described by the angle DBD', and this is called a shear strain. If, as before, a solid of arbitrary shape is divided into a large number of small cubes, each of the cubes can in the most general case undergo dilatations or contractions in the directions of the three cube edges; in addition, each of the squares bounding the cube can undergo a shear strain so that the cube is deformed into a parallelepiped. The three dilatations in the directions of the coordinate axes and the three shear strains in the coordinate planes describe the deformation of the elementary cube completely; and, if these six strain components are known as functions of the coordinates, they are known at every point of the solid and the deformation of the solid is then fully described.

There is one important condition that must be satisfied by the strain components. If they were entirely independent of each other, then each elementary cube might be deformed in such a manner that, after deformation, the cubes would no longer fit together without gaps, and the continuity of the solid would be destroyed. In order to ensure the continuity of the solid after deformation, the functions representing the strain components must satisfy a set of six second-order partial differential equations, known as the compatibility conditions.

Compatibility conditions

These six compatibility conditions are obtained in the form of second-order partial differential equations for the strain functions (see 72 and 73), in which e_{xx}, e_{yy}, e_{zz} are the three normal strains and $e_{xy} = e_{yx}$, $e_{xz} = e_{zx}$, $e_{yz} = e_{zy}$ are the three shear strains. The compatibility conditions can be derived from the fact that the components of strain are derivatives of the components of the displacement vector. The displacement vector is a measure of the displacement of each cubical element in space with regard to a fixed system of coordinate axes. If the components of the displacement vector in the three directions x, y, z of the coordinate system are designated by u, v, w, respectively, the components of strain are derived as six expressions involving partial derivatives of u, v, and w (see 74). The compatibility conditions can be derived from these six equations by eliminating the three components of the displacement vector. These six equations are therefore equivalent to the compatibility equations and are frequently used in their place.

Relation between stress and strain. It is a fact of experience that a wire subjected to equal and opposite forces P at its ends extends. The extension is considered a consequence of the stresses produced in the wire by the forces. There exists, therefore, a relation of cause and effect between stress and strain, or strain can be considered a function of stress. This functional relationship must be expressible by some mathematical law, and two distinct possibilities for postulating such a law present themselves.

The one possibility would be to determine the relationship experimentally and then find a mathematical expression representing this relationship. This meets with formidable difficulties because such measurements are possible only for relatively simple states of stress and deformation and cannot be carried out for general states of deformation resulting from triaxial states of stress. Measurements on some metal wires and metal pillars, however, show that for certain simple types of stress, when the stresses do not exceed certain limits, the strain is proportional to stress so that a linear relationship holds.

For other metals and such materials as concrete, stone or timber, the relation even for simple states of stress is not linear. Such empirical stress–strain relations could be applied with confidence only within the region for which they have been experimentally confirmed.

The second possibility, and this is the one adopted in the classical theory of elasticity, is to postulate a relationship that will result in a mathematical theory aesthetically satisfying and at the same time simple. Only the assumption of a linear law permits this.

The linear relation between stress and strain is known as Hooke's law, so named after its discoverer, Robert Hooke of England (about 1676). It is applicable in many solid materials, provided the strains are sufficiently small. This law of proportionality has to be generalized for triaxial states of stress, and six equations can be written expressing the six strain components as linear functions of stress. This would involve 36 constants of proportionality.

Hooke's law

The proportionality between stress and strain does not involve elasticity, this being the property of the solid to recover its original shape completely after the forces producing the deformation are removed. Elasticity, on

the other hand, does not necessarily require proportionality between stress and strain. Modern spring steels recover their shape completely after deformation, but stress and strain do not obey even approximately a linear law. Despite these fundamental conceptual differences between stress–strain proportionality and elasticity, the two concepts are used indiscriminately in such a way that elasticity implies proportionality and vice versa. The theory of elasticity is not primarily concerned with elasticity, that is, with the behaviour of solids after removal of the load, but is essentially employed as a linear theory with proportionality when the load is applied.

The assumption of elasticity, however, permits the application of certain thermodynamic principles that enable the reduction of the 36 independent elastic constants to 21. This number can be further reduced to 15, and if it is assumed that the material is isotropic (see below) with regard to its elastic properties, the number of elastic constants can be reduced to 2. These are the modulus of elasticity, generally denoted by the symbol E, and Poisson's ratio, μ, which has already been mentioned. The modulus of elasticity E is also called Young's modulus after Thomas Young, an 18th–19th-century English scientist, who first defined it. It is, for a given material, the ratio between stress and strain for a simple uniaxial state of stress. Isotropy with regard to elastic properties means that the properties of the material are independent of direction. This is far from being true for most engineering materials. It is obviously not so for wood, which has different properties along and across the grain. Metals produced by casting, rolling, forging, or extrusion, however, will also exhibit some directionality with regard to elastic properties.

Young's modulus

The six equations of proportionality (or Hooke's law) for a general triaxial state of stress are obtained in the following forms: (a) Strains are expressed in terms of stresses in six relations involving E and G (see 75 and 76). In these equations, G is introduced as a further elastic constant, the modulus of rigidity. This, however, is not a third independent constant, but is a function of Young's modulus E and Poisson's ratio μ (see 77). (b) Stresses are expressed in terms of strains in six equations in which the constants G, μ, and e appear (see 78). In these equations e, written as a sum of three terms (see 79), is the change in volume of the unit cube of the material under the given state of stress, the cubical dilatation.

For a state of hydrostatic stress for which the three normal stress components $X_x = Y_y = X_z$ are equal to a uniform pressure p, and the three shear stress components $X_y = X_z = Y_z = 0$, the cubical dilatation from summation of equations (75) is obtained as two expressions relating e, μ, p, k, and E (see 80) and k is called the modulus of compression or bulk modulus of the material. The value for Poisson's ratio μ cannot exceed 0.5, because for values of μ greater than 0.5 the bulk modulus becomes negative, which would require the solid to increase its volume under hydrostatic pressure.

The bulk modulus

Principal stresses and strains; invariants and strain energy. The six stress components describing the state of stress and the six strain components describing the state of strain at every point are, of course, functions of the coordinate axes and will change if a different system of coordinate axes is used. They form what are called tensors in mathematics, and from this fact two important deductions can be made with regard to the effect of transformations of the coordinate axes.

The first deduction is that there are always three orthogonal directions for which, when they are chosen as the coordinate axes, the expressions for the shear stresses and shear strains become zero and the values of the normal stresses are extrema. These directions, by definition, are those of principal stress (or strain). The largest of the three principal stresses (or strains) will be the absolute maximum value of the stress (or strain) at the given point, and the smallest of the three will be the absolute minimum value. If the state of stress or strain is written in the form of a matrix (see 81), then for the three principal directions $x \to 1$, $y \to 2$, $z \to 3$, the state of stress will be fully characterized by three principal stress components, diagonal matrix components X_1, Y_2, Z_3 (see 82), and the state of strain by three principal strain components, diagonal matrix elements expressed in terms of e_{ij} (see 83). In the case of simple tension or compression already treated, the direction of the tensile or compressive stress will be a direction of principal stress, because the shear stresses in planes at right angles and parallel to this direction must be zero.

The equations of equilibrium (71), the compatibility conditions (72 and 73), and the equations of proportionality (75, 76, and 78) can be written in terms of principal stresses without losing generality.

In tensor analysis, it is shown that certain expressions can be formed from the components of tensors of the type used for the description of states of stress or strain that are invariants with regard to transformations of the coordinate axes. This is the second important deduction, because a physical quantity in which the formal structure and/or numerical value do not depend upon the choice of a particular coordinate system is evidently—*i.e.*, by definition—of more general validity than an expression the form and value of which do change from one coordinate system to another. Invariance of a physical quantity under coordinate transformation is generally taken to reveal some profound aspect, or intrinsic property, of the system considered.

Importance of invariance

In the present context, one set of invariants involving stresses and strains, respectively, is given by two linear equations (see 84). The first of these two expres-

(75) $$e_{xx}=\frac{1}{E}[X_x-\mu(Y_y+Z_z)], \qquad e_{xy}=\frac{1}{G}X_y$$

(76) $$\begin{cases} e_{yy}=\frac{1}{E}[Y_y-\mu(X_x+Z_z)], & e_{xz}=\frac{1}{G}X_z \\ e_{zz}=\frac{1}{E}[Z_z-\mu(X_x+Y_y)], & e_{yz}=\frac{1}{G}Y_z \end{cases}$$

(77) $$G=\frac{E}{2(1+\mu)}$$

(78) $$\begin{cases} X_x=2G\left(e_{xx}+\mu\frac{e}{1-2\mu}\right), & X_y=Ge_{xy} \\ Y_y=2G\left(e_{yy}+\mu\frac{e}{1-2\mu}\right), & X_z=Ge_{xz} \\ Z_z=2G\left(e_{zz}+\mu\frac{e}{1-2\mu}\right), & Y_z=Ge_{yz} \end{cases}$$

(79) $$e=e_{xx}+e_{yy}+e_{zz}$$

(80) $$e=3p\frac{1-2\mu}{E}=kp, \qquad k=3\frac{1-2\mu}{E}$$

(81) $$\begin{pmatrix} X_x & X_y & X_z \\ Y_x & Y_y & Y_z \\ Z_x & Z_y & Z_z \end{pmatrix}, \qquad \begin{pmatrix} e_{xx} & e_{xy} & e_{xz} \\ e_{yx} & e_{yy} & e_{yz} \\ e_{zx} & e_{zy} & e_{zz} \end{pmatrix}$$

(82) $$\begin{pmatrix} X_1 & 0 & 0 \\ 0 & Y_2 & 0 \\ 0 & 0 & Z_3 \end{pmatrix}$$

(83) $$\begin{pmatrix} e_{11} & 0 & 0 \\ 0 & e_{22} & 0 \\ 0 & 0 & e_{33} \end{pmatrix}$$

(84) $$\begin{cases} I_1^x=X_1+Y_2+Z_3 \\ I_1^e=e_{xx}+e_{yy}+e_{zz} \end{cases}$$

sions indicates that the sum of the three principal stresses is a constant. The second expression is the cubical dilatation; *i.e.*, a volume change. This change is independent of the coordinate system employed.

A second set of invariants involving the components of the stress and strain tensors (81), respectively, is two quadratic equations (see 85). Combining now the two relations (85) yields another invariant, namely U, equal to an expression involving E and μ (see 86), which can be shown to represent the strain energy per unit volume of the strained solid. The source of this energy is the work done by the external forces against the resistance of the solid to being deformed. This work, stored as elastic energy within the solid, is released when the external forces producing the deformation are removed. The unwinding of a clock spring is a good example of stored strain energy.

By combining the invariants—*i.e.*, the first equations, (84) and (85)—a similar expression for the strain energy density U can be obtained as three equations that identify U, U_1, and U_2 in terms of components of X, Y, and Z (see 87). Here, the term U_1 represents that part of the total strain energy U that is absorbed in changing the volume of the solid. The other term, U_2, called the shear strain energy, or the energy of distortion, is that part of U that is absorbed in a change in form without a change in volume.

Energy of distortion

The strain energy U plays an important role in the solution of many problems in the theory of elastic solids.

(J.F.Br./R.Wk.)

STATISTICAL MECHANICS

Basic definitions. Statistical mechanics is the branch of theoretical physics that seeks to draw a connection between the macroscopic viewpoint of matter and its microscopic constitution; the method is mathematical and quantitative. The behaviour of bulk macroscopic matter is described in such terms as volume, density, pressure, temperature, etc., none of which is meaningful when applied to single atoms or molecules. Yet, bulk laboratory matter does consist entirely of atoms and molecules. The momentary states of these are described classically by giving coordinate values, and velocities or momenta. The motions are governed by the laws of mechanics. The task of statistical mechanics is to infer the behaviour of bulk matter, the macroscopic system, from a knowledge of the structure of the molecules composing it, using the laws of mechanics.

One of the simplest of thermodynamic systems, a fixed mass of some particular monatomic gas such as neon, in a given volume V, having a known energy E, will serve as an example. Instead of mass, m, use of the thermodynamic variable N denoting the number of atoms is preferred. The three specifications of V, E, N suffice to define the thermodynamic state of this simple system at equilibrium; it will have a definite temperature T, pressure P, and chemical potential per atom, given by the Greek letter mu, μ. The number of atoms N is extremely large. Even for as little as a milligram of gas, it will be of order 10^{20}.

The mechanical state, or microstate, of such a gas at a given instant of time is described classically for each of the N atoms (specified by i, in which i has all values between 1 and N) by giving $3N$ Cartesian coordinates, x_i, y_i, z_i, and $3N$ vector components of velocity v_{xi}, v_{yi}, v_{zi}. Actually, the laws of mechanics in Hamiltonian form are used in statistical mechanics; *i.e.*, the coordinates and momenta, p_{xi}, p_{yi}, and p_{zi}, p_{xi}, $= m_i v_{xi}$, etc. (*e.g.*, momentum p_{xi} equals mass m_i times velocity v_{xi}), must be used instead of coordinates and velocities. This choice greatly simplifies the mathematical treatment of more complicated systems. The combination of $3N$ coordinates and $3N$ momenta is called the phase space of the system, and following the terminology introduced by a 20th-century Austrian-Dutch physicist, Paul Ehrenfest, is known as gamma-space (γ-space), for gas phase space. In contrast, the mu-space (μ-space) for molecule space is the six-dimensional phase space for a single molecule described by six coordinates x, p_x, y, p_y, z, and p_z. For molecules consisting of two or more atoms, internal coordinates and their conjugate momenta (so-called generalized momenta related to potential and kinetic energy) are used in addition to the three Cartesian coordinates of the centre of mass. The total necessary number, γ, of coordinates is called the number of degrees of freedom and is three times the total number of atoms. The microstate of the gas is represented at any time t by a single point in γ-space, or by N points, one for each molecule, in μ-space. The number of coordinates in μ-space is equal to three times the number of atoms per molecule, and as many momenta.

Phase space

When only the macrostate V, E, N is specified, then the allowed γ-space positions include states of abnormal properties. For instance, it includes many positions in which all, or a great preponderance, of the molecules are in one half of the volume, leaving the other half almost empty; or in which all of the molecules have nonzero momentum along the x-axis only, thus exerting zero pressure on the walls parallel to this axis.

The procedure of statistical mechanics is to calculate the probability or, more precisely, the probability density (see below) that the system will be found at a given point in γ-space. In the example of fixed V, E, N this can be shown to be constant in all parts of γ-space consistent with the restraints on total volume, total energy, and number of molecules. For each point in γ-space, the mechanical properties of the system are exactly determined; for instance, the pressure on each element of wall area, caused by collisions of the molecules with the wall. From this and the probability density in γ-space, the average of any mechanical property, such as the pressure, is computable. Furthermore, the probability that the pressure will differ from the predicted average by a specified amount can also be calculated. In agreement with empirical observation, one finds that for fractional deviations greater than the errors of measurement, the probabilities are negligibly small so long as N is large. A rough rule is that, ordinarily, fractional fluctuations appreciably greater than about $10/\sqrt{N}$ have quite negligible probabilities.

Probability. The concepts of probability enter fundamentally into all of statistical mechanics. For a given population such as the independent molecules of a perfect gas at temperature T, the quantity W_ν may be defined as the probability that a randomly selected molecule will be in the internal quantum state designated by ν. The sum of probabilities W_ν over all quantum states must be one (unity), because each molecule must be in some particular state; *i.e.*, the sum of W_ν is one (see 88), in which the mathematical notation Σ is used to indicate summation over all values of ν. If f is any quantity and any particular value f_ν is known for each state ν, then the average, written as $\langle f \rangle$, for the population is a weighted sum of f_ν (see 89) and the average of the square of f is a weighted sum of the squares of f_ν (see 90).

If, instead of an integer variable (*e.g.*, the number of molecules), the variable under consideration is continuous—*i.e.*, infinitely divisible—as in the case of the energy E, then the concept of probability density $W(E)$ is used. This quantity is defined in such a way that $W(E)\Delta E$ is the probability that the energy lies between E and $E + \Delta E$, in the limit as ΔE approaches zero. The quantity $W(E)$ is thus seen to be a probability per unit energy interval, ΔE; it is for this reason that $W(E)$ is termed a probability density (in close analogy with the more familiar mass density, or simply density, meaning mass per unit volume).

Probability density

A relation analogous to equation (88) can now be obtained by introducing an auxiliary quantity $W^*(E)$, defined as the probability that the energy is less than E or, formally, an equation relating W and W^* (see 91). Because E here is a continuous variable, the summation in equation (88) must be replaced by integration, so that the integral of W as a function of E is one (see 92), in which the integration is extended over the allowed range of the variable. The average, $\langle E \rangle$,

of the variable is an integral with integrand the product of E and W (see 93); the average of its square is an integral with the integrand containing as factors the square of E and the function W (see 94).

Perfect gas pressure and energy density. If E_k is the total kinetic energy of the molecular centre of mass in a perfect gas of volume V, then the pressure, P, is expressed (see 95) in terms of the number of molecules, N, and $\langle \varepsilon_k \rangle$, their average kinetic energy, which is the average of one-half the product of m times the square of v (see 96).

This relation may be derived as follows. Given a rectangular container with rigid walls, it is provisionally assumed that all molecules move with the same speed v, one-sixth of them along the positive and one-sixth along the negative directions of each of the three Cartesian axes, x, y, z. Without loss of generality, the system of coordinate axes may be chosen so that the positive x-axis is perpendicular to the right-hand wall of the container. The number of molecules striking unit area of this wall in a time t is given by $(1/6)(N/V)vt$. This is so because all molecules moving along the positive x-direction will in time t reach the unit area under consideration provided they are inside a cylindrical volume equal to $vt \times 1$; *i.e.*, a cylinder of length vt along the x-direction and of unit cross sectional area. The total number of such molecules within the cylinder is equal to the product of two factors: the number of these molecules per unit volume, and the volume in question; *i.e.*, $(N/6V) \times vt$.

Upon elastically colliding with the wall, each molecule undergoes a change in momentum equal to $2mv$. By Newton's second law of mechanics, therefore, the force exerted by the gas per unit area of the wall, or the pressure P, is expressed in terms of m, v, N, and V (see 97), or, in terms of the molecule's kinetic energy (see 95), P equals two-thirds the product of ε_k times the ratio of N to V (see 98).

The restrictions of the same speeds for all molecules and motions only along the Cartesian axes will now be removed. The pressure is the sum of the momentum changes of all molecules impinging on unit area in unit time. Denoting by $N(v_x)dv_x$ the number of molecules in V with x-component of velocity between v_x and $v_x + dv_x$, the number $N(v_x)/V$ (in a parallelepiped of unit base and height $v_x t$, of volume $v_x t \times 1$) will, if v_x is positive, strike a unit wall area perpendicular to the x-axis in an interval of time of duration t. An equal number will leave the wall with negative v_x in that time. Their summed momentum change is $mv^2tN(v_x)/V$, which, divided by t and summed by integration over v_x from minus infinity to plus infinity, is the pressure (see 99). Because the average squares of velocity components are equal (see 100) and their sum is $\langle v^2 \rangle$, $\langle v_x^2 \rangle = \frac{1}{3} \langle v^2 \rangle$ and the proof of (97) and (98) is complete.

Definition of a perfect, or ideal, gas

A perfect, or ideal, gas is defined as one that obeys an equation in which the product of V and P is a product of N, k, and T (see 103), in which T is the ideal gas temperature and k is Boltzmann's constant, equal to $1.380\,54 \times 10^{-16}$ erg per degree. If N_0 is Avogadro's number, equal to $6.022\,52 \times 10^{23}$, then R, the gas constant per mole, is expressed as an equality that relates R, N_0, and k (see 104), with numerical value 8.314 3 joule per mole. From (96) it follows that the average of ε_k is 3/2 of the product of k and T (see 105).

Maxwell-Boltzmann distribution. The probability density, W, of molecules in three-dimensional velocity space (denoted v, with its three components v_x, v_y, and v_z, which are parallel, respectively, to the three direction coordinates x, y, and z) is known as the Maxwell-Boltzmann distribution law, which will be derived below. It is written in exponential form (106) in terms of the velocity squared (v^2), which is the square of the absolute value of v, equal to the sum of the squares of the component velocities (see 107) and in terms of the mass (m), absolute temperature (T), and Boltzmann's constant (k). (The notation exp [a] for exponential a is the same as e^a, or the Naperian number $e = 2.718 \cdots$ raised to the power a.) In momentum space, in which momentum p equals mv, $W(p)$ is expressible as a negative exponential in which the square of p appears in the argument (see 108), with p^2 equal to $|p|^2$. In

(85) $$\begin{cases} I_2^x = X_x Y_y + X_x Z_z + Y_y Z_z - X_x^2 - Y_z^2 - X_z^2 \\ I_2^e = e_{xx}e_{yy} + e_{xx}e_{zz} + e_{yy}e_{zz} - e_{xy}^2 - e_{xz}^2 - e_{yz}^2 \end{cases}$$

(86) $$U = \frac{E(1-\mu)}{2(1+\mu)(1-2\mu)}(I_1^e)^2 - \frac{E}{1+\mu} I_2^e$$

(87) $$\begin{cases} U = U_1 + U_2 \\ U_1 = \dfrac{1-2\mu}{6E}(X_x + Y_y + Z_z)^2 \\ U_2 = \dfrac{1+\mu}{6E}[(X_x - Y_y)^2 + (Y_y - Z_z)^2 + (Z_z - X_x)^2] + \\ \qquad + \dfrac{1}{2G}(X_y^2 + X_z^2 + Y_z^2) \end{cases}$$

(88) $$\sum_\nu W_\nu = 1$$

(89) $$\langle f \rangle = \sum_\nu f_\nu W_\nu$$

(90) $$\langle f^2 \rangle = \sum_\nu f_\nu^2 W_\nu$$

(91) $$W(E) = \frac{dW^*(E)}{dE}$$

(92) $$\int W(E)\, dE = 1$$

(93) $$\langle E \rangle = \int dE\, E\, W(E)$$

(94) $$\langle E^2 \rangle = \int dE\, E^2\, W(E)$$

(95) $$P = \tfrac{2}{3} E_k / V = \tfrac{2}{3} \langle \varepsilon_k \rangle (N/V)$$

(96) $$\varepsilon_k = \langle \tfrac{1}{2} mv^2 \rangle$$

(97) $$P = [\tfrac{1}{6}(N/V)\, vt](2mv/t) = \tfrac{1}{3} mv^2 (N/V)$$

(98) $$P = \tfrac{2}{3}(\tfrac{1}{2} mv^2)(N/V) = \tfrac{2}{3} \varepsilon_k (N/V)$$

(99) $$P = \int_{-\infty}^{+\infty} dv_x\, mv_x^2 [N(v_x)/V] = m(N/V)\langle v_x^2 \rangle$$

(100) $$\langle v_x^2 \rangle = \langle v_y^2 \rangle = \langle v_z^2 \rangle,\ \langle v^2 \rangle = \langle v_x^2 + v_y^2 + v_z^2 \rangle$$

(101) $$\langle v_x^2 \rangle = \tfrac{1}{3} \langle v^2 \rangle$$

(102) $$\begin{cases} P = \sum_\nu \tfrac{2}{3} \varepsilon_{k,\nu} (N_\nu / V) \\ P = \tfrac{2}{3} \langle \varepsilon_k \rangle (N/V) = \tfrac{2}{3} E_k / V \end{cases}$$

(103) $$PV = NkT$$

(104) $$R = N_0 k, \qquad k = R/N_0$$

(105) $$\langle \varepsilon_k \rangle = \tfrac{3}{2} kT$$

(106) $$W(v_x, v_y, v_z) = W(v) = \left(\frac{m}{2\pi kT}\right)^{3/2} \exp\left(-\frac{m}{2kT}\right) v^2$$

(107) $$v^2 = |v|^2 = v_x^2 + v_y^2 + v_z^2$$

(108) $$W(\mathbf{p}) = (2\pi mkT)^{-3/2} \exp\left(-\frac{1}{2mkT}\right) p^2$$

$$W(p) = 4\pi p^2 (2\pi mkT)^{-3/2} \exp\left(-\frac{1}{2mkT}\right) p^2 \tag{109}$$

$$W(\varepsilon) = 2\pi \varepsilon_k^{1/2} (\pi kT)^{-3/2} \exp(-\varepsilon_k/kT) \tag{110}$$

$$\langle \varepsilon_k^2 \rangle = (15/4)(kT)^2 \tag{111}$$

$$\begin{cases} N(\rightarrow) = Ak\, W(\boldsymbol{v}_1) W(\boldsymbol{v}_2) \\ N(\leftarrow) = Ak\, W(\boldsymbol{v}_1') W(\boldsymbol{v}_2') \end{cases} \tag{112}$$

$$|\boldsymbol{v}_1|^2 + |\boldsymbol{v}_2|^2 = |\boldsymbol{v}_1'|^2 + |\boldsymbol{v}_2'|^2 \tag{113}$$

$$W(\boldsymbol{v}_1) W(\boldsymbol{v}_2) = W(\boldsymbol{v}_1') W(\boldsymbol{v}_2') \tag{114}$$

$$W(\boldsymbol{v}) = c \exp[-\gamma |\boldsymbol{v}|^2] \tag{115}$$

$$\int_{-\infty}^{+\infty} \iiint dv_x\, dv_y\, dv_z\, W(\boldsymbol{v}) = 1 \tag{116}$$

$$W(\boldsymbol{v}_1) W(\boldsymbol{v}_2) = c^2 \exp[-\gamma(|\boldsymbol{v}_1|^2 + |\boldsymbol{v}_2|^2)] \tag{117}$$

$$H(q^{(\nu)}, p^{(\nu)}) = \sum_{i=1}^{i=N} \frac{1}{2m_i} (p_{xi}^2 + p_{yi}^2 + p_{zi}^2) + U(x_1, \cdots, z_N) \tag{118}$$

$$r = ([x_a - x_b]^2 + [y_a - y_b]^2 + [z_a - z_b]^2)^{1/2} \tag{119}$$

$$\begin{cases} \cos\theta = (z_b - z_a)/r \\ \sin\phi = (y_b - y_a)/[(x_a - x_b)^2 + (y_a - y_b)^2] \end{cases} \tag{120}$$

$$\dot{q}_\nu = (\partial H/\partial p_\nu), \qquad \dot{p}_\nu = -(\partial H/\partial q_\nu) \tag{121}$$

$$\Delta\omega = \prod_{\nu=1}^{\nu=\gamma} \Delta q_\nu \Delta p_\nu \tag{122}$$

terms of the magnitude of momentum equal to $|p|$, W as a function of p is a negative exponential in which the square of p appears in the argument and also in the factor that multiplies the exponential (see 109); and, as a function of the kinetic energy, ε_k, the probability density is an exponential times a term that includes ε_k (see 110). Using equation (93) for the average of ε_k gives equation (105) as it must; and using (99) the average of the square of ε_k is proportional to the square of k times T (see 111).

One derivation of equation (106) is obtained by invoking the principle of microscopic reversibility, which asserts that, at equilibrium, any process and its reverse proceed at equal rates. An example is the collision of two molecules of equal mass and respective velocities $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, resulting in velocities $\boldsymbol{v}_1'$ and $\boldsymbol{v}_2'$ after collision. To an observer moving with the velocity of their centre of mass, the two molecules have equal speeds, namely, $\frac{1}{2}|\boldsymbol{v}_2 - \boldsymbol{v}_1|$, and they approach each other on a line from opposite directions, collide, and leave each other on a line at some angle with the line of approach, again with equal speeds and opposite directions. The reverse process is the collision of molecules with velocities $\boldsymbol{v}_1'$ and $\boldsymbol{v}_2'$ to produce velocities $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Viewed from the centre of mass this process differs from the previous one only in the sign of the angle between the line of approach and the line of departure. According to the principle cited (and using the convention of arrows to denote direction), the intrinsic rate constants $k(\rightarrow)$ and $k(\leftarrow)$ for the two processes must therefore be equal; *i.e.*, $k(\rightarrow) = k(\leftarrow) = k$, in which k is a constant. The total numbers $N(\rightarrow)$ and $N(\leftarrow)$ of the two events will be individually proportional to k and to the product of the two probability densities, $W(\boldsymbol{v})$, evaluated at the initial colliding velocity values. With A representing a constant, it is thus found that the total numbers $N(\rightarrow)$ and $N(\leftarrow)$, which must be equal at equilibrium, are proportional to the product of two probability densities (see 112). Energy conservation requires that the sum of absolute squares of the $\boldsymbol{v}_i$ equal the sum of the absolute squares of $\boldsymbol{v}'_i$ (see 113), whereas equating $N(\rightarrow)$ to $N(\leftarrow)$ in (112) requires that the product of W evaluated at $\boldsymbol{v}_i$ be equal to the product of W evaluated at the arguments $\boldsymbol{v}'_i$ (see 114) as long as (113) applies. A solution of (114) is W as a function of $\boldsymbol{v}$, an exponential in which the exponent is proportional to $\boldsymbol{v}^2$ (see 115), and in which c—a constant—and γ—a function of mass (m), absolute temperature (T), and the Boltzmann constant (k)—are to be determined. A triple integral over all space in which the integrand is the function W must have the numerical value one (see 116). It follows that the product of W evaluated at two points is an exponential in which the exponent includes a sum of absolute squares (see 117) and, by (113), equals also $W(\boldsymbol{v}_1')W(\boldsymbol{v}_2')$. It can be shown that this is the only solution of (114) subject to (113).

The condition corresponding to equation (92)—that is, the total probability in $\boldsymbol{v}$-space is equal to unity (see 116)—requires that the constant c equal $(\gamma/\pi)^{3/2}$. The average kinetic energy $\langle \varepsilon_k \rangle$ is obtained by integrating the quantity $\frac{1}{2}m|\boldsymbol{v}|^2 W(\boldsymbol{v})$, and it is found to be equal to $\frac{3}{4}(m/\gamma)$, from which, with equation (105), γ becomes $(m/2kT)$. This completes the derivation of equation (106).

Gibbs's ensembles. The European investigators of the 19th century most frequently interpreted the bulk material properties of a thermodynamic system as being given by the time-averaged value of a single isolated system. J. Willard Gibbs, a physicist of the United States, however, introduced the concept of ensemble average. A Gibbsian ensemble consists of a large number, K, of systems, with the same thermodynamic specifications, and average values of physical quantities are calculated in the limit that K approaches infinity. This average value is then interpreted as the expectation value of a physical measurement on a real equilibrium system in the given thermodynamic state.

Ensemble average

One important distinction exists between describing the states of systems composing an ensemble and the description in classical thermodynamic theory. The thermodynamic state of an equilibrium system composed of one chemical component is defined (*i.e.*, specified) if its volume (V), energy (E), and number (N) of molecules are given. The system is then assumed to have an exact pressure (P), temperature (T), and chemical potential (μ). Any of six other choices of variable sets, however, can be used to define the same thermodynamic state in classical theory. These sets are: P, E, N; V, T, N; V, E, μ; or P, T, N; P, E, μ; and V, T, μ. One of the so-called extensive variables V, E, or N must be retained to fix the size of the system. Any of the other two can be substituted by its conjugate (corresponding) intensive variable, P, T, or μ. (An intensive variable is one the value of which is not changed when only part, instead of an entire equilibrium system, is being considered. Examples are pressure or temperature. Variables the values of which do change, such as volume or energy, are extensive.)

Gibbsian ensembles, on the other hand, are not identical for such different choices of variables. The ensemble consisting of isolated systems of fixed V, E, and N values is called the microcanonical ensemble, according to Gibbs's nomenclature. That of fixed V, T, and N values is called the petite canonical or simply the canonical ensemble and consists of systems of fixed V and N, in thermal equilibrium with an infinite heat reservoir at fixed temperature (T). The energies of the systems can fluctuate. Their average energy is interpreted as the thermodynamic value. The grand canonical ensemble, for which the thermodynamic state is described by V, T, and μ, comprises systems of fixed volume (V) in equilibrium through walls, permeable to both heat and molecules, with an infinite reservoir at

temperature (T) and with molecules of the same species at chemical potential (μ). This ensemble is sometimes referred to as one of open systems. Both energy and numbers of molecules can fluctuate; their average values are identified with the thermodynamic values.

Ensembles corresponding to other choices of variables are sometimes used, but for most problems one of the three Gibbsian ensembles is generally suited best. For systems containing several chemical species, a, b, c, $\cdots$, the minimum number of variables in the thermodynamic specification is equal to the number of components plus two. The number, N_a, N_b, $\cdots$, of molecules of the different species may be used or one or all replaced by the chemical potentials, μ_a, μ_b, $\cdots$.

Other extensive variables such as the electric or magnetic moments of the systems or the extensive strain tensor may enter the thermodynamic state specifications. Their conjugate intensive state variables are electric and magnetic fields and the stress tensor, respectively. These two sets of variables correspond to generalized displacements and generalized external forces, respectively, entirely analogous to the relation between volume and pressure. If attention is to be focussed on the electric, magnetic, or elastic properties of the system, introduction of the corresponding variables poses no fundamental difficulties. In the absence of such forces, the potential existence of these variables can be ignored.

Returning to the simple one-component system without external forces other than uniform pressure, it is found that for each choice of the triad of state variables, there is a thermodynamic function that is a minimum at equilibrium. For fixed V, E, and N this function is what may be defined as negative entropy ($-S$), a function of V, E, and N, namely, $-S(V, E, N)$. For fixed V, T, and N, it is the Helmholtz free energy $A(V, T, N)$ (see THERMODYNAMICS, PRINCIPLES OF), and for fixed V, T, and μ it is the negative pressure-volume product $-VP(T, \mu)$. The symbols $q^{(\gamma)}$, $p^{(\gamma)}$ are used to denote the entire set, $q^{(\gamma)} \equiv q_1, q_2, \cdots q_\gamma$ of γ coordinates, and conjungate momenta $p^{(\gamma)} \equiv p_1, p_2, \cdots p_\gamma$, necessary to describe the microstate. For each ensemble, Gibbs writes the probability density $W(q^{(\gamma)}, p^{(\gamma)})$ over the complete phase space (γ-space) for finding a random member-system of the ensemble at $p^{(\gamma)}$, $q^{(\gamma)}$. Each of these probability densities contains a constant factor. For the microcanonical ensemble of fixed V, E, N, the factor is $\exp[-S/k]$. For the canonical ensemble of fixed V, T, N, the factor is $\exp[A/kT]$, and for the grand canonical ensemble it is $\exp[-PV/kT]$. The condition that the integral of the probability density over the allowed range of the $p^{(\gamma)}$, $q^{(\gamma)}$ must be unity fixes the value of the functions $-S(V, E, N)/k$, $A(V, T, N)/kT$, $VP(T, \mu)/kT$ in the three ensembles, respectively. The thermodynamic characteristics of any system are completely determined if any one of the functions, $S(V, E, N)$, $A(V, T, N)$ or $VP(T, \mu)$, is known over the complete range of its variables. The choice of ensemble is generally based on considerations of convenience and of mathematical tractability.

Helmholtz free energy

The Hamiltonian. For the canonical and grand canonical ensembles, the Hamiltonian $H(p^{(\gamma)}, q^{(\gamma)})$ enters the expression for the probability density. The Hamiltonian for a system composed of atoms and molecules with purely conservative forces (*i.e.*, forces that can be written as the gradient of potential energy) is simply the expression for the energy, expressed as a function of the coordinate set $q^{(\gamma)}$ and the conjugate momenta $p^{(\gamma)}$.

The atoms composing the molecules of the system are treated as point masses. If Cartesian coordinates x_i, y_i, z_i, and momenta p_{xi}, p_{yi}, p_{zi}, $p_{xi} = mv_{xi}$, $\cdots$ are used for the ith particle of mass m_i, the Hamiltonian is the sum of kinetic energies expressed in terms of momenta and the potential energy functions (U), which depend on the coordinates $x_1, \cdots x_i, y_i, z_i, \cdots, z_N$ only (see 118). Actually, for molecules consisting of more than one atom, a transformation is made to other coordinates and momenta in order to simplify the expression for the potential, U. For instance, for a diatomic molecule comprised of atoms of types a and b with masses m_a and m_b, and total mass $m = m_a + m_b$, the coordinates x, y, and z of the centre of mass are $x = (1/m)(m_a x_a + m_b x_b)$, etc.; the distance r between the atoms is the square root of a sum of squares of differences (see 119) and two Eulerian angles, θ, ϕ, necessary to define the direction of the vector pointing from atom a to atom b, are given by the cosine and sine functions (see 120). The potential energy of an isolated molecule is then a function of r alone.

If a transformation is made to generalized coordinates, generalized momenta conjugate to the new coordinates must be used. These momenta are defined as follows. Using the notation $\dot{q}_\nu$ for velocity along the coordinate q_ν, ($\dot{q} = dq_\nu/dt$), the kinetic energy, K, is evaluated as a function $K(q^{(\gamma)}\dot{q}^{(\gamma)})$ of the set $q^{(\gamma)}$ of the generalized coordinates, $q_1, q_2, \cdots, q_\nu, \cdots, q_\gamma$, and $\dot{q}^{(\gamma)}$ of the corresponding velocities. The momentum p_ν conjugate to the coordinate q_ν is $p_\nu = \partial K/\partial \dot{q}_\nu$. The Hamiltonian of (118) is then transformed into a function of the generalized coordinates and momenta.

Generalized coordinates

The product of a single coordinate q_ν with its conjugate momentum p_ν *is* $q_\nu p_\nu = q_\nu \partial K/\partial \dot{q}_\nu$. Because K is the kinetic energy, the physical dimensions of $q_\nu p_\nu$ are always those of energy times time or, equivalently, of mass times length squared divided by time, which are the dimensions (*i.e.*, physical units) of Planck's constant h. The γ-space of 2γ-dimensions made up of γ coordinates and their γ conjugate momenta therefore always has a volume element of physical dimensions h^γ, irrespective of choice of coordinates. In addition, the Jacobian that expresses the ratio of the volume element $dq_1 dp_1 \cdots dq_\gamma dp_\gamma$ in one set of coordinates and momenta, $q^{(\gamma)}$, $p^{(\gamma)}$, to that, $dq_1' dp_1' \cdots dq_\gamma' dp_\gamma'$, of another set, $q'^{(\gamma)}$, $p'^{(\gamma)}$, is always unity. This means that the integral of any function of the coordinates and momenta over a physically specified region of γ-space is invariant under a transformation from one coordinate-momenta set to another. The integral does, however, depend on the units of mass, length, and time.

One great advantage of the Hamiltonian formulation is that also the canonical equations of motion—two equations identifying the time derivatives q_ν and p_ν with partial derivatives of H with respect to p_ν and q_ν (see 121)—remain invariant in form under such transformations.

When applied to the Hamiltonian (118) expressed in terms of Cartesian coordinates and momenta, the first of equations (121) gives the familiar relation $\dot{x} = p_x/m$ between velocity $\dot{x} = v_x$ and momentum p_x along the x-axis; and the second equation expresses Newton's law of motion, because the partial differentials $-\partial H/\partial x = \partial U/\partial x = f_x$, the force along the x-axis, so that $\dot{p}_x$ is equal to f_x.

Liouville's theorem. With the aid of Hamilton's canonical equations of motion (see 121), it can be shown that the flow of any ensemble probability density in γ-space is like that of an incompressible fluid, in the following sense. At some initial time $t = 0$, a given volume in γ-space will contain a certain number of points, each representing a member-system of the ensemble. Liouville's famous theorem asserts that although the shape of the volume under consideration changes as time progresses, the volume enclosing these points and no others remains unchanged; and, hence, also the density of system points in γ-space remains constant, all in exact analogy with the density of particles constituting an incompressible fluid (see above *Fluid mechanics*).

To prove Liouville's theorem, which is of fundamental importance in statistical mechanics and in related parts of mathematical physics, the initial volume may be taken to be a 2γ-dimensional parallelepiped, so that (in analogy with the familiar three-dimensional case in which $\Delta V = \Delta x \Delta y \Delta z$) one can write $\Delta\omega = \Delta q_1 \Delta p_1 \Delta q_2 \Delta p_2, \cdots, \Delta q_\gamma \Delta p_\gamma$; or, more concisely, $\Delta\omega$ equals a product extended over γ terms of Δ of q_ν and Δ of p_ν (see 122). Here the symbol Π (pi) means "the product of all terms like this, a typical one." Defining the quantity Δa_ν as a product of

Δq_ν and Δp_ν (see 123), it is seen that the time-rate of change of Δa_ν is a sum of terms each involving a product of a time derivative with a nondifferentiated term (see 124). But by Hamilton's canonical equations (121), time derivatives of Δq_ν and Δp_ν are expressible in terms of second partial derivatives of H and undifferentiated terms (see 125). Substituting the expressions (125) into equation (124) shows that $\dot{\overline{\Delta a}}_\nu$ is equal to zero for every ν. It follows from this result and from equations (122) and (123) that time derivatives of $\Delta\omega$ and Δa_ν are zero (see 126); *i.e.*, the volume in γ-space remains constant in time, which completes the proof of the first part of the theorem. The second part, that the number of system points within $\Delta\omega$ remains constant, can now be established by showing that system points initially inside the volume cannot leave $\Delta\omega$, and similarly, those initially outside cannot enter $\Delta\omega$. In order to leave $\Delta\omega$, an initially interior point would evidently have to cross the surface S bounding $\Delta\omega$. This is impossible, however, because once such a point reached the surface, it would coincide with one of the system points defining S; *i.e.*, at that instant both points would have the same position and momentum in γ-space, so that by the laws of classical mechanics both would continue to move in identical manner: The initially interior point would always remain at the boundary and hence could never leave $\Delta\omega$. By exactly analogous reasoning it follows that also an initially exterior system point can never enter $\Delta\omega$, and thus the proof of Liouville's theorem is complete.

In general, the probability density W of an ensemble in γ-space will be a function, $W(t, q^{(\gamma)}, p^{(\gamma)})$, of time, t, and of the phase space variables. It follows from Liouville's theorem that W remains constant in time if followed along the path of motion of the systems. That is, the total derivative is zero (see 127). Using the equations of motion (121) to substitute for $\dot{q}_\nu$ and $\dot{p}_\nu$ in equation (127), it is found that a time rate of change of W, at a fixed position in γ-space, is equal to the summation over gamma terms of a difference of products of partial derivatives of the functions H and W (see 128).

If W, at each point $q^{(\gamma)}, p^{(\gamma)}$ in phase space, depends only on the Hamiltonian, $H(q^{(\gamma)}, p^{(\gamma)})$, then the partial derivatives of W are expressible in terms of partial derivatives of H (see 129), so that by equation (128) the partial derivative of W with respect to t is by calculation zero (see 130). Because the probability density of an ensemble of equilibrium systems must be independent of time, $\partial W/\partial t$ must be zero. It follows from the reasoning leading to equation (130) that an equation for W, in which its dependence on position in γ-space is through the Hamiltonian, is an acceptable solution.

In any case, W must be constant along the path actually traversed by any isolated system starting from any initial point in γ-space within the macrostate specification V, E, and N. Such a path is at constant value of the Hamiltonian.

The ergodic surmise

The notion that every path (in the mathematical sense of almost every path) covers every point in γ-space before returning to its initial starting point is known as the ergodic surmise. Despite the expenditure of much effort, it has thus far been impossible to prove in a manner useful in physics. Moreover, in some cases, the surmise is in fact incorrect. The nature of the actual physical system, however, is always adequate for the appropriate region of γ-space to be specified. Assuming the ergodic surmise, the probability density W must be constant for all γ-space corresponding to a given energy, that is, to a given value of the Hamiltonian H.

Equations for ensemble probability densities. For the three Gibbsian ensembles, the probability density for the microcanonical of fixed V, E, and N is a constant—the function W with arguments $q^{(\gamma)}$ and $p^{(\gamma)}$ is an exponential with argument the negative of S divided by k (see 131); for the petite canonical of fixed V, T, and N the function W with arguments $q^{(\gamma)}$ and $p^{(\gamma)}$ is an exponential with argument depending upon H as a function of $q^{(\gamma)}$ and $p^{(\gamma)}$ (see 132); and for the grand canonical ensemble described by V, T, and μ, W as the function of $q^{(\gamma)}$ and $p^{(\gamma)}$ is an exponential with three terms in the argument, one constituted of the function H (see 133). The thermodynamic functions $-S/k$, A/kT, and $-PV/kT$ are then determined by the condition that the sum (or integral) of each $W(q^{(\gamma)}, p^{(\gamma)})$ must be unity.

The last statement needs amplification. The functions W above do not give the probability density that molecule number one is at x_1, y_1, z_1, molecule number two

(123) $$\Delta a_\nu = \Delta q_\nu \, \Delta p_\nu$$

(124) $$\dot{\overline{\Delta a}}_\nu = \dot{\overline{\Delta q}}_\nu \Delta p_\nu + \Delta q_\nu \dot{\overline{\Delta p}}_\nu$$

(125)
$$\begin{cases} \dot{\overline{\Delta q}}_\nu = \dfrac{\partial H(\cdots q_\nu + \Delta q_\nu \cdots)}{\partial p_\nu} - \dfrac{\partial H(\cdots q_\nu \cdots)}{\partial p_\nu} \\ \qquad = \dfrac{\partial^2 H}{\partial p_\nu \partial q_\nu} \Delta q_\nu \\ \dot{\overline{\Delta p}}_\nu = -\dfrac{\partial H(\cdots p_\nu + \Delta p_\nu \cdots)}{\partial q_\nu} + \dfrac{\partial H(\cdots p_\nu \cdots)}{\partial q_\nu} \\ \qquad = -\dfrac{\partial^2 H}{\partial p_\nu \partial q_\nu} \Delta p_\nu \end{cases}$$

(126) $$\dot{\overline{\Delta\omega}} = \Delta\omega \sum_{\nu=1}^{\nu=\gamma} (\dot{\overline{\Delta a}}_\nu / \Delta a_\nu) = 0$$

(127) $$\dot{W} = \frac{dW}{dt} = \frac{\partial W}{\partial t} + \sum_{\nu=1}^{\nu=\gamma} \left[\left(\frac{\partial W}{\partial q_\nu}\right) \dot{q}_\nu + \left(\frac{\partial W}{\partial p_\nu}\right) \dot{p}_\nu \right] = 0$$

(128) $$\frac{\partial W}{\partial t} = \sum_{\nu=1}^{\nu=\gamma} \left(\frac{\partial H}{\partial q_\nu}\frac{\partial W}{\partial p_\nu} - \frac{\partial H}{\partial p_\nu}\frac{\partial W}{\partial q_\nu} \right)$$

(129)
$$\begin{cases} \partial W/\partial p_\nu = (\partial H/\partial p_\nu)\, dW/dH \\ \partial W/\partial q_\nu = (\partial H/\partial q_\nu)\, dW/dH \end{cases}$$

(130) $$\frac{\partial W}{\partial t} = \sum_{\nu=1}^{\nu=\gamma} \left[\frac{\partial H}{\partial q_\nu}\frac{\partial H}{\partial p_\nu} - \frac{\partial H}{\partial p_\nu}\frac{\partial H}{\partial q_\nu} \right] \frac{dW}{dH} = 0$$

(131) $$W(q^{(\gamma)}, p^{(\gamma)}) = \exp[-S/k]$$

(132) $$W(q^{(\gamma)}, p^{(\gamma)}) = \exp[(1/kT)\{A - H(q^{(\gamma)}, p^{(\gamma)})\}]$$

(133) $$W(q^{(\gamma)}, p^{(\gamma)}) = \exp[(1/kT)\{-PV + N\mu - H(q^{(\gamma)}, p^{(\gamma)})\}]$$

(134)
$$\begin{cases} Q(V, E, N) = \\ \qquad \lim_{\Delta E \to 0} \{\Delta E^{-1} \int\int\cdots\int (N!h^\gamma)^{-1}\, dq^{(\gamma)} dp^{(\gamma)}\} \\ E \leqslant H(q^{(\gamma)} p^{(\gamma)}) \leqslant E + \Delta E \end{cases}$$

(135) $$Q(V, T, N) = \int_0^\infty dE\, Q(V, E, N) e^{-E/kT}$$

(136) $$Q(V, T, \mu) = \sum_{N=0}^{N=\infty} Q(V, T, N) e^{N\mu/kT}$$

(137) $$S = k \ln Q(V, E, N)$$

(138) $$A = -kT \ln Q(V, T, N)$$

(139) $$PV = kT \ln Q(V, T, \mu)$$

(140) $$Q(V, T, N) = \int (N!h^\gamma)^{-1} dq^{(\gamma)} dp^{(\gamma)} \exp[-(1/kT) H(q^{(\gamma)} p^{(\gamma)})]$$

(141) $$\mathfrak{H}\Psi_K = E_K \Psi_K$$

at x_2, y_2, z_2, $\cdots$, number N at x_N, y_N, z_N, with the corresponding momentum and internal coordinate and momentum values. What they do give are the probability densities that the position sets labeled 1, 2, $\cdots$, N are occupied by molecules, and no other molecules are in volume V. For every such assignment of position in γ-space, there are $N!$ (*e.g.*, 4! equals $4 \times 3 \times 2 \times 1$) identical points differing only in the order of labelling the coordinates and momenta. Hence the integral over the phase space must be divided by $N!$ If the system comprises different species, a, b, $\cdots$ of molecules, then the integral must be divided by $N_a!\, N_b! \cdots$.

As remarked above, the volume element in γ-space depends on the units used for mass, length, and time. Thus, the computed values of S/k and of A/kT equal to the difference $(E/kT) - (S/k)$ will depend on the physical units used. (In the case of the grand canonical ensemble, it is the value of μ for a given physical system that would change rather than that of pressure P.) This is consistent with the concepts of classical thermodynamics in which only differences of entropy are defined. Classically, entropy has no absolute value.

In the early part of the 20th century, a German physical chemist, W.H. Nernst, postulated that entropy has an absolute value such that crystalline materials have zero entropy at the absolute zero of temperature. This is now well established and known as the third law of thermodynamics. It applies not only to crystals but also to liquid helium at 0° K and to hypothetical perfect gases at zero temperature. The law is a consequence of quantum mechanics. Instead of integrating over phase space, a summation should be performed over the quantum states of the system. The classical Hamiltonian, $H(q^{(\gamma)}, p^{(\gamma)})$, must be replaced by the energy, E_K, of the system in the quantum state K, and summations are to be made over all quantum states of the system. If, however, along every one of the γ degrees of freedom of the system, the energy change is small compared to kT for a change $\Delta q_\nu \Delta p_\nu$ equal to Planck's constant h in the values of the coordinate and momentum, then the classical equations are asymptotically correct (correct in the limit as the energy change divided by kT goes to zero) if the integral is divided by h^γ. This follows from the fact that in a volume $\Delta\omega$ equal to $\Delta q_1 \Delta p_1 \times \Delta q_2 \Delta p_2 \times \Delta q_3 \Delta p_3 \times \cdots$, for which $\Delta q_\nu \Delta p_\nu$ is much greater than h for every ν-value from 1 up to and including γ, there are approximately $\Delta\omega / N! h^\gamma$ allowed quantum states.

Thus, the condition for semiclassical evaluation of the thermodynamic functions from any of the three ensemble probability equations (see 131–133) is to require that the integrals over γ-space, divided by $N! h^\gamma$, be unity.

Partition functions

The results, usually called partition functions and denoted by Q, are functions of the sets of variables defining the thermodynamic state of the systems in the microcanonical (see 134), petite canonical (see 135), and grand canonical (see 136) ensembles. The dimensionless thermodynamic functions S/k, $-A/kT$, and PV/kT, which take maximum values at equilibrium for the given state variables, are the natural logarithms (denoted ln; $\ln e^x = x$) of the corresponding partition functions. The equations (137), (138), and (139) follow.

Comparing the formal structures of equations (134), (135), and (136) might suggest that the first would be the easiest to use. This, however, is rarely the case, because integration within boundaries determined by the function $H(q^{(\gamma)}, p^{(\gamma)})$ is difficult to perform. The expression (135) can be written such that Q with arguments V, T, and N is an integral in which the exponential appears with the function H as argument (see 140), which is usually easier to evaluate than (134).

Justification of the equations. A rigorous derivation of the preceding equations is beyond the scope of this article. Hence, only an outline of one method is given here. It was stated above that for an isolated system of fixed V, E, and N, the probability density at equilibrium must be independent of position in γ-space. It can also be shown that $\ln Q(V, E, N)$ is proportional to the size of the system (for the sum of two independent systems, it is the sum of their individual ln Q's). Both position-independence and size-proportionality are properties of the entropy. If a natural process is permitted to occur in a previously constrained system of constant V, E, and N, such as opening a stopcock between two flasks of gas, making thermal contact between two previously isolated parts, or introducing a catalyst for a chemical reaction, then new regions of phase space are made accessible to the system. Thus, $\ln Q(V, E, N)$ always increases or (at worst) stays constant for such removal of a constraint; and so ln Q will be a maximum at equilibrium, as will S, by equation (137). Considering the exchange of energy through thermal contact between two systems, the sum of their energies remaining constant, the condition that $\ln Q(V, E, N)$ be a maximum requires that for the two systems to be at thermal equilibrium the quantity $\partial(k \ln Q)/\partial E$ must have the same value for both systems and that, in attaining this maximum, energy flows from that system for which the quantity $\partial(k \ln Q)/\partial E$ has the lower value to the system for which it is higher. It follows that $\partial(k \ln Q)/\partial E$ may be considered as the reciprocal of a possible temperature scale. Any two systems of equal $\partial(\ln Q)\partial E$ thus have the same temperature. Similarly, by considering volume exchange of two systems of equal temperature connected by a movable piston, $\partial(k \ln Q)/\partial V$ is found to be proportional to pressure. From the first law of thermodynamics for reversible processes, $T\, dS = dE + P\, dV$, it follows that the partial derivatives of the entropy S possess precisely these two characteristics; *i.e.*, $\partial S/\partial E = T^{-1}$ and $\partial S/\partial V = P/T$. Indeed, for a perfect gas, $\partial(\ln Q)/\partial E$ actually is $1/T$ and $\partial(k \ln Q)/\partial V$ is P/T. One therefore identifies $k \ln Q(V, E, N)$ with entropy (see 137).

For the other two ensembles, in which the systems are in contact with a reservoir, the case that the systems plus reservoir have constant V, E, and N is considered. The total entropy at equilibrium is a maximum, as before; from this condition equations (132) and (133) are derived for the ensembles of the systems alone.

Experimental values of the ratio of S/k for macroscopic systems are large. For normal laboratory conditions they are of order N, which will be 10^{20} or larger. Thus the partition function $Q(V, E, N)$ has prodigious values like $10^{10^{20}}$. This is essential for the present treatment because for sufficiently small numbers of molecules (N) the equations are inconsistent with thermodynamic relations. They are strictly compatible with these relations only in the limit as N approaches infinity. The discrepancies are numerically of order $(\ln N)/N$, which for N equal to 10^{20} is about 10^{-18}, a negligible quantity.

Quantum statistics. The mathematically more elegant formulation of quantum statistics uses the so-called density matrix formulation. A simpler formalism, however, is adequate for most equilibrium problems. The microstate of a quantum mechanical system is given by a single-valued continuous state function, $\Psi(q^{(\gamma)})$, of the coordinates. This function can always be expressed as a linear combination of the time-independent solutions Ψ_K of the Schrödinger equation. Using the symbol $\mathfrak{H}$ for the Hamiltonian quantum mechanical operator obtained from the classical Hamiltonian function $H(q, p)$ by replacing each p_ν by $(h/2\pi i)(\partial/\partial q_\nu)$ (see below *Quantum mechanics*) and using E_K for the energy of the system in the quantum state designated by K, the time-independent Schrödinger equation becomes a form in which the Hamiltonian quantum mechanical operator operating upon Ψ_K is equal to a constant times Ψ_K (see 141). The number K requires the specification of as many single quantum numbers as there are coordinates, the number of degrees of freedom, γ.

The Hamiltonian quantum mechanical operator

Not all formal solutions of equation (141) are physically allowed states. Only state functions, Ψ_K, that are antisymmetric under exchange of coordinates of any pair of identical fermions (protons, neutrons, or electrons), occur in nature. That is, the exchange of each such coordinate pair must change the function Ψ_K into its negative. If, however, the sum of protons, neutrons, and electrons in an atom or molecule is even (*i.e.*, $2n$ in

which n is any integer), the function is multiplied by $(-1)^{2n}$ equal to 1 when there is a pair exchange of identical atoms or molecules. The function is then said to be symmetric. Thus helium-4 ($^4_2$He), consisting of two protons, two neutrons, and two electrons, requires symmetric state functions, and helium-4 is said to be a boson and forms a Bose-Einstein system, whereas helium-3 ($^3_2$He), having one neutron less and hence an odd number of particles, forms a Fermi-Dirac system requiring antisymmetric functions.

Because the Hamiltonian is necessarily symmetric in exchange of identical entities, every one of the $N!$ functions obtained by making all $N!$ permutations on Ψ_K will also be a solution to (141), each with the same energy E_K. Only one linear combination of these, namely the sum of all-permuted functions, is symmetric. The sum with negative sign for odd numbers of pair permutations is antisymmetric. If, however, the original function is already symmetric in some pair exchanges, then fewer than $N!$ different functions will be produced by the permutations. Indeed, the antisymmetric sum will then be zero because, under these conditions, terms differing only in pair exchange at two mutually symmetrical positions will cancel pairwise.

Thus only (approximately) the fraction $1/N!$ of the solutions of equation (141) are allowed states. There are somewhat more than $1/N!$ for bosons demanding symmetric functions, and somewhat fewer for Fermi-Dirac systems. For both cases, the quantum mechanical analogue of equation (134) is Q as a function of Z, E, and N equal to a limit as ΔE approaches zero of a product of terms depending upon ΔE (see 142). Equation (135) is valid with (142) for the partition functions for $Q(V, E, N)$, but can now be written so that Q as a function of V, T, and N is a sum of exponentials (see 143), whence (136) becomes a form in which Q as a function of V, T, and μ is a double sum of exponential type (see 144). (J.E.Ma.)

ELECTROMAGNETIC THEORY

Electromagnetic theory is the science principally concerned with the causal interrelation between time-varying electric and magnetic fields and with the propagation of these fields in space and time. The theory is one of exceptional beauty and vastness of scope. The set of four famous Maxwell equations most elegantly and concisely summarizes the fundamentals of the entire theory. From these equations, James Clerk Maxwell proved that electromagnetic fields propagate *in vacuo* with a speed equal to that of light in empty space. His theory thus brought about one of those great (and rare) syntheses in physical science: The phenomena of electromagnetism and of optics, theretofore considered totally unrelated and distinct from each other (except by a few; see below), were for the first time demonstrated to be, in fact, manifestations of the same fundamental processes (*e.g.*, visible light is an electromagnetic wave). Moreover, though unknown to Maxwell, his theory also satisfies the stringent requirements (of Lorentz invariance) that were to be demanded 50 years later by Einstein's theory of relativity (see below). The profound significance of Maxwell's theory, *per se*, is closely paralleled by the vast range of its applicability. For it is now known that the electromagnetic-wavelength spectrum spans at least some 20 orders of magnitude, ranging in wavelengths from about 10^{-10} centimetre for nuclear gamma-rays, up to about 10^{10} centimetres for radio waves, with visible light constituting but a minute fraction.

Origin of the electromagnetic wave concept. Historically, the development of science has always been an essentially collective effort. Even the greatest of advances were made possible, in part, by the work of earlier investigators. In the case of mechanics, Isaac Newton, born in 1642, had drawn upon the work of his antecedents, notably on that of Galileo (died 1642). Similarly, Maxwell's theory, too, is the culmination in part of his predecessors' efforts. Already in 1752, the Swiss-born mathematician and physicist Leonhard Euler had envisioned all electrical, light, and gravitational phenomena as being associated with states of stress in a hypothetical, tenuous, all-pervading medium, the ether, in which these states were assumed capable of being propagated with one fixed velocity. Michael Faraday, an English experimental physicist of the 19th century, held similar views, but added the notion of electric lines of force originating on positive charges and terminating on negative charges of electricity. Magnetic lines of force, on the other hand, he held to form closed curves. He further postulated, with remarkable insight, that light consists of electric and magnetic lines of force in motion, and later suggested that both light and radiant heat might well consist of transverse vibrations of the stresses portrayed by the lines of force. In 1848, Gustav Robert Kirchhoff in Germany noted, as a coincidence, that the ratio between the electromagnetic and electrostatic units of charge is equal to the velocity of light; and in 1858, a German mathematician, Bernhard Riemann, presented a paper at the Göttingen academy in which he assumed a finite velocity of propagation of electric effects and deduced that it must be equal to the same ratio of units.

Faraday's concept of lines of force

Maxwell's field theory. *Basic concepts.* The following definitions and mathematical theorems are used in the formulation of Maxwell's equations. The electric field (or electric intensity) $\boldsymbol{E}$ is defined as the force per unit electric charge held fixed at a given point (it is often printed in boldface type because it is a vector, having direction). Electric charge in motion is said to constitute an electric current (or conduction current) $\boldsymbol{I}$, the magnitude of which is determined by the amount of charge passing through a given surface area per unit time. If the surface (S) is plane and perpendicular to the velocity of the moving charge, then $\boldsymbol{I}/S$ is called the average current density over S. In the limit as S tends to zero at some point P, the ratio $\boldsymbol{I}/S$ is called the current density ($\boldsymbol{J}$) at P. Conduction currents are typically found in metals, in which there are many electrons free to move under the impressed $\boldsymbol{E}$.

In other materials, called dielectrics, the electrons remain ordinarily bound to their parent-atom or -molecule. If an electric field is applied to such materials, some of the electrons will, nevertheless, become displaced relative to the positively charged nuclei of their parent-atoms. Furthermore, if a thin metal plate is introduced into the dielectric (or into empty space), then, under the influence of the electric field, some of the electrons in the plate will be displaced from one face to the other. For some particular orientation of the plate, this displacement of electrons will be a maximum. The ratio of the charge, thus displaced, to the area of the plate (or the surface-charge density) is called the displacement vector $\boldsymbol{D}$ at that particular place in the field. In noncrystalline, isotropic media, $\boldsymbol{D}$ is proportional to the electric field $\boldsymbol{E}$, so that $\boldsymbol{D}$ equals a product of a constant ε times $\boldsymbol{E}$ (see 145), in which the proportionality constant ε is called the dielectric coefficient of the material.

Three other quantities, namely, $\boldsymbol{B}$, $\boldsymbol{H}$, and μ, are similarly related: $\boldsymbol{B}$ equals a product of a constant times $\boldsymbol{H}$ (see 146). Here, $\boldsymbol{H}$ is called the magnetic field intensity; μ is the permeability of the medium; and $\boldsymbol{B}$, called the magnetic induction, may be defined by the equation $\boldsymbol{F}_m = q\boldsymbol{v} \times \boldsymbol{B}$, in which $\boldsymbol{F}_m$ is the force experienced by the charge q moving with velocity $\boldsymbol{v}$ in the presence of $\boldsymbol{B}$. The quantity $\boldsymbol{v} \times \boldsymbol{B}$, called the vector product (or cross product) of $\boldsymbol{v}$ and $\boldsymbol{B}$, is equal to the product of their respective magnitudes, v and B, and the sine of the angle θ between them, and is a vector in the direction of $\boldsymbol{n}$—*i.e.*, vB (sin θ) $\boldsymbol{n}$—here $\boldsymbol{n}$ is a unit vector perpendicular to both $\boldsymbol{v}$ and $\boldsymbol{B}$.

Two further definitions and two theorems, all involving vectors, are needed. These are as follows. The del (for delta) operator, which, in Cartesian coordinates, is a formal sum of products of unit vectors times partial derivative operators (see 147), in which $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$ are unit vectors that are directed along the positive x-, y-, and z-axis, respectively.

The scalar product (or dot product) of two vectors A, B, is the product of the magnitudes A and B and the cosine of the angle θ between the two vectors (see 148). Using relations (147) and (148), Gauss's theorem asserts the equivalence of two integrals involving a closed surface, in which A is a vector; S is the closed surface, of which dS is a typical element; and τ is the volume enclosed by S (see 149). This relation allows a volume integral to be transformed into a surface integral, and *vice versa*. A similar relation, Stokes's theorem, permits a surface integral to be transformed into a line integral, and conversely (see 150). Here the integral on the left is over the closed curve s, of which ds is a typical element. The surface S is arbitrary except that it must be bounded by the curve s.

Maxwell's equations. The four celebrated equations (in differential form) are described as follows (see 151–154). Interpreting each of these briefly, equation 151 asserts that (in the parlance of *Fluid mechanics*, above) the divergence (div or $\nabla\cdot$) of the lines of displacement D is equal to the electric charge density ρ; *i.e.*, the charge contained in unit volume. Recalling that divergence refers to the net time-rate of outflow per unit volume, it is seen that there will be a net number of lines of D leaving (or entering) a given volume τ provided that the latter contains a net positive (or negative) amount of electric charge. If ρ is zero, then div D is also zero; that is to say, the net number of lines of D issuing from (or entering into) the volume is zero, and conversely.

Equation (152) states that the divergence of the magnetic induction (B) is always zero: The net number of lines of B leaving a given volume always vanishes. Geometrically, this will happen if, as Faraday had suggested, the lines of B form closed curves. Physically, it means that the magnetic charge density ρ_m is always zero; that is, there are no single isolated magnetic poles: north poles and south poles always occur in pairs, in contradistinction to the well-known existence of positive electric charge (as on a proton), separate and apart from negative charge (as on an electron).

Absence of isolated magnetic poles

According to equation (153) (in which the dot denotes differentiation with respect to time), the existence of a time-varying magnetic induction (B) will give rise to an electric field (E). In beautiful formal symmetry, equation (154) similarly informs that the presence of a time-varying displacement D (the displacement current $\dot{D}$) and a conduction current J gives rise to a magnetic field H. (The fundamental importance of the term $\dot{D}$ is explained below.)

The basis and derivation of each of equations (151) to (154) follow. According to Gauss's law (itself crucially dependent upon Coulomb's law that the force between two point-charges q_1, q_2, a distance r apart, k being a constant the numerical value of which depends upon the units used, is $F = kq_1q_2/r^2$), thoroughly verified experimentally, the net number of lines of D leaving any closed surface S is numerically equal to the net electric charge contained within the volume bounded by S. Mathematically, the integral around S of D dotted with the differential of S is the integral over τ of ρ (see 155), all symbols having their previously defined meanings. Applying Gauss's theorem (149) to the left-hand member of equation (155) gives a final equality (see 156), whence, by considering the volume itself to be arbitrarily small, it follows that the integrands are equal (see 151).

The observational fact that the displacement lines of B are always closed means that the net number of lines issuing from or entering a closed surface S is zero (see 157), which, by another application of equation (149), can also be written as an integral over τ of $\nabla\cdot B$ integrated with respect to the differential of τ equalling zero (see 158). If τ again is chosen to be an infinitesimal volume, it follows that the divergence of the magnetic induction is zero, as asserted by equation (152).

According to Faraday's law of electromagnetic induction, the electromotive force E is equal to the negative time-rate of change of flux Φ; *i.e.*, $E = -\Phi$. Substituting for E and Φ their respective defining expressions gives an equality between integrals, one a line integral of E dotted into the differential of s, one the negative of the integral over S of the time derivative of B dotted into the differential of S (see 159). Applying Stokes's theorem (150) allows (159) to be written as an equality between two integrals over S (see 160), so that choice of infinitesimal dS assures equality of the integrands, as stated in equation (153).

The empirical basis for (154) is given in part by Ampère's circuital law, according to which the line integral involving H is equal to the surface integral involving J (see 161). Treating the left-hand integral as in (159) above gives $\nabla\times H$ equal to J (see 162), which is not quite equation (154). The reason for this is that, un-

(142) $$Q(V,E,N)=\lim_{\Delta E\to 0}\left[\Delta E^{-1}\times\left(\begin{matrix}\text{Number of allowed states } K\\ \text{for which } E\leqslant E_K=E+\Delta E\end{matrix}\right)\right]$$

(143) $$Q(V,T,N)=\sum_K e^{-E_K/kT}$$

(144) $$Q(V,T,\mu)=\sum_N\sum_K e^{+(1/kT)(\mu N-E_K)}$$

(145) $$D=\varepsilon E$$

(146) $$B=\mu H$$

(147) $$\nabla=\mathbf{i}\frac{\partial}{\partial x}+\mathbf{j}\frac{\partial}{\partial y}+\mathbf{k}\frac{\partial}{\partial z}$$

(148) $$A\cdot B=AB\cos\theta$$

(149) $$\iint_S A\cdot dS=\int_\tau \nabla\cdot A\,d\tau$$

(150) $$\oint_s A\cdot ds=\int_S \nabla\times A\cdot dS$$

(151) $$\nabla\cdot D=\rho$$

(152) $$\nabla\cdot B=0$$

(153) $$-\dot{B}=\nabla\times E$$

(154) $$\dot{D}+J=\nabla\times H$$

(155) $$\oint_S D\cdot dS=\int_\tau \rho\,d\tau$$

(156) $$\int_\tau \nabla\cdot D\,d\tau=\int_\tau \rho\,d\tau$$

(157) $$\int_S B\cdot dS=0$$

(158) $$\int_\tau \nabla\cdot B\,d\tau=0$$

(159) $$\oint_s E\cdot ds=-\int_S \dot{B}\cdot dS$$

(160) $$\int_S \nabla\times E\cdot dS=\int_S -\dot{B}\cdot dS$$

(161) $$\oint_s H\cdot ds=\int_S J\cdot dS$$

(162) $$\nabla\times H=J$$

known until Maxwell discovered and remedied it, Ampère's law (160) is not correct. (The particular conditions prevailing during Ampère's experimentation were such as to make the error too small to be noticed.) That his law is inconsistent with the equation of continuity (see 163; see also above *Mechanics of fluids*) is readily seen by forming the divergence of the conduction current ($\nabla \cdot \boldsymbol{J}$) using equation (162). This gives $\nabla \cdot \boldsymbol{J}$ equal to zero (see 164); *i.e.*, the expression $\nabla \cdot \nabla \times \boldsymbol{A}$ is identically zero for any vector $\boldsymbol{A}$ (and hence also for $\boldsymbol{H}$). Yet, the right-hand sides of equations (163) and (164) should not equal zero, because equation (151) after differentiation with respect to time, shows that ∇ dotted into the time derivative of $\boldsymbol{D}$ is equal to the time derivative of ρ (see 165), and hence $\dot{\rho}$ is different from zero if $\nabla \cdot \dot{\boldsymbol{D}}$ is. It was one of Maxwell's greatest contributions to electromagnetic theory that he correctly resolved this fundamental difficulty by introducing the displacement current $\dot{\boldsymbol{D}}$ into the right-hand side of Ampère's law (159), so that $\nabla \times \boldsymbol{H}$ is a sum of $\boldsymbol{J}$ and the time derivative of $\boldsymbol{D}$ (see 166).

Forming the divergence of the conduction current with this augmented form gives minus the divergence of the displacement current, or $\nabla \cdot \boldsymbol{J}$ is the negative of ∇ dotted into the time derivative of $\boldsymbol{D}$ (see 167); or, by equations (151) or (165), $\nabla \cdot \boldsymbol{J}$ is the negative of the time derivative of ρ (see 168), correctly, as demanded by the equation of continuity (163). It is seen that equation (166) is identical with (154).

Velocity of electromagnetic waves. It remains to show that the electric and magnetic fields, $\boldsymbol{E}$ and $\boldsymbol{B}$, do satisfy the wave equation and that the velocity of propagation of these waves, as predicted by Maxwell's theory, is indeed equal to the empirical 3×10^{10} centimetres per second (cm/sec).

To perform these calculations, Maxwell's equations (151) to (154) are used, with both ρ and $\boldsymbol{J}$ equal to zero, as appropriate for empty space. Thus differentiation of (154) with respect to time (in which the double dot denotes a second differentiation with respect to time) gives the second-order time derivative of $\boldsymbol{D}$ as a cross product between ∇ and the time derivative of $\boldsymbol{H}$ (see 169), or, using equations (146) and (153), the second-order time derivative of $\boldsymbol{D}$ is a cross product between ∇ and a constant times a cross product between ∇ and $\boldsymbol{E}$ (see 170), in which the subscript 0 refers to empty space. But by equation (145), $\boldsymbol{D} = \varepsilon_0 \boldsymbol{E}$, so the second derivative of $\boldsymbol{E}$ with respect to time is proportional to a double cross product, terms in the products being ∇, ∇, and $\boldsymbol{E}$ (see 171).

A well-known vector identity states that an equality between a double cross product and a difference of $\nabla(\nabla \cdot \boldsymbol{E})$ and the formal square of the ∇ operator exists (see 172), in which $\nabla^2 = \nabla \cdot \nabla$. But the first term on the right-hand side of equation (172) vanishes because, in equation (151), ρ is zero. Equation (171) therefore becomes a form in which ∇^2 operating upon $\boldsymbol{E}$ is proportional to the second derivative of $\boldsymbol{E}$ with respect to time (see 173), which indeed is the wave equation; *i.e.*, the differential equation known to have solutions of the form in which the x component of $\boldsymbol{E}$ is the sum of two terms each with a factor f_i and each with a factor involving x, v, and t; similar expressions hold for other components (see 174), representing transverse waves propagating with velocity v given by the coefficient of $\ddot{\boldsymbol{E}}$ raised to the power $-\frac{1}{2}$, v is the square root of the reciprocal of the product of ε_0 and μ_0 (see 175). Substitution of the appropriate numerical values for ε_0, μ_0 gives $v = 3 \times 10^{10}$ identical with the observed value of the velocity of light in empty space.

The differential equations for velocity

To show also that $\boldsymbol{B}$ obeys the wave equation, equation (153) is differentiated with respect to time, giving the second derivative with respect to time of $\boldsymbol{B}$ equal to the cross product between ∇ and the time derivative of $\boldsymbol{E}$ (see 176). But by equations (145) and (154), again with $\boldsymbol{J}$ equal to 0, ε_0, ∇, times the time derivative of $\boldsymbol{E}$ is the cross product between ∇ and $\boldsymbol{H}$ (see 177), so that, using equation (146), an equality between the negative of the second derivative of $\boldsymbol{B}$ and a triple cross product involving $\frac{1}{\varepsilon_0}$, ∇, $\boldsymbol{B}$, and μ_0 is found (see 178). Application of the vector identity (172) and of the definition (146) gives a form in which ∇^2 operating on $\boldsymbol{B}$ is a constant times the second derivative of $\boldsymbol{B}$ (see 179), showing that also the magnetic induction $\boldsymbol{B}$ does satisfy the wave equation and is propagated with the same speed v, as given in (175). This concludes the proof.

For more detailed accounts, and for the treatment of other parts of the theory, the reader is referred to the articles ELECTRICITY and ELECTROMAGNETIC RADIATION and also to the various references given in the *Bibliography*. (Ed.)

RELATIVITY THEORY

Maxwell's electromagnetic theory (see above), published in England in 1855, had brilliantly unified not only electricity and magnetism, but also electromagnetism and optics. Closely related, and in a manner remarkably parallel to this, there followed Einstein's theory of relativity, published in Germany beginning in 1905, demonstrating the fundamental unity not only of space and time, but also that of mass and energy. In this section, edited from an article originally written by Einstein for the 13th edition of *Encyclopædia Britannica*, the concepts of space and time are presented. Other mathematical aspects of the theory, namely Riemannian geometry; Lorentz-transformation, -invariance, -contraction, and time-dilation, are treated in three sections that follow below. (See also the article on RELATIVITY.) (Ed.)

All thoughts and concepts are called up by sense experiences and have a meaning only in reference to these sense experiences. On the other hand, however, they are products of the spontaneous activity of the human mind; they are thus in no wise logical consequences of the contents of these sense experiences. Therefore, to grasp the essence of a complex of abstract notions it is necessary for the one part to investigate the mutual relationships between the concepts and the assertions made about them; for the other, to investigate how they are related to the experiences.

As far as the way is concerned in which concepts are connected with one another and with the experiences, there is no difference of principle between the concept systems of science and those of daily life. The concept systems of science have grown out of those of daily life and have been modified and completed according to the objects and purposes of the science in question.

The more universal a concept is, the more frequently it enters into human thought; and the more indirect its relation to sense experience, the more difficult it is to comprehend its meaning; this is particularly the case with prescientific concepts that are customarily used from the time of childhood. Consideration of the concepts referred to in the words "where," "when," "why," "being," to the elucidation of which innumerable volumes of philosophy have been devoted, illustrate the difficulty. The philosopher fares no better in speculations than a fish which should strive to become clear as to what is water.

Space. In the present section the meaning of "where," that is, of space, is of concern. It appears that there is no quality contained in individual primitive sense experiences that may be designated as spatial. Rather, what is spatial appears to be a sort of order of the material objects of experience. The concept "material object" must therefore be available if concepts concerning space are to be possible. It is the logically primary concept. This is easily seen if an analysis is made of the spatial concepts, for example, "next to," "touch," and so forth, that is, if their equivalents in experience are sought. The concept "object" is a means of taking into account the persistence in time or the continuity, respectively, of certain groups of experience complexes. The existence of objects is thus of a conceptual nature, and the meaning of the concepts of objects depends wholly on their being connected (intuitively) with groups of elementary sense experiences. This connection is the basis of the illusion that makes primitive

experience appear to provide information directly about the relation of material bodies (which exist, after all, only insofar as they are thought).

In the sense thus indicated there is (the indirect) experience of the contact of two bodies. Attention only is called to this; nothing is gained for the present purpose by singling out the individual experiences to which this assertion alludes. Many bodies can be brought into permanent contact with one another in manifold ways. In this sense the term "position relationships" of bodies is used. The general laws of such position relationships are essentially the concern of geometry. This holds, at least, if it is not desired to regard the propositions that occur in this branch of knowledge merely as relationships between empty words that have been set up according to certain principles.

The notion of "space" in prescientific thought

Now, it is important to examine the meaning of the concept "space" which is also encountered in prescientific thought. The concept of space in prescientific thought is characterized by the sentence: "It is possible to think away things but not the space which they occupy." It is as if, without having had experience of any sort, man had a concept, even a presentation, of space; and as if he ordered his sense experiences with the help of this concept present, a priori. On the other hand, space appears as a physical reality, as a thing which exists independently of thought, like material objects. Under the influence of this view of space the fundamental concepts of geometry, the point, the straight line, the plane, were even regarded as having a self-evident character. The fundamental principles that deal with these configurations were regarded as being necessarily valid and as having at the same time an objective content. No scruples were felt about ascribing an objective meaning to such statements as "three empirically given bodies (practically infinitely small) lie on one straight line," without demanding a physical definition for such an assertion. This blind faith in evidence and in the immediately real meaning of the concepts and propositions of geometry became uncertain only after non-Euclidean geometry had been introduced.

If a beginning is made from the view that all spatial concepts are related to contact experiences of solid bodies, it is easy to understand how the concept "space" originated, namely, how a thing independent of bodies and yet embodying their position possibilities was posited. If there is a system of bodies in contact and at rest relatively to one another, some can be replaced by others. This property of allowing substitution is interpreted as "available space." Space denotes the property in virtue of which rigid bodies can occupy different positions. The view that space is something with a unity of its own is perhaps due to the circumstance that in prescientific thought all positions of bodies were referred to one body (reference body), namely the earth. In scientific thought the earth is represented by the coordinate system. The assertion that it would be possible to place an unlimited number of bodies next to one another denotes that space is infinite. In prescientific thought the concepts "space" and "time" and "body of reference" are scarcely differentiated at all. A place or point in space is always taken to mean a material point on a body of reference.

The notion of "distance" in Euclidean geometry

If Euclidean geometry is considered, it is seen that it refers to the laws regulating the positions of rigid bodies. It turns to account the ingenious thought of tracing back all relations concerning bodies and their relative positions to the very simple concept "distance." Distance denotes a rigid body on which two material points (marks) have been specified. The concept of the equality of distances (and angles) refers to experiments involving coincidences; the same remarks apply to the theorems on congruence. Now, Euclidean geometry, in the form in which it has been handed down from Euclid, uses the fundamental concepts "straight line" and "plane" which do not appear to correspond, or at any rate, not so directly, with experiences concerning the position of rigid bodies. (On this it must be remarked that the concept of the straight line may be reduced to that of distance. A hint of this is contained in the theorem: "The straight line is the shortest connection between two points." This theorem served well as a definition of the straight line, although the definition played no part in the logical texture of the deductions.) Moreover, geometricians were less concerned with bringing out the relation of their fundamental concepts to experience than with deducing logically the geometrical propositions from a few axioms enunciated at the outset.

A brief outline may be made to show how perhaps the basis of Euclidean geometry may be gained from the concepts of distance. To start from the equality of distances (the axiom of the equality of distances) makes it possible to suppose that of two unequal distances one is always greater than the other. The same axioms are to hold for the inequality of distances as hold for the inequality of numbers. Three distances $\overline{AB'}$, $\overline{BC'}$, $\overline{CA'}$ may, if CA' be suitably chosen, have their marks BB', CC', AA' superposed on one another in such a way that a triangle ABC results. The distance CA' has an upper limit for which this construction is still just possible. The points A, (BB') and C then lie in a "straight line" (definition). This leads to the concepts: producing a distance by an amount equal to itself; dividing a distance into equal parts; expressing a distance in terms of a number by means of a measuring rod (definition of the space interval between two points).

When the concept of the interval between two points or the length of a distance has been gained in this way, it is necessary to require only the following axiom in order to arrive at Euclidean geometry analytically. To every point of space (body of reference) three num-

$$\nabla \cdot J = -\dot{\rho} \tag{163}$$

$$\nabla \cdot J = \nabla \cdot \nabla \times H \equiv 0 \tag{164}$$

$$\nabla \cdot \dot{E} = \dot{\rho} \tag{165}$$

$$\nabla \times H = J + \dot{D} \tag{166}$$

$$\nabla \cdot J = -\nabla \cdot \dot{D} \tag{167}$$

$$\nabla \cdot J = -\dot{\rho} \tag{168}$$

$$\ddot{D} = \nabla \times \dot{H} \tag{169}$$

$$\ddot{D} = \nabla \times \left(-\frac{1}{\mu_0} \nabla \times E \right) \tag{170}$$

$$\varepsilon_0 \ddot{E} = -\frac{1}{\mu_0} \nabla \times (\nabla \times E) \tag{171}$$

$$\nabla \times (\nabla \times E) = \nabla(\nabla \cdot E) - \nabla^2 E \tag{172}$$

$$\nabla^2 E = (\varepsilon_0 \mu_0) \ddot{E} \tag{173}$$

$$E_x = f_1(x - vt) + f_2(x + vt), \cdots \tag{174}$$

$$v = (1/\varepsilon_0 \mu_0)^{1/2} \tag{175}$$

$$-\ddot{B} = \nabla \times \dot{E} \tag{176}$$

$$\varepsilon_0 \dot{E} = \nabla \times H \tag{177}$$

$$-\ddot{B} = \nabla \times \left(\frac{1}{\varepsilon_0} \nabla \times \frac{B}{\mu_0} \right) \tag{178}$$

$$\nabla^2 B = (\varepsilon_0 \mu_0) \ddot{B} \tag{179}$$

bers (coordinates) x, y, z may be assigned—and conversely—in such a way that for each pair of points A (x_1, y_1, z_1) and B (x_2, y_2, z_2) Pythagoras' theorem holds (see 180).

All further concepts and propositions of Euclidean geometry can then be built up purely logically on this basis, in particular also, the propositions about the straight line and the plane. These remarks are not, of course, intended to replace the strictly axiomatic construction of Euclidean geometry. It is merely wished to indicate plausibly how all conceptions of geometry may be traced back to that of distance. It would have been equally well to have epitomized the whole basis of Euclidean geometry in the last theorem above. The relation to the foundations of experience would then be furnished by means of a supplementary theorem. The coordinate may and *must* be chosen so that two pairs of points separated by equal intervals, as calculated by the help of Pythagoras' theorem, may be made to coincide with one and the same suitably chosen distance (on a solid). The concepts and propositions of Euclidean geometry may be derived from Pythagoras' proposition without the introduction of rigid bodies; but these concepts and propositions would not then have contents that could be tested. They are not "true" propositions but only logically correct propositions of purely formal content.

A serious difficulty is encountered in the above represented interpretation of geometry in that the rigid body of experience does not correspond *exactly* with the geometrical body. There are no absolutely definite marks, and, moreover, temperature, pressure, and other circumstances modify the laws relating to position. It is also to be recollected that the structural constituents of matter (such as atom and electron) assumed by physics are not in principle commensurate with rigid bodies, but that nevertheless the concepts of geometry are applied to them and to their parts. For this reason consistent thinkers have been disinclined to allow real contents of facts to correspond to geometry alone. They considered it preferable to allow the content of experience to correspond to geometry and physics conjointly.

This view is certainly less open to attack than the one represented above; as opposed to the atomic theory, it is the only one that can be consistently carried through. Nevertheless, it would not be advisable to give up the first view, from which geometry derives its origin. This connection is essentially founded on the belief that the ideal rigid body is an abstraction that is well rooted in the laws of nature.

The question arises: what is a priori certain or necessary, respectively, in geometry (doctrine of space) or its foundations? Formerly it was thought everything; nowadays it is thought nothing. Already the distance concept is logically arbitrary; there need be no things that correspond to it, even approximately. Something similar may be said of the concepts straight line, plane, of three-dimensionality and of the validity of Pythagoras' theorem. Even the continuum doctrine is in no wise given with the nature of human thought, so that from the epistemological point of view no greater authority attaches to the purely topological relations than to the others.

It is necessary yet to deal with those modifications in the space concept which have accompanied the advent of the theory of relativity. For this purpose the space concept of the earlier physics can be considered from a point of view different from that above. If the theorem of Pythagoras to infinitely near points is applied, it involves ds (see 181), in which ds denotes the measurable interval between them. For an empirically given ds the coordinate system is not yet fully determined for every combination of points by this equation. Besides being translated, a coordinate system may also be rotated. This signifies analytically: The relations of Euclidean geometry are covariant with respect to linear orthogonal transformations of the coordinates.

In applying Euclidean geometry to pre-relativistic mechanics, a further indeterminateness enters through the choice of the coordinate system: the state of motion of the coordinate system is arbitrary to a certain degree, namely, in that substitutions of the coordinates (see 182) also appear possible. On the other hand, earlier mechanics did not allow coordinate systems to be applied of which the states of motion were different from those expressed in these equations. In this sense the term "inertial systems" is used. In these favoured inertial systems there is a new property of space as far as geometrical relations are concerned. Regarded more accurately, this is not a property of space alone but of the four-dimensional continuum consisting of time and space conjointly.

Interrelationship of space and time

Time. At this point, time enters explicitly into the discussion for the first time. In their applications space (place) and time always occur together. Every event that happens in the world is determined by the space coordinates x, y, z and the time coordinate t. Thus the physical description was four-dimensional right from the beginning. But this four-dimensional continuum seemed to resolve itself into the three-dimensional continuum of space and the one-dimensional continuum of time. This apparent resolution owed its origin to the illusion that the meaning of the concept "simultaneity" is self-evident, and this illusion arises from the fact that news is received of near events almost instantaneously owing to the agency of light.

This faith in the absolute significance of simultaneity was destroyed by the law regulating the propagation of light in empty space or, respectively, by the Maxwell-Lorentz electrodynamics. Two infinitely near points can be connected by means of a light signal if a simple relation (see 183) holds for them. It further follows that ds has a value that, for arbitrarily chosen infinitely near space-time points, is independent of the particular inertial system selected. In agreement with this, it is found that, for passing from one inertial system to another, linear equations of transformation hold which do not in general leave the time values of the events unchanged. It thus became manifest that the four-dimensional continuum of space cannot be split up into a time continuum and a space continuum except in an arbitrary way. This invariant quantity ds may be measured by means of measuring rods and clocks.

The four-dimensional continuum

On the invariant ds a four-dimensional geometry may be built up which is in a large measure analogous to Euclidean geometry in three dimensions. In this way physics becomes a sort of statics in a four-dimensional continuum. Apart from the difference in the number of dimensions, the latter continuum is distinguished from that of Euclidean geometry in that ds^2 may be greater or less than zero. Corresponding to this, differentiation is made between timelike and spacelike line elements. The boundary between them is marked out by the element of the "light-cone" $ds^2 = 0$ which starts out from every point. If only elements which belong to the same time value are considered, an expression for $-ds^2$ is obtained that is formally identical with that of ds^2 given by the Pythagorean theorem (see 181). These elements ds may have real counterparts in distances at rest and, as before, Euclidean geometry holds for these elements.

This is the modification which the doctrine of space and time has undergone through the restricted theory of relativity. The doctrine of space has been still further modified by the general theory of relativity, because this theory denies that the three-dimensional spatial section of the space-time continuum is Euclidean in character. Therefore, it asserts that Euclidean geometry does not hold for the relative positions of bodies that are continuously in contact.

The empirical law of the equality of inertial and gravitational mass led to the interpretation of the state of the continuum, insofar as it manifests itself with reference to a noninertial system, as a gravitational field, and to the treatment of noninertial systems as equivalent to inertial systems. Referred to such a system, which is connected with the inertial system by a nonlinear transformation of the coordinates, the metrical in-

variant ds^2 assumes a general form (see 184) in which certain functions $g_{\mu\nu}$ are functions of the coordinates and in which the sum extended over the indices of these functions is to be taken over the indices for all combinations 11, 12, · · · , 44. The variability of the $g_{\mu\nu}$'s is equivalent to the existence of a gravitational field. If the gravitational field is sufficiently general, it is not possible at all to find an inertial system, that is, a coordinate system with reference to which ds^2 may be expressed in the simple form already given (see 185); but in this case, too, there is in the infinitesimal neighbourhood of a space-time point a local system of reference for which the last-mentioned simple form for ds holds. This state of the facts leads to a type of geometry which Riemann's genius created more than half a century before the advent of the general theory of relativity of which Riemann divined the high importance for physics.

Riemann's geometry of an n-dimensional space bears the same relation to Euclidean geometry of an n-dimensional space as the general geometry of curved surfaces bears to the geometry of the plane. For the infinitesimal neighbourhood of a point on a curved surface, there is a local coordinate system in which the distance ds between two infinitely near points is given by an equation (see 186). For any arbitrary (Gaussian) coordinate system, however, an expression of a known form (see 187) holds in a finite region of the curved surface. If the $g_{\mu\nu}$'s are given as functions of x_1 and x_2, the surface is then fully determined geometrically. For from this formula it is possible to calculate for every combination of two infinitely near points on the surface the length ds of the minute rod connecting them; and with the help of this formula all networks that can be constructed on the surface with these little rods can be calculated. In particular, the "curvature" at every point of the surface can be calculated; this is the quantity that expresses to what extent and in what way the laws regulating the positions of the minute rods in the immediate vicinity of the point under consideration deviate from those of the geometry of the plane.

This theory of surfaces by Gauss has been extended by Riemann to continua of any arbitrary number of dimensions and has thus paved the way for the general theory of relativity. For it was shown above that corresponding to two infinitely near space-time points there is a number ds which can be obtained by measurement with rigid measuring rods and clocks (in the case of timelike elements, indeed, with a clock alone). This quantity occurs in the mathematical theory in place of the length of the minute rods in three-dimensional geometry. The curves for which $\int ds$ has stationary values determine the paths of material points and rays of light in the gravitational field, and the "curvature" of space is dependent on the matter distributed over space.

Just as in Euclidean geometry the space concept refers to the position possibilities of rigid bodies, so in the general theory of relativity the space-time concept refers to the behaviour of rigid bodies and clocks. The space-time continuum, however, differs from the space continuum in that the laws regulating the behaviour of these objects (clocks and measuring rods) depend on where they happen to be. The continuum (or the quantities that describe it) enters explicity into the laws of nature, and conversely these properties of the continuum are determined by physical factors. The relations that connect space and time can no longer be kept distinct from physics proper. Nothing certain is known of what the properties of the space-time continuum may be as a whole. Through the general theory of relativity, however, the view that the continuum is infinite in its timelike extent but finite in its spacelike extent has gained in probability.

Subjective notions of time

The physical time concept derives from the time concept of the nonscientific mind. Now, the latter has its root in the time order of the experiences of the individual, and this order must be accepted as something primarily given. Man experiences the moment "now," or, expressed more accurately, the present sense experience combined with the recollection of (earlier) sense experiences. That is why the sense experiences seem to form a series, namely the time series indicated by "earlier" and "later." The experience series is thought of as a one-dimensional continuum. Experience series can repeat themselves and can then be recognized. They can also be repeated inexactly, wherein some events are replaced by others without the character of the repetition becoming lost. In this way man forms the time concept as a one-dimensional frame that can be filled in by experiences in various ways. The same series of experiences answer to the same subjective time intervals.

The transition from this "subjective" time to the time concept of prescientific thought is connected with the formation of the idea that there is a real external world independent of the subject. In this sense the (objective) event is made to correspond with the subjective experience. In the same sense, there is attributed to the "subjective" time of the experience a "time" of the corresponding "objective" event. In contrast with experiences, external events and their order in time claim validity for all subjects.

This process of objectification would encounter no difficulties were the time order of the experiences corresponding to a series of external events the same for all individuals. In the case of the immediate visual perceptions of man's daily life, this correspondence is exact. That is why the idea that there is an objective time order became established to an extraordinary extent. In working out the idea of an objective world of external events in greater detail, it was found necessary to make events and experiences depend on each other in a more complicated way. This was at first done by means of rules and modes of thought instinctively gained, in which the conception of space plays a particularly prominent part. This process of refinement leads ultimately to natural science.

The measurement of time is effected by means of clocks. A clock is a thing which automatically passes in succession through a (practically) equal series of events (period). The number of periods (clock time) elapsed serves as a measure of time. The meaning of this definition is at once clear if the event occurs in the immediate vicinity of the clock in space; for all observers then observe the same clock time simultaneously with the event (by means of the eye) independently of their position. Until the theory of relativity was propounded it was assumed that the conception of simultaneity had an absolute objective meaning also for events separated in space.

This assumption was demolished by the discovery of the law of propagation of light. For if the velocity of light in empty space is to be a quantity that is independent of the choice (or, respectively, of the state of motion) of the inertial system to which it is referred, no

(180) $$AB = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$$

(181) $$ds^2 = dx^2 + dy^2 + dz^2$$

(182) $$\begin{cases} x' = x - vt \\ y' = y \\ z' = z \end{cases}$$

(183) $$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2 = 0$$

(184) $$ds^2 = \sum_{\mu\nu} g_{\mu\nu}\, dx_\mu dx_\nu$$

(185) $$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$$

(186) $$ds^2 = dx^2 + dy^2$$

(187) $$ds^2 = g_{11}dx_1^2 + 2g_{12}dx_1dx_2 + g_{22}dx_2^2$$

absolute meaning can be assigned to the conception of the simultaneity of events that occur at points separated by a distance in space. Rather, a special time must be allocated to every inertial system. If no coordinate system (inertial system) is used as a basis of reference there is no sense in asserting that events at different points in space occur simultaneously. It is in consequence of this that space and time are welded together into a uniform four-dimensional continuum. (A.Ei.)

RIEMANNIAN GEOMETRY

As indicated in the preceding section, the geometry developed by Riemann plays a fundamental role in the mathematical formulation of relativity theory. This non-Euclidean geometry is concerned with the properties of a coordinate space $(x^1, \cdots, x^n)$ in which there is a nondegenerate quadratic differential form called the element of arc (see 188). This geometry reduces to Euclidean geometry if the element of arc takes the special form $ds^2 = (dx^1)^2 + \cdots + (dx^n)^2$. The two-dimensional case had been considered before Riemann by Carl Friedrich Gauss (born 1777) as the intrinsic geometry on a surface in ordinary Euclidean space. In pure mathematics, the differential form ds^2 is generally supposed to be positive definite, an assumption that is essential to many of the important consequences. In applications to general relativity, however, in which the Riemannian space is the physical universe, ds^2 is supposed to be hyperbolic; *i.e.*, reducible to a sum of squares minus the square of a linear differential form (see GEOMETRY, NON-EUCLIDEAN).

Basic properties. The quantity ds^2 in equation (188) allows the definition of the arc length of a curve $x^i = x^i(t)$ as an integral (see 189). Geodesics are curves that are shortest between any two of their points, sufficiently near to each other. An n-dimensional domain has a volume given by the integral $\int\sqrt{|g|}\, dx^1 \cdots dx^n$, in which g is the determinant of g_{ik}.

Geodesics

The coordinates x^i are local coordinates in the sense that they can be subjected to a general differentiable transformation $x'^i = f^i(x^1, \cdots, x^n)$, in which $i = 1, \cdots, n$, with nonvanishing Jacobian determinant (see GEOMETRY, DIFFERENTIAL). For ds^2 to be invariant under such changes of local coordinates, the quantities g_{ik} should transform according to certain equations (see 190) and are, therefore, the components of a symmetric covariant tensor field of order two. The elements g^{jl} of the inverse matrix of the matrix (g_{ik}), which are related to the elements g_{ik} by equations (see 191), are the components of a symmetric contravariant tensor field of order two. The fundamental tensors g_{ik} and g^{jl} can be used to derive new tensors from those given. Thus, if $A^{ij}{}_k$ is a given tensor field then $A^i{}_{jk}$ (see 192) is a new tensor field. The two fields are said to be associated to each other, and the process is described as that of raising or lowering indices. The length of a vector A^i is defined (see 193). The angle θ between two vectors A^i, B^j, of lengths different from zero, is defined by a formula (see 194).

From the fundamental (or metric) tensor g_{ik}, Elwin Bruno Christoffel, a German mathematician (see ANALYSIS, VECTOR AND TENSOR), constructed (in 1869) a quantity expressed in terms of partial derivatives (see 195). These are not the components of a tensor, but can be used to define the covariant derivative of a tensor field, giving a tensor field of one more covariant order. In the cases of contravariant and covariant vector fields and that of a mixed tensor field, their covariant derivatives are defined symbolically (see 196). Similar expressions are valid for the covariant derivatives of more general tensor fields. The covariant derivative of g_{ik} vanishes identically. Conversely, $\Gamma_i{}^j{}_k$ can be characterized as the set of quantities that are symmetric in i, k and for which the covariant derivative of the fundamental tensor is zero. In geometrical language, $\Gamma_i{}^j{}_k$ defines a parallel displacement of vectors along curves: the vector field A^k is parallel along a curve $x^i = x^i(t)$ if a condition holds (see 197). This is called the Levi-Civita parallelism (named after Tullio Levi-Civita, a 19th–20th-century Italian mathematician and physicist). It is an affine connection in the general terminology of connections, but is related in a special way to the Riemannian metric. Under the parallelism of Levi-Civita, the scalar product of two vectors remains unchanged. A geodesic is either a curve of zero length or one along which the unit tangent vector is parallel. The former possibility does not appear in the case of a positive definite metric.

Levi-Civita parallelism

The notion of curvature arises in studying the parallel displacement of a vector around a closed curve. When the closed curve is an infinitesimal parallelogram, the new position of the vector can be obtained from the initial position by an infinitesimal rotation. Analytically, this association is described by the Riemann-Christoffel curvature tensor (see 198). The components of the curvature tensor have symmetry properties (see 199). As a result, there are $n^2(n^2-1)/12$ independent components; *e.g.*, six for a three-space, 20 for a four-space, and one for a two-space. In the latter case the quotient $K = -R_{1212}/g$ is a scalar, called the Gaussian

(188) $$ds^2 = \sum_{1 \le i,k \le n} g_{ik}(x)\, dx^i dx^k$$

(189) $$s = \int_{t_0}^{t_1} \sqrt{\left|\sum_{i,k} g_{ik} \frac{dx^i}{dt}\frac{dx^k}{dt}\right|}\, dt$$

(190) $$g_{ik} = \sum_{j,l} g'_{jl} \frac{\partial x'^j}{\partial x^i}\frac{\partial x'^l}{\partial x^k} \quad \text{in which} \quad i,j,k,l = 1, \cdots, n$$

(191) $$\begin{cases} \sum_k g_{ik}\, g^{kl} = \delta_i^l = \begin{cases} 1, & i = l \\ 0, & i \neq l \end{cases} \\ \text{in which} \quad i,k,l = 1, \cdots, n \end{cases}$$

(192) $$A^i_{jk} = \sum_l g_{ij} A^{il}_k, \qquad A^{ijk} = \sum_l g^{kl} A^{ij}_l$$

(193) $$|A| = \sqrt{\left|\sum_{i,k} g_{ik} A^i A^k\right|}$$

(194) $$\cos\theta = \sum_{i,k} g_{ik} A^i B^k / (|A| \cdot |B|)$$

(195) $$\Gamma_{ijk} = \tfrac{1}{2}\left(\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j}\right), \qquad \Gamma^j_{ik} = \sum_l g^{jl}\Gamma_{ilk}$$

(196) $$\begin{cases} A^i_k = \dfrac{\partial A^i}{\partial x^k} + \sum_j \Gamma^i_{jk} A^j \\ A_{i,k} = \dfrac{\partial A_i}{\partial x^k} - \sum_j \Gamma^j_{ik} A_j \\ B^i_{j,k} = \dfrac{\partial B^i_j}{\partial x^k} + \sum_l (\Gamma^i_{lk} B^l_j - \Gamma^l_{jk} B^i_l) \end{cases}$$

(197) $$\begin{cases} \dfrac{dA^i}{dt} + \sum_{j,k} \Gamma^i_{jk} \dfrac{dx^j}{dt} A^k = 0 \\ \text{in which} \quad i,j,k = 1, \cdots, n \end{cases}$$

(198) $$\begin{cases} R^j_{ikl} = \dfrac{\partial \Gamma^j_{il}}{\partial x^k} - \dfrac{\partial \Gamma^j_{ik}}{\partial x^l} + \sum_m (\Gamma^m_{il}\Gamma^j_{mk} - \Gamma^m_{ik}\Gamma^j_{ml}) \\ R_{ijkl} = \sum_m g_{jm} R^m_{ikl} \end{cases}$$

(199) $$\begin{cases} R_{ijkl} = -R_{jikl} = -R_{ijlk} \\ R_{ijkl} = R_{klij} \\ R_{ijkl} + R_{iklj} + R_{iljk} = 0 \end{cases}$$

curvature. If a vector is given a parallel displacement around a closed curve on a surface, it returns with its direction changed by an amount equal to $\int K d\sigma$, the Gaussian curvature of the surface integrated over the surface area bounded by the closed curve. As a consequence, the excess (over π) of the angle sum in any geodesic triangle is $\int K d\sigma$, a theorem proposed by Gauss.

In the case of three or more dimensions, the curvature properties can no longer be expressed by a single magnitude, but require for their description the whole curvature tensor. The concept of Gaussian curvature is replaced by that of Riemannian curvature, or sectional curvature. This is, at any point p and any plane element (two-dimensional) through p, the Gaussian curvature of the geodesic surface through p and tangent to the plane element. If the latter is spanned by the vectors A^i, B^j, the Riemannian curvature is given by a formula (see 200). In general, K will depend both on the point and on the plane element through it. In other words, with regard to curvature, the space may be non-homogeneous as well as anisotropic. For $n \geq 3$, however, Schur's theorem (named after Friedrich Heinrich Schur, a 19th–20th-century Polish-German mathematician) says that if K is everywhere isotropic (*i.e.*, independent of the choice of the plane element at every point), then it is constant throughout. Riemannian spaces, for which the Riemannian curvature is constant for all points and all plane elements through them, are said to be of constant curvature. They include the Euclidean and non-Euclidean spaces. They are characterized by analytic conditions (see 201) with constant K. The curvature tensor gives, by contraction, the symmetric tensor (see 202) called the Ricci tensor (named after Curbastro Gregorio Ricci). A further contraction gives the scalar invariant (see 203).

Riemannian curvature

The space in general relativity, in particular (see RELATIVITY), is a four-dimensional Riemannian space with an indefinite metric that can be written locally as $ds^2 = \omega_1^2 + \omega_2^2 + \omega_3^2 - \omega_4^2$ ($\omega_1, \cdots, \omega_4$ are linear differential forms). The components g_{ik} of the fundamental tensor define the gravitational potential. Thus, light rays are null curves defined by the differential equation $ds^2 = 0$, whereas the trajectories of particles, without the action of external forces, are the geodesics. In this way, the geometry of the space is not a priori given, but is determined by matter. An essential restriction on the metric is given by Einstein's field equation $R_{ik} = -H(T_{ik} - \frac{1}{2}g_{ik}T)$, in which T_{ik} is the energy-momentum tensor, T is the contraction of T_{ik} relative to the fundamental tensor, and H is a constant. Many attempts have been made, some by Einstein himself in his late years, to develop a unified field theory that would include both the gravitational and the electromagnetic fields. This leads to a generalization of Riemannian geometry to more general geometric objects, such as the Weyl geometry, the Kaluza five-dimensional metric, projective relativity of Oswald Veblen. The results are not conclusive. (S.S.C.)

The unified field theory

Lorentz transformation and Lorentz invariance. It will now be shown explicitly that, though valid in Newtonian mechanics, the so-called Galilean transformation given by equations (182) and augmented by $t' = t$, is not compatible with relativity theory, but must be replaced by the Lorentz-transformation (named after the 19th–20th-century Dutch physicist Hendrik Antoon Lorentz; see 204). Here, the primed and unprimed coordinates and times (t', t) are associated with reference frames S' and S, respectively; c is the velocity of light in empty space; $\beta = v/c$, and $\gamma = (1 - \beta^2)^{-1/2}$, with v denoting the constant velocity of S' with respect to S. The two frames are supposed to be oriented so that the space axes of S' are parallel to the corresponding axes of S, and x- and x'-axes also being collinear with each other and with v.

The inverse transformation—*i.e.*, from S' to S—is readily obtained from equations (204) by interchanging primed and unprimed symbols and replacing v by $-v'$; *i.e.*, β by $-\beta$ (see 205). It will be observed that, as $\beta \to 0$ and $\gamma \to 1$—*i.e.*, as the relative velocity v becomes progressively smaller when compared to that of light, c—the Lorentz transformations (see 204 and 205) do reduce to the Galilean form (and $t' = t$; see 181).

$$K = -\frac{\sum_{i,j,k,l} R_{ijkl} A^i A^k B^j B^l}{\sum_{i,j,k,l} (g_{ik}g_{jl} - g_{il}g_{jk}) A^i A^k B^j B^l} \tag{200}$$

$$R_{ijkl} = -K(g_{ik}g_{jl} - g_{il}g_{jk}) \tag{201}$$

$$R_{ik} = \sum_j R^j_{ijk} \tag{202}$$

$$R = \sum_i R^i_i \tag{203}$$

$$\begin{cases} x' = \gamma(x - \beta ct), & y' = y, \quad z' = z \\ t' = \gamma(t - \beta x/c) \end{cases} \tag{204}$$

$$\begin{cases} x = \gamma(x' + \beta ct'), & y = y', \quad z = z' \\ t = \gamma(t' + \beta x'/c) \end{cases} \tag{205}$$

$$x^2 + y^2 + z^2 = c^2t^2 \tag{206}$$

$$(x' + vt')^2 + y'^2 + z'^2 = c^2t'^2 \tag{207}$$

$$\gamma^2(x' + \beta ct')^2 + y'^2 + z'^2 = c^2\gamma^2(t' + \beta x'/c)^2 \tag{208}$$

$$x'^2 + y'^2 + z'^2 = c^2t'^2 \tag{209}$$

$$s^2 \equiv x^2 + y^2 + z^2 - c^2t^2 = x'^2 + y'^2 + z'^2 - c^2t'^2 \tag{210}$$

As had been shown by Maxwell (see above *Electromagnetic theory*), light and all other electromagnetic waves propagate, in empty space, with velocity $c = 3 \times 10^{10}$ cm/sec relative to a given reference frame S. Hence, if a light signal is emitted from the origin O of S at time $t = 0$, then after a time t has elapsed, a spherical wave will have spread out to a distance ct (see 206). If the origin O' of system S' was in coincidence with O at time $t = 0$, then the Galilean transformation gives the equation (see 207) of a sphere centred at $x' = -vt' = -vt$. Along the $+x'$-direction of S', the wave will therefore have travelled only a distance $ct - vt = (c - v)t$, and a distance $(c + v)t$ along the negative x'-direction. Each of these deductions, however, is invalidated by an extensive series of experiments (the earliest of which was that by U.S. physicists A.A. Michelson and E.W. Morley, reported in 1887) showing that the velocity of light has the same constant value c in every reference frame.

This discrepancy and that between the analytic forms (206), (207) are what led Einstein (in 1905) to enunciate the so-called postulate of equivalence of special relativity. This states that all reference systems with constant velocities relative to one another are equivalent with regard to the mathematical form of the laws of nature; that is, the latter are to be invariant in form under transformation of coordinates from one reference frame to another. The transformations satisfying this requirement are linear in the four variables x, y, z, and t, as shown by equations (204), (205).

A corresponding postulate of general relativity (*i.e.*, relating to frames of reference in arbitrary motion relative to one another) was formulated by Einstein in 1916. In this case, invariance of the laws of nature requires that the equations of transformation be nonlinear. The following treatment will be limited to the special theory only.

Applying the transformations (205) instead of the Galilean transformation to (198) yields an equation (see 208) that, after straightforward expansion and cancellation, is found to reduce (see 209). Equation (209) represents a spherical light wave, centred at O', propagating with speed c, and has exactly the same

mathematical form in S' as the corresponding expression (205) has with respect to S. The quantity expressing the square of s in terms of the coordinates x, y, and z (see 210) is therefore Lorentz-invariant, and is thus seen to satisfy Einstein's equivalence principle and also the requirement that the velocity of light have the same value c in all uniformly moving coordinate systems. The Lorentz transformation is, therefore, the relativistically correct one. In general, quantities such as s^2 are said to be Lorentz-invariant. (Maxwell's equations of electromagnetism constitute another example.)

Denoting the variables x, y, z, and ict, respectively, by x_1, x_2, x_3, and x_4, with $i = \sqrt{-1}$, equation (210) can be rewritten: s squared equals a sum of squares of x_μ and a sum of squares of x'_μ (see 211), which, in analogy with the corresponding three-dimensional expression $(x^2 + y^2 + z^2)$, is interpreted as the invariant length squared of a so-called four-vector in four-space; *i.e.*, in the four-dimensional space-time continuum. Equations (204), (205), and (211) exhibit clearly the essentially inseparable interdependence of the coordinates of space and time.

Lorentz contraction and time dilation. These two are striking consequences of the Lorentz transformation. Thus, if $l_0 = x'_2 - x'_1$ be the length of a rod at rest in S', then by the first of equations (204), l_0 is a difference of x'_2 and x'_1 and is equal to an expression involving γ, the unprimed coordinates, and β (see 212); but the measurement of the rod's length in S must be made at the same time $t = t_1 = t_2$. It therefore follows from equation (212) that l is proportional to l_0, the proportionality constant involving β (see 213), in which $l = x_2 - x_1$. Equation (213) shows that the observer in S will judge the length of the rod to be shortened by the factor $\sqrt{1 - \beta^2}$, relative to the length l_0 determined by the observer at rest in S'. This reduction in length is known as the contraction of Lorentz and Fitzgerald (George Francis Fitzgerald, an Irish physicist).

Finally, if a given event observed in S' is found to take a time $t_0 = t'_2 - t'_1$, then it follows from the fourth of equations (204) that t_0 is a difference of primed coordinates and equals γ times an expression involving γ, β, and unprimed coordinates (see 214); but the time measurement in S is made with a clock at the fixed location $x = x_1 = x_2$, so that t is proportional to t_0, the proportionality constant being expressed in terms of β (see 215), in which $t = t_2 - t_1$ represents the time taken by the event as measured by the clock (or observer) at rest in S. Equation (215) thus shows that t, as measured in S, exceeds the corresponding determination in S' by the factor $\sqrt{1 - \beta^2}$. This phenomenon, abundantly verified by experiment, is called time-dilation. (The same factor affects also the determination of mass, if it is moving relative to the observer. For other aspects of relativity theory, see the article RELATIVITY and the *Bibliography*.)

QUANTUM MECHANICS

Quantum mechanics is the mathematical description of physical systems and phenomena based on the explicit premise that, like matter, energy too is atomistic or quantized; *i.e.*, extant in discrete amounts, or quanta, only.

Transition from classical to quantum mechanics. Although classical mechanics has proved to have an accuracy and a range of empirical validity far beyond the experimental evidence available at the time of Newton, the theory was found to be inapplicable to systems of atomic or subatomic dimensions, and also when the speeds involved are comparable to that of light. In some instances, both of these features can occur simultaneously, as in the case of electrons moving in the immediate vicinity of a heavy atomic nucleus.

Planck's quantum hypothesis. The notion that energy is quantized was introduced in 1900 by the German physicist Max Planck in a study of blackbody radiation. (A blackbody is one that absorbs all radiation incident upon it.) The next major advance came in 1905, when Albert Einstein showed that the energy of escape of an electron ejected photoelectrically from a metal could be explained by assuming that a light quantum of energy $h\nu$ is absorbed, in which ν is the frequency and h is a constant of proportionality, now called Planck's constant, equal to 6.625×10^{-27} erg-sec (see 216).

Self-consistent formulations of quantum mechanics. A great leap for theoretical physics occurred when, in 1913, the Danish physicist Niels Bohr devised a mathematical model of the atom depending crucially on Planck's quantum hypothesis. Yet, although this theory introduced certain novel concepts that have proved of lasting value to modern physics, and although the theory successfully accounted for a number of atomic phenomena not previously understood, it proved inadequate in other respects and in some instances blatantly disagreed with experiment. It thus became clear that despite its modern (*i.e.*, nonclassical) concepts, the difficulties of Bohr's quantum theory stemmed from its remaining still too closely connected to classical mechanics, with the quantum conditions grafted on, rather than forming an intrinsic part of, the basic theoretical framework. What was needed was a rational, self-consistent quantum mechanics that started afresh and evolved a mathematical framework and philosophy all its own. This was discovered in 1926, an epoch-making year in the history of physics.

The inadequacy of Bohr's atomic model

This remarkable new mechanics was the result of independent efforts by some of the leading theoretical physicists of the time. It was developed in a diversity of mathematical formulations, which, though different in superficial mathematical structure, are basically identical in physical content.

The three main mathematical forms are: (1) the wave mechanics of Erwin Schrödinger, an Austrian physicist, the outgrowth of de Broglie's concept (formulated by Louis de Broglie, a French physicist) of phase waves; (2) the matrix theory of the German physicists Max Born, Werner Heisenberg, and Pascual Jordan; and (3) the so-called transformation theory, based on kinematical indeterminism, developed by P.A.M. Dirac, a British physicist, and Jordan, and interpreted by Heisenberg.

Wave mechanics. *Hamiltonian analogy and de Broglie waves.* The version most universally used, and perhaps easiest to understand, is the so-called wave formulation. The impetus to its development came from de Broglie's interpretation of the classical Hamiltonian analogy between optics and dynamics. Readily demonstrated from Fermat's principle in optics and the principle of least action in mechanics, this analogy establishes a formal similarity between the path of the particle in a mechanical system and a hypothetical optical system in which the refractive index is correlated with the potential energy $V(x,y,z)$ in the mechanical problem. This comparison always seemed highly artificial, because the velocity in the optical system was different from that in the mechanical one. If, however, it is somehow possible to introduce the concept of wavelength into the analogy, so that it is meaningful to speak of dispersion, then there exists besides the phase velocity, also a group velocity; *i.e.*, the velocity with which signals or modulations are propagated.

De Broglie showed that, although different from the phase velocity, the velocity of the mechanical particle becomes identical with the group velocity, provided the wavelength λ is arbitrarily assumed to have the same value as the ratio of h to the product of v and m (see 217). The second form of (217) is obtained by noting that the energy constant E is the sum of the kinetic energy $mv^2/2$ and the potential energy $V(x, y, z)$. Equation (217) gives the celebrated de Broglie wavelength, characteristic of quantum-mechanical diffraction problems. When de Broglie first propounded these ideas in 1924, it was not clear whether they had actual physical significance. Two years later, however, the existence of the de Broglie wavelength was confirmed in a most striking fashion through the experiments of the U.S. physicists Clinton J. Davisson and Lester H. Germer, who showed that beams of electrons passing through a crystal are diffracted in many ways like X-rays, but with the wavelength given by de Broglie's formula $\lambda = h/mv$.

Existence of de Broglie waves confirmed

Schrödinger wave equation. Already before the Davisson-Germer experiments, however, Schrödinger recognized the great portent of the de Broglie wavelength, and by means of it was led to the fundamental equation of wave mechanics, usually called the wave equation. Schrödinger conjectured that true quantum mechanics bears the same relation to classical mechanics as physical optics does to geometrical optics: Just as the ray-tracing method of the latter fails to explain the phenomena of diffraction and interference of light waves, so, Schrödinger reasoned, ordinary mechanics could not explain atomic phenomena, the reason in both cases being that the dimensions of apertures through which light passes in diffraction experiments, or atoms, are not large compared to the wavelength. Hence, he sought to establish a procedure analogous to that used in physical optics, in which the fundamental wave equation is a second-order partial differential equation for f (see 218). Here the phase velocity u may be a function of x, y, z. In the monochromatic, or periodic, case, the solutions can be written in the form $f = \psi(x, y, z) \exp(2\pi i\nu t)$ so that the space factor ψ satisfies a second-order partial differential equation (see 219) with $\lambda = u/\nu$. For the wavelength λ, Schrödinger inserted the de Broglie value (11). Equation (219) then becomes a second-order partial differential equation for ψ in which the difference E minus V appears (see 220). This is the Schrödinger wave equation. It may be questioned how quantized phenomena can be extracted from (220), because a partial differential equation generally suggests solutions that are continuous rather than discrete. The answer is that only those solutions are physically allowed that are single-valued and continuous, together with their first derivatives; that vanish at infinity; and that are finite, except possibly at a finite number of points at which ψ may become infinite; but this in only such a way that the integrals of the absolute square of ψ and the sum of absolute squares of the partial derivatives of ψ (see 221) exist. Such physically admissible solutions are called characteristic functions, proper functions, or eigenfunctions. Now, to be sure, (220) possesses an infinity of solutions, because the latter involve arbitrary constants of integration and even arbitrary functions. Nevertheless, it often happens that not one of these solutions has the properties demanded of a characteristic function. Instead, an admissible solution is usually obtained only if the energy E is given certain particular values, called characteristic values, proper values, or eigenvalues.

The Schrödinger wave equation

A simple illustration of how the condition that the solution be physically admissible limits the parameter E is provided by the two-dimensional rotator; *i.e.*, rotating dumbbell (diatomic molecule) the motion of which is confined to a plane. Here the wave equation (220) becomes a second-order differential equation including I, E, and h (see 222), as is seen by specializing (220) to one coordinate, with $V = 0$, and noting that in a rotational problem the moment of inertia I and angle ϕ replace the mass and a linear displacement, respectively. The solution of (222) is an expression for ψ in terms of A and the cosine of a function involving I, E, h, ϕ, and ε (see 223), in which A and ε are arbitrary constants. The requirement of single-valuedness as a function of position on the circular path of rotation demands that $\psi(\phi + 2\pi) = \psi(\phi)$. This is possible only if the coefficient of ϕ, in the argument of the cosine in (223), is an integer n; that is, if E equals a certain ratio involving n, h, and I (see 224). This expression for the energy of a rotator is exactly the same as that obtainable from Bohr's old quantum theory. At first sight it would thus appear that the new quantum mechanics fails to remove the difficulty of Bohr's theory requiring half-integers for obtaining agreement with experiment on rotational energy levels. Actually, however, the end-over-end rotation of a molecule should be represented by a rotator confined not to a particular plane, but rather by one moving in three-dimensional space. When this extra degree of freedom is taken into account, it is found that the factor n^2 in (224) is replaced by $(n + \frac{1}{2})^2 - \frac{1}{4}$, so that, apart from an unimportant additive constant, the result is the same as though half quantum numbers were used in the Bohr theory.

(211) $$s^2 = \sum_{\mu=1}^{4} x_\mu^2 = \sum_{\nu=1}^{4} x_\nu'^2$$

(212) $$l_0 = x_2' - x_1' = \gamma(x_2 - \beta c t_2 - x_1 + \beta c t_1)$$

(213) $$l = l_0\sqrt{1-\beta^2}$$

(214) $$t_0 = t_2' - t_1' = \gamma\left(t_2 - t_1 - \frac{\beta x_2}{c} + \frac{\beta x_1}{c}\right)$$

(215) $$t = t_0/\sqrt{1-\beta^2}$$

(216) $$E = h\nu$$

(217) $$\lambda = \frac{h}{mv} = \frac{h}{\sqrt{2m(E-V)}}$$

(218) $$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \frac{1}{u^2}\frac{\partial^2 f}{\partial t^2}$$

(219) $$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} = -\frac{4\pi^2}{\lambda^2}\psi$$

(220) $$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} + \frac{8\pi^2 m}{h^2}(E-V)\psi = 0$$

(221) $$\iiint |\psi|^2 \, dv \quad \text{and} \quad \iiint \left\{\left|\frac{\partial \psi}{\partial x}\right|^2 + \left|\frac{\partial \psi}{\partial y}\right|^2 + \left|\frac{\partial \psi}{\partial z}\right|^2\right\} dv$$

(222) $$\frac{\partial^2 \psi}{\partial \varphi^2} + \frac{8\pi^2 IE}{h^2}\psi = 0$$

(223) $$\psi = A\cos[(8\pi^2 IE/h^2)^{1/2}\varphi - \varepsilon]$$

(224) $$E = n^2h^2/8\pi^2 I$$

(225) $$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2}(E - \tfrac{1}{2}ax^2)\psi = 0$$

(226) $$\begin{cases} A = \dfrac{4\pi E}{h}\left(\dfrac{m}{a}\right)^{1/2}, \qquad u = \dfrac{m^{1/4}a^{1/4}(2\pi)^{1/2}}{h^{1/2}}x \\ \psi = e^{-u^2/2}g \end{cases}$$

(227) $$\frac{d^2 g}{du^2} - 2u\frac{dg}{du} + (A-1)g = 0$$

(228) $$g = c_k u^k + c_{k+2}u^{k+2} + c_{k+4}u^{k+4} + \cdots$$

In most problems, the mathematical analysis is, of course, more complicated than in this simple illustration of the rotator. When exact closed expressions are obtained for the characteristic values of a system with one degree of freedom, they are usually derivable by the so-called polynomial method. In this, the solution is expressed as a power series and is shown to be physically admissible if the series terminates. The termination condition then usually restricts the energy parameter E to certain particular eigenvalues.

Use of the polynomial method

The polynomial method may be illustrated by the one-dimensional harmonic oscillator. Here the wave equation (220) takes the form of a second-order ordinary

differential equation for ψ in which the difference between E and one-half ax^2 appears (see 225). On the introduction of the appropriate abbreviations and changes of variables (see 226), equation (217) becomes an ordinary second-order differential equation for g in which A appears (see 227). A solution is obtainable by assuming that g is a power series in u with constant coefficients (see 228). After the indicated operations are performed, the entire left side of (227) can be expressed in the form of a series. The coefficient of each power of u in the latter must vanish if (227) is to be satisfied. The vanishing of the coefficient of the lowest power (*i.e.*, u^{k-2}) requires either $k = 0$ or $k = 1$. On equating to zero the coefficient of u^n, one obtains a recurrence formula (see 229) for determining the ratio c_{n+2}/c_n. The series (228) will terminate at the term $c_n u^n$ if $A = 2n + 1$ or, in other words, by (226), if the expressions (230) hold for E and ν_0. It can be shown that if the series does not terminate, then ψ increases without limit as x becomes arbitrarily large, and the solution is not physically admissible. Hence, the only characteristic values of the harmonic oscillator are those given by (230).

It is of interest to note that the new quantum mechanics gives the same energy levels for the harmonic oscillator as does the old Bohr theory, except for an additive constant (equivalent to the use of half-integers in the old quantum theory). Usually it is unimportant whether the extra constant is included, but by studying the shift between the electronic spectra of diatomic molecules differing only in isotopic composition, it is possible to decide experimentally between Bohr's result and (230). The results unequivocally favour the latter.

More general form of the wave equation. Equation (220) is the form of the wave equation for a conservative system consisting of a single particle and having a definite energy. A rule will now be given for forming the wave equation in other cases, involving either more than one particle or an indefinite energy, or both. (In classical theory a conservative system necessarily has a definite, constant energy, but in quantum mechanics this need not be so. As a result, it is no longer redundant to say that a conservative system has a definite energy.) The more general scheme for forming the wave equation given here also applies to the nonconservative case in which the potential energy V contains the time explicitly.

The procedure is as follows. First of all, the Hamiltonian technique is to be used; *i.e.*, the system is to have associated with it a classical Hamiltonian function (see 231), which is a function of the coordinates and momenta. One replaces p_k by an operator involving the partial derivative with respect to q_k (see 232) and in this fashion forms a linear differential equation (see 233). This is the general wave equation. One is led to it primarily by comparison with other forms of quantum mechanics, especially transformation theory, rather than through the de Broglie and Schrödinger interpretation of the Hamiltonian analogy.

If the system is conservative, so that t does not enter explicitly, (233) has a solution (234) of the form of a product of an exponential involving t times a function of ψ_m with arguments $q_1, \cdots, q_f$, in which ψ_m is independent of t, and satisfies an equation in which an operator operating upon ψ_m is zero (see 235).

An important property of (235) is that its solutions can be shown to be orthogonal, provided the operator is of the so-called self-adjoint type. This condition is one that can practically always be met. If, in addition, the arbitrary proportionality constants in the wave function are so chosen that they are normalized to unity, then the integral of the product of $\psi_l^* \psi_m$ is a δ function (see 236) provided the integration is extended over all the coordinates. Thus $d\tau$ in (236) denotes $dx_1 \cdots dz_n$ if there are n particles. The expression δ_m^l is the Kronecker delta and has the meaning that $\delta_m^m = 1$ and that $\delta_m^l = 0$ $(l \neq m)$. The asterisk is used here and elsewhere to denote the complex conjugate. It is necessary to distinguish between ψ_m and its conjugate ψ_m^* because the solutions of the Schrödinger wave equation are not necessarily real. In (236), the case $l \neq m$ expresses the orthogonality, and $l = m$ embodies the normalization.

Matrix form of quantum mechanics. *Matrices and their physical significance.* A few months before Schrödinger's discovery of wave mechanics, another, superficially much different, approach to quantum mechanics was developed by Born and Heisenberg. The impetus to this approach was Born's repeated emphasis, to his colleagues at Göttingen, that the reason the old quantum theory was then (1925) failing was that it sought to use the same kinematical concepts of space and time within the atom as in ordinary measurable large-scale events. After all, the concepts of space and time have a meaning only when a realizable procedure for their measurement is specified. Ordinary measuring rods and clocks are clearly not usable at atomic dimensions. Guided by this philosophy based on the so-called operational viewpoint, Heisenberg discovered his matrix formulation of quantum mechanics. He did not originally recognize it as such; it was Born who saw first that the equations were best interpreted in the language of matrices. The basic idea of the matrix mechanics is that the coordinates and momenta associated with the electrons in the atom are to be treated as matrices rather than as quantities having a definite numerical value. A matrix is an array (see 237). Most matrices involved in quantum mechanics are infinite; *i.e.*, have an infinite number of rows and columns.

The meaning of matrix mechanics

The Heisenberg matrix elements are intimately connected with the concept of the transition probability (probability that an atom or nucleus will change from one energy state to another) in the old quantum theory and, alternatively, with Fourier coefficients in classical mechanics. In a simple system having only one degree of freedom, the matrix index can be identified with the quantum number, and a matrix element $a(lm)$ is correlated with a transition between states of quantum numbers l and m. The expression for the transition probability is given in (238). Here $x(lm)$ is an element of the x coordinate matrix and is the quantum analogue of the Fourier coefficient x_{l-m} involved in the corresponding rough formula arrived at by means of Bohr's correspondence principle. This important part of Bohr's theory states that, in the limit of large quantum numbers, the quantum-mechanical results must reduce to those of classical mechanics. As presented here, and as Heisenberg reasoned historically, (238) is obtained by analogy with classical theory, but it can also be derived independently from the radiation theory of Dirac, which applies quantum mechanics to the degrees of freedom of the ether. By means of (238) the intensity of a spectral line can be calculated if the Heisenberg matrices are known. To compute the matrix elements needed for use in (238), it is necessary to transcribe atomic mechanics into matrix language; for this a matrix algebra and calculus are necessary.

The sum of two matrices is formed by adding corresponding elements, so that $(a + b)(lm) = a(lm) + b(lm)$. Matrices are multiplied in the same way as determinants (see 239). There is, however, the difference that, whereas a determinant has a numerical value, a numerical meaning is attached only to the individual entries of a matrix and not to its entirety. This loss of numerical value corresponds to the abandonment of ordinary kinematical relationships within the atom. The multiplication is in general noncommutative; *i.e.*, $(ab)(lm) \neq (ba)(lm)$, or more symbolically: $\boldsymbol{ab} \neq \boldsymbol{ba}$, if boldface type is used to designate the entire matrix rather than a single entry. By repeated additions and multiplications it is possible to construct functions of matrices that are in turn matrices, and thus develop a matrix algebra.

It is further necessary to have a rule for constructing the derivative of one matrix with respect to another. A consistent mathematical framework is obtained if the derivatives of a function $f(q_1, \cdots, q_f, p_1, \cdots, p_f)$ are defined as in (240). The special cases $f = p_i$ or q_i yield the so-called commutation relations (see 241).

A diagonal matrix is one that has zero entries except along the principal diagonal—*i.e.*, $a(lm) = 0$, $l \neq m$—whereas a unit matrix has the additional property $1(ll) = 1$. A diagonal matrix is the quantum analogue of a Fourier series that contains but one single constant term. Hence, in a conservative system with a definite energy, the energy matrix must be a diagonal one.

Diagonal matrices

It will be recalled that the Heisenberg matrices are infinite arrays; *i.e.*, they involve an infinite number of rows or columns. Physically, this means that each index (number of row or column) is correlated with a stationary state, and that there is an infinite number of such states. Furthermore, the permitted stationary states often form a continuous rather than a discrete manifold. In the continuous region it is necessary to introduce the (rather recondite) concept of a matrix with continuously varying indices, so that it cannot then even be tabulated in terms of rows and columns. When there is a continuum, the summation in (239) must be replaced by an integration.

Relation of matrix and wave theories. It is possible to work out the energy levels of the harmonic oscillator, rotator, hydrogen atom, etc., by means of pure matrix theory, and the results are always the same as those obtained with wave mechanics. It turns out that the two approaches are interrelated. If the Schrödinger wave functions are known, the Heisenberg matrix elements can be computed by a straightforward integration. Thus the matrix elements of a function $f(q_1, \cdots, q_f, p_1, \cdots, p_f)$ of the coordinates and momenta may be desired. Then the appropriate formula f as a function of the product lm is an integral (see 242), in which the integrand contains an operator f formed by replacing an operator composed of the partial derivative with respect to q_k (see 243). Equation (242) will not be proved, but one or two simple consequences will be noted. In the first place, the commutation relations are satisfied. For instance, if $f = p_k q_k - q_k p_k$, then f times ψ_m is by calculation proportional to ψ_m (see 244) and, except for a factor $h/2\pi i$, the integral (242) becomes identical with (236) so that the matrix elements of this function are proportional to a δ function (see 245). This result agrees with the first relation of (241) because the elements of a unit matrix are equivalent to a Kronecker delta. The remaining relations of (241) are obviously satisfied. The third relation, for instance, follows because the second partial of ψ can be taken with the two variables in either order (see 246).

Satisfaction of the wave equation (233) for a conservative system with definite energy is tantamount to making the energy a diagonal matrix. To see this it may be noted that by (233) the operator H is equivalent to multiplication by a constant E_m. Hence, if f in (244) is taken to be the Hamiltonian operator H, the integral differs from (236) merely by a constant factor and so has the value $E_m \delta_m^l$.

If one is wedded to the extreme matrix viewpoint, one regards the wave functions merely as mathematical auxiliaries used to compute the Heisenberg matrix elements by means of the matrix computing machine. Such a viewpoint is undoubtedly too one-sided and is offensive to physicists who take comfort from the parallelism of wave mechanics to classical mathematical physics, and to optics, in particular.

Transformation theory. The general interpretation of quantum mechanics as essentially a theory of probability amplitudes has been developed particularly by Jordan and by Dirac. The details of this so-called transformation theory are much too intricate and abstract to be treated here, but this formulation of quantum mechanics has elegance, generality, and a unifying influence on the interpretation of the various other viewpoints. The Schrödinger wave equation, from the standpoint of transformation theory, may be regarded as an equation for finding the probability amplitude for distribution-in-position when the energy is known. One might ask for the probability amplitude for the distribution of functions other than positional coordinates—for instance, momenta—and the known quantity might not necessarily be the energy, as certain experiments correspond to fixation of quantities other than the energy. Such more general problems can be solved by means of transformation theory.

(229) $$\frac{c_{n+2}}{c_n} = \frac{2n+1-A}{(n+1)(n+2)}$$

(230) $$E = (n+\tfrac{1}{2})h\nu_0 \quad \text{with} \quad \nu_0 = \frac{1}{2\pi}\left(\frac{a}{m}\right)^{1/2}$$

(231) $$H(q_1, \cdots, q_f, p_1, \cdots, p_f, t)$$

(232) $$\frac{h}{2\pi i}\frac{\partial}{\partial q_k}$$

(233) $$\left[H\left(q_1, \cdots, q_f, \frac{h}{2\pi i}\frac{\partial}{\partial q_1}, \cdots, \frac{h}{2\pi i}\frac{\partial}{\partial q_f}, t\right) + \frac{h}{2\pi i}\frac{\partial}{\partial t}\right]\Psi = 0$$

(234) $$\Psi = e^{-2\pi i E_m t/h}\psi_m(q_1, \cdots, q_f)$$

(235) $$\left[H\left(q_1, \cdots, q_f, \frac{h}{2\pi i}\frac{\partial}{\partial q_1}, \cdots, \frac{h}{2\pi i}\frac{\partial}{\partial q_f}\right) - E_m\right]\psi_m = 0$$

(236) $$\int \psi_l^* \psi_m d\tau = \delta_m^l$$

(237) $$a = \begin{Bmatrix} a(11) & a(12) & a(13) & \cdots \\ a(21) & a(22) & a(23) & \cdots \\ a(31) & a(32) & a(33) & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & a(nn) \end{Bmatrix}$$

(238) $$A_{l\to m} = (4e^2/3hc^3)(2\pi)^4\nu(lm)^3[|x(lm)|^2 + |y(lm)|^2 + |z(lm)|^2]$$

(239) $$(ab)(lm) = \Sigma_n a(ln)b(nm)$$

(240) $$\frac{\partial f}{\partial p_k} = \frac{2\pi i}{h}(fq_k - q_k f), \qquad \frac{\partial f}{\partial q_k} = -\frac{2\pi i}{h}(fp_k - p_k f)$$

(241) $$\begin{cases} p_k q_k - q_k p_k = (h/2\pi i)1, & p_k q_j - q_j p_k = 0, \quad (k \neq j) \\ p_k p_j - p_j p_k = 0, & q_k q_j - q_j q_k = 0 \end{cases}$$

(242) $$f(lm) = \int \psi_m^* f\left(q_1, \cdots, q_f, \frac{h}{2\pi i}\frac{\partial}{\partial q_1}, \cdots, \frac{h}{2\pi i}\frac{\partial}{\partial q_f}\right) \times \psi_l d\tau$$

(243) $$\frac{h}{2\pi i}\frac{\partial}{\partial q_k}$$

(244) $$f\psi_m = \frac{h}{2\pi i}\left[\frac{\partial}{\partial q_k}(q_k\psi_m) - q_k\frac{\partial \psi_m}{\partial q_k}\right] = \frac{h}{2\pi i}\psi_m$$

(245) $$f(lm) = (h/2\pi i)\delta_m^l$$

(246) $$\partial^2\psi/\partial q_j \partial q_k = \partial^2\psi/\partial q_k \partial q_j$$

Statistical significance of the wave function. In the following, attention will be focussed on the physical meaning of the solution of the Schrödinger wave equation, as distinct from its merely formal aspects. The wave function satisfying equation (233) has an important physical interpretation: The square of its modulus, $|\Psi|^2$, which is Ψ times its complex conjugate, is proportional to what may be called a statistical charge density. In the simple case of a system having only one

The statistical charge density

particle, the probability that the particle will be in the volume element $dxdydz$ at time t may be taken to be a differential with factor the absolute square of Ψ (see 247), provided the total probability is normalized to unity (see 248). This manner of interpretation is consistent, for it can be shown to be a consequence of the differential equation (233) that the integral in (248) is independent of the time. Otherwise, the normalization condition (248) could not be imposed. Similarly, when there are n particles, $|\Psi|^2$ determines the probability in a $3n$-dimensional space, the volume element $d\tau$ of which is $dx_1, \cdots, dz_n$.

It is to be emphasized that the density given by (247) is inherently statistical in character. The particle is not said to be in the element $dxdydz$ at time t, but merely that there is a certain probability of the particle's being there. The places where $|\Psi|$ or $|\psi_m|$ is large are those in which the particle is likely to be located, but it is never possible to say when it is there. The situation may be compared to the information obtainable from the time exposure of a firefly on a dark night. Wherever the photographic plate reveals the greatest brightness, the firefly must have spent most of the time, but one cannot infer just when the insect was at a particular spot. Similarly in quantum mechanics, the information about particles is inherently restricted because of its statistical nature.

That quantum mechanics must be interpreted statistically should be regarded as an advantage rather than as a drawback. For an analysis of atomic experiments shows that the measurements always involve statistical features.

Probability distribution in momentum space. Again confining attention to one particle only, one may wish to know the probability that the components of its momentum fall within a certain range at time t. The symmetrical roles of coordinates and momenta in the Hamiltonian technique suggest that there is a function χ such that a differential with factor absolute square of χ (see 249) determines the probability that the momenta are in the element dp_x, dp_y, dp_z at time t. There is indeed such a function, and the relation between the two functions Ψ, χ is similar to that between a function and its Fourier transform. Namely, it can be shown that Ψ is an integral of the product of χ and an exponential (see 250), and χ is an integral of the product Ψ and an exponential (see 251). Equation (250) has an intuitively very plausible form, for it is natural to regard any disturbance as being built up of a superposition of de Broglie waves, each of the structure $\exp(2\pi ipx/h)$ in one dimension, or $\exp[2\pi i(p_x x + p_y y + p_z z)/h]$ in three. In any principle of superposition in physics, the square of the amplitude governs the intensity, and so it is natural to assume that $|\chi|^2$ gives the probability. It can be shown from the mathematical theory of Fourier transforms (Plancherel's theorem) that two integrals are equal in which integrands are absolute squares of ψ and χ (see 252). If (252) were not satisfied, the physical interpretation given Ψ, χ would lack consistency, for it is necessary that the total probability for either coordinate or momentum distribution be unity; in other words, the normalization in coordinate space must also imply a corresponding normalization in momentum space. Equations (253) and (254), respectively, can be shown to give the mean value of the nth power of the x-component of momentum and of displacement when n is an integer.

It is to be noticed that in (247) or (249), expressions have been given for the probability of the particle being in an element of volume $dxdydz$ of position space, or in an element of momentum space, but that there is no probability function for a concurrent occurrence of given values of position and momentum associated with a volume element $dxdydzdp_x dp_y dp_z$. This fact is formally expressed in the Heisenberg uncertainty principle, which limits the accuracy with which it is possible to specify simultaneously the values of a coordinate and its canonically conjugate momentum. If Δp_x and Δx be the limits of precision within which p_x and x can simultaneously be determined, there exists the inequality in which the product of Δx and Δp_x is greater than or equal to the ratio of h to 4π (see 255). Thus it is meaningless to talk about the probability of a simultaneous occurrence of a given value of a coordinate and its conjugate momentum, for by (255) there is infinite uncertainty in a momentum if the position is known with complete precision, or vice versa.

The uncertainty principle should not be regarded as an additional hypothesis, artificially injected into quantum mechanics. Instead, it can be proved to be a consequence of the latter.

Application of quantum mechanics to atomic phenomena is further treated in the article MECHANICS, QUANTUM. The preceding section has covered the mathematical formalism only of nonrelativistic quantum mechanics. Dirac has shown that inclusion of relativity necessitates replacing the second-order Schrödinger equation by four simultaneous first-order equations. The solutions of these equations account for the properties of electron spin and for the existence of the positron. When spin is considered, employing the so-called Pauli exclusion principle, the structure of Mendeleyev's periodic table emerges (see PERIODIC LAW and CHEMICAL ELEMENTS). (Ed.)

Relativistic quantum fields. In this section, the structure of relativistic quantum field theories will be described. The most successful of these quantum electrodynamics is the theory describing the electromagnetic interactions of electrons, positrons, and photons, consistent with quantum theory and with relativity.

The task of quantum electrodynamics may be divided into three parts: first, to develop a formalism for describing the particles in the absence of interactions; second, to discover the nature of the interactions between the particles; and third, given this interaction, to develop a scheme for calculating any quantity of interest.

The problem of describing free—*i.e.*, non-interacting—particles has a fairly straightforward solution. In nonrelativistic quantum mechanics, a free particle is described by a known wave function $\psi(x, t)$, which is already a field. It is straightforward to generalize it to the relativistic case. Just as certain quantities (like energy) are quantized in the passage from classical to quantum mechanics, so here the field is quantized (this process is called second quantization).

Second quantization

In the course of doing this, an infinity of variables is encountered. This arises because of particle production, which is an intrinsic feature of relativistic interactions, and one that distinguishes them sharply from the non-relativistic interactions. In the relativistic case, it is possible to begin with two particles and produce many others; for example, a negative electron and a positron (positive electron) can collide to produce an arbitrarily large number of electrons, positrons, and photons. Any formalism designed to account for such reactions must possess enough variables to describe all the particles produced, and thus needs an infinite number of variables. The quantized field formalism provides this flexibility.

Having developed the free-field formalism, the next problem is to introduce the interaction. The choice is guided by analogy with classical physics. There, the interaction energy is given in terms of the charge e of an electron by $ej_\mu(x)a^\mu(x)$, in which $ej_\mu(x)$ is the electric current produced by electrons and positrons, and $a^\mu(x)$ is the electromagnetic potential. In the free-field theory, a quantized field composed of electron and positron fields, which behaves like a current of free particles, is defined by $J_\mu(x)$. The analogue of $a^\mu(x)$ turns out to be the field associated with photons, denoted by $A^\mu(x)$. One then writes the energy operator H (called the Hamiltonian) as a sum of two terms, one labelled H_{free}, one labelled $H_{interaction}$ (see 256), in which H_{free} is the energy operator of freely moving particles. It is to be noted that e determines the strength of the interaction.

The final and most difficult problem is to develop a calculational scheme for exploring the consequences of the assumed energy operator. The solution of this problem was given by physicists Shin'ichirō Tomonaga of

Japan and Julian Schwinger and Richard Feynman of the United States; its extraordinary success depends on the exceedingly small value of the so-called fine-structure constant, namely, the dimensionless parameter $\alpha = 2\pi e^2/hc$ ($\alpha \cong 1/137$); so that H_{int} (the interaction Hamiltonian) is small compared to H_{free} (the free Hamiltonian). Each physical quantity M is written as a power series (see 257), and a scheme is given for calculating the coefficients M_n. As the integer n increases, the calculation of M_n becomes increasingly difficult. But because the fine-structure constant α is a small number and the M_n are not expected to be large, the error made in neglecting terms with large n is quite small. It is thus sufficient for practical purposes to calculate only the first few terms of (257).

Feynman diagrams and rules

For this purpose, Feynman associated with each M_n a set of diagrams (Feynman diagrams) and for each diagram he gave a prescription (Feynman rules) for evaluating the contribution of that diagram to M_n. These diagrams can be considered as mnemonic devices for keeping track of the various terms contributing to M_n. They also, however, allow a physical interpretation. One can think of electrons as continually absorbing and emitting photons. Two electrons interact when one of them emits a photon, which travels to the location of the second, where it is absorbed. (It is thus seen that local interactions have replaced action at a distance.) It can happen that a photon is emitted and absorbed before it can be detected. Such a photon is termed a virtual photon, as opposed to a physical photon that can, in principle, be detected. Each Feynman diagram is a pictorial representation of the manner in which virtual particles can be emitted and absorbed.

Unfortunately, there are several serious difficulties with this formalism. The operator H_{free} does not really represent freely moving physical electrons, positrons, and photons. A physical electron, even in isolation, has a charge that sets up an electric field with which the electron itself interacts. Pictorially, each electron emits and reabsorbs its own photons. This self-interaction is ignored in H_{free}, and arises from H_{int}. Thus, H_{free} represents not a physical free electron but a hypothetical bare electron, in which all self-interactions are ignored. The mass and charge of the bare electron differ from those of the physical electron. In place of (257) H can be written as the sum of two terms (see 258), in which H_{free}^{phys} represents the free Hamiltonian of the physical electrons. H_{free}^{phys} and H_{free} differ only in that the electron mass is different. The change of mass is called mass renormalization. Similarly, the bare charge becomes renormalized to the physical charge. It should be noted that the bare mass and charge are not directly measurable, because they refer to hypothetical bare electrons.

Mass renormalization

Feynman's arguments can then be repeated, this time involving the physical mass and physical charge. In either the renormalized or unrenormalized form, however, M_n seems to become infinite for sufficiently large n, so that Feynman's prescription cannot be correct. The clue toward overcoming this severe difficulty lies in noting that the bare masses and charges are physically unobservable. It turns out that one can choose the bare mass and charge to be infinite in such a way that they exactly cancel the other infinities in the coefficients M_n. Then M_n can be expressed in terms of the physical quantities only; *i.e.*, not involving the bare mass and charge, and M_n is now finite. (It is remarkable that the infinities in any term M_n can be avoided by a unique choice of only two parameters, the bare mass and charge of the electron.) Of course, it is these finite values of M_n that must be compared with experiment, and which have been so successful. For example, the prediction of the magnetic moment of the electron has been verified to an accuracy of nine significant digits.

It is possible to extend the entire development outlined above to theories other than electrodynamics. Given an interaction Hamiltonian in terms of some fields, one can discover the Feynman diagrams and the Feynman rules appropriate to this interaction. The calculations and difficulties in these theories are analogous to those in electrodynamics. In particular, attempts have been made to study strong interactions—those responsible for nuclear forces, for example—by means of these techniques. The difficulty is that the expansion parameter, the analogue of α, is now large (approximately 15), so that subsequent terms in the expansion (257) do not fall off rapidly, and hence the first few terms cannot adequately represent the quantity M. Besides this, the appearance of infinities, even in the formalism of electrodynamics, raises the question of whether such theories are mathematically well defined.

In response to these problems, an alternate approach to quantum field theory has been developed. In this approach, several general principles that are believed valid for any field theory are postulated. These principles include (1) relativistic invariance; (2) the postulate that all physical states have positive energy; (3) existence of a unique vacuum state; (4) causality—*i.e.*, the statement that signals do not propagate faster than the speed of light; (5) a connection between the fields and the particles in the theory (the so-called asymptotic condition); (6) assumptions about the mathematical nature of the field operators.

These general principles form a framework that is flexible enough to accommodate many different quantum field theories. Consequently, these principles do not lead to detailed numerical predictions. There are, nevertheless, some properties that emerge from these principles alone and that should, therefore, be true for any field theory. For example, it follows that for every particle of a given mass and charge, there exists another one with the same mass and opposite charge (antimatter).

Axiomatic field theory

This general approach, often called axiomatic field theory, is developed with a much higher degree of

(247) $$|\Psi|^2\,dxdydz$$

(248) $$\iiint |\Psi|^2\,dxdydz = 1$$

(249) $$|\chi(p_x, p_y, p_z, t)|^2\,dp_x dp_y dp_z$$

(250) $$\begin{cases} \Psi = h^{-3/2} \iiint \chi(p_x, p_y, p_z, t) e^{2\pi i(p_x x + p_y y + p_z z)/h}\,dv \\ dv = dp_x dp_y dp_z \end{cases}$$

(251) $$\begin{cases} \chi = h^{-3/2} \iiint \Psi(x, y, z, t)\, e^{-2\pi i(p_x x + p_y y + p_z z)/h}\,dv \\ dv = dxdydz \end{cases}$$

(252) $$\iiint |\Psi|^2\,dxdydz = \iiint |\chi|^2\,dp_x dp_y dp_z$$

(253) $$\begin{cases} \iiint p_x^n |\chi|^2\,dp_x dp_y dp_z = \iiint \Psi^* \left(\frac{h}{2\pi i}\frac{\partial}{\partial x}\right)^n \Psi\,dv \\ dv = dxdydz \end{cases}$$

(254) $$\iiint x^n |\Psi|^2\,dxdydz = \iiint \chi^* \left(-\frac{h}{2\pi i}\frac{\partial}{\partial p_x}\right)^n \chi\,dp_x dp_y dp_z$$

(255) $$\Delta x \Delta p_x \geq h/4\pi$$

(256) $$\begin{cases} H = H_{\text{free}} + H_{\text{interaction}} \\ \quad = H_{\text{free}} + eJ_\mu(x)A^\mu(x) \end{cases}$$

(257) $$M = \sum_n M_n \alpha^n$$

(258) $$H = H_{\text{free}}^{\text{phys}} + H_{\text{interaction}}^{\text{phys}}$$

mathematical rigour than was the case with old-fashioned field theory. As such, it has illuminated much of the mathematical foundations of field theory. The disturbing fact remains, however, that, thus far, no example of an interacting relativistic field theory satisfying all the required principles has been constructed, nor has it been shown that such a theory indeed exists. (R.Ro.)

DIMENSIONAL ANALYSIS

General considerations. Numerical quantities, the significance of which depends upon the system of units chosen for their measurement, are said to be dimensional. Other quantities (also the result of measurements) having values independent of the system of units used are called dimensionless, abstract, or pure numbers.

Thus, the size of a given angle, for example, may be specified equivalently as a_1 radians, a_2 degrees, a_3 right angles, etc. When considered under these circumstances, the given angle is therefore a dimensional quantity. If, however, consideration is limited to only such systems of measurement wherein angles are invariably measured in radians, then the size of an angle may be considered a dimensionless quantity, or a pure number. (Radian measure of an angle, a pure number, is defined as the ratio a/r, in which a is the length of the circumferential arc of a circle, intercepted by the given angle; r, its radius; and the vertex of the angle coincides with the centre of the circle. Because a and r both have the same dimensions—*i.e.*, a length—these cancel when the ratio a/r is formed, thus resulting in a pure number.)

Systems of measurement and their associated units (cgs = centimetre, gram, second; mks = metre, kilogram, second; foot, pound, second; BTU, or British thermal unit; calorie, etc.) play a fundamental role in dimensional analysis. The use of different systems of (dimensional) units inevitably leads to different numerical coefficients in the corresponding results; *e.g.*, one foot equals 12 inches equals 30.48 centimetres, approximately, etc. Intercomparison of measured and calculated quantities is much encumbered, however, unless all such quantities are expressed in the same system of units. With this objective in mind, some efforts on the part of the scientific community have therefore been made to adopt universally one standard system of units; *e.g.*, rationalized mks units in the study of electricity and magnetism. In addition to these, considerations of another and more fundamental kind are involved in the selection of a system of units. One of these is the intrinsic nature of the physical system to be dimensioned. Thus, although a fermi (10^{-13} centimetre) is a unit suitable for describing lengths relating to the atomic nucleus, a unit 100,000 times larger, called an angstrom (10^{-8} centimetre), is appropriate for expressing lengths related to atomic events. At the opposite end of the scale of spatial extension, the astronomical unit (equal to the mean Sun-to-Earth distance, or about 93,000,000 miles) is useful for describing distances within the solar system, but it is the light-year (distance traversed by light during one year) that is most appropriate for expressing interstellar distances.

It is apparent that in the examples cited, the unit definitions were made in a frankly *ad hoc* manner; *i.e.*, primarily for the purpose of keeping the associated numerical coefficients from becoming excessively small or excessively large. In this sense, such systems of measurement are artificial and arbitrary. With regard to certain other units, however, there exists some controversy as to whether they, too, are arbitrary or whether they possess a more profound significance indicative of some underlying physical reality. The so-called atomic unit of length (0.529×10^{-8} centimetre), and the energy unit called a rydberg (approximately 13.5 electron volts) are examples. The atomic unit (according to the Bohr model) is the radius of the energetically lowest circular electron orbit in the hydrogen atom. The rydberg represents the minimum energy required to remove this electron to infinity.

In any case, in some sense the selection of a system of units is comparable to the choice of a coordinate system appropriate to the physical situation at hand: When dealing with a system bounded by a cubical enclosure, rectangular Cartesian coordinates are indicated; but if the system's geometry is spherical (*e.g.*, a stellar model comprising homogeneous concentric shells) then the choice of spherical polar coordinates is naturally suggested. More generally, just as the form of a physical law remains unchanged regardless which particular one of a suitable set of coordinate systems is used to verify the law, so also are physical regularities generally independent of the particular system of units of measurement selected from among a set of such systems. (Physical laws formulated so that they possess this kind of form-invariance are said to be complete.) That this should be so is plausible: In view of the distinctly arbitrary considerations operative in the definition of a large number of systems of measurement, it is so exceedingly improbable as to be practically impossible that any physical regularity could be discovered if it did depend on one particular system of measurement. Conversely: If a physical regularity is discovered despite the arbitrary selection and use of one out of an infinitude of possible systems of measurement, then it is so exceedingly improbable as to be practically impossible that the manifestation of the observed regularity had any essential connection with the particular system of units employed.

Basis for choosing a particular system of units

The pi theorem. This independence of a physical law from the particular system of units employed (also called completeness) is expressed by the so-called pi (π) theorem, often attributed to Edgar Buckingham, a United States physicist. If $q_1, q_2, \cdots, q_n$ denote n physical parameters and dimensional constants (such as pressure, volume, thermal conductivity, and the acceleration of gravity) contained in a physical law, then the latter can always be expressed as a functional relationship of the form in which f as a function of n arguments q equals zero (see 259). The pi theorem asserts that the above relation can also be written as g as a function of arguments π_k equal to zero (see 260), in which each argument π is an independent dimensionless product of some of the q's. The physical law is thus demonstrably independent of any particular system of measurement, a statement previously merely held plausible. The theorem is subject to the restriction, generally of little practical significance, that there exists but one independent functional relationship between the q's.

An important result obtainable from the pi theorem is the so-called principle of dimensional homogeneity, often referred to as the fundamental theorem of dimensional analysis. If equation (260) is solved for one of the arguments, say π_1, and if π_1 is further written in terms of the physical quantities q (see 261), then equation 260 can be modified (see 262). Because h is a function of the dimensionless arguments π only, the function h is itself dimensionless. But π_1 can also be written as a ratio (see 263), and because this ratio (equal to π_1) is dimensionless, it follows that the dimensions of q_1^a in the numerator [and on the left-hand side of equation (254)] are necessarily the same as those of $q_2^b q_3^c \cdots$ in the denominator [and on the right-hand side of equation (262)]. The dimensions of all the terms in equation (262) are therefore the same; *i.e.*, the physical law is dimensionally homogeneous, which was to be proved.

Principle of dimensional homogeneity

Applications. *Use in developing an equation.* Dimensional analysis yields an amount of information dependent on the skill and experience of the analyst. Positive errors are never introduced by failure to recognize special situations—it is only that less than the maximum information is obtained. In the limiting case, in which nothing whatever will be contributed to a knowledge of the nature of the physical system, dimensional analysis does not give incorrect results, but only trivial results.

A conventional introductory problem is the simple pendulum. The physical parameters defining the simple pendulum may be taken to be the mass of the bob and

the length of the suspending rod. The problem is obviously one in mechanics, and the fundamental units may be taken as mass, length, and time. A detailed solution of the equations of motion will give the time of any complete oscillation, and this oscillation may be characterized by its angular amplitude. It is sufficient to show that the time of oscillation for any particular amplitude must be uniquely determined because the bob starts from rest at the extreme point of the oscillation and the boundary conditions and the equation of motion determine a unique solution. The equations of motion will involve the force on the bob, and this will involve the acceleration of gravity; that is, in the equations of motion will appear the dimensional constant g. The detailed solution should therefore give a relationship between mass of bob, m, length of pendulum, l, angular amplitude of oscillation, θ, acceleration of gravity, g, and the period of oscillation, τ. This is a complete relationship, being deduced by mathematical manipulations from a complete equation, and therefore the pi theorem applies. The parameters are as follows:

Parameter	Dimensional formula
m	M
l	L
θ	0
g	LT^{-2}
τ	T

Requirements for use of the pi theorem

The pi theorem requires the formation of all the independent dimensionless products. There will in general be two because the number of parameters is here five and the number of primary units is here three. One of these products, θ (theta), is obvious on inspection; and one dimensionless product must be formed from the four remaining parameters. The exponents in the dimensionless product are unknowns, written as α, β, γ, δ. These four quantities must now be found so that $m^\alpha l^\beta g^\gamma \tau^\delta$ is dimensionless. The dimensional formulas are substituted for the parameters. $M^\alpha L^\beta (LT^{-2})^\gamma T^\delta$ must be dimensionless, or the total exponents of M, L, and T must vanish. This condition gives three algebraic equations (see 264). There are three equations and four unknowns, which means that one unknown may be assumed arbitrarily. The exponent δ may be taken as the arbitrary one. Then the solution is: $\alpha = 0$, $\gamma = \delta/2$, $\beta = -\delta/2$. The dimensionless product is $l^{-\delta/2} g^{\delta/2} \tau^\delta$. The pi theorem states that the solution is some functional relationship between the two independent products (see 265), in which the functional form is completely undetermined.

This equation may be rearranged by solving for the quantity in parentheses (see 266), in which f is completely undetermined. This again may be rearranged (see 267). That is, the time of swing is independent of the mass, proportional to the square root of the length, and inversely proportional to the square root of the acceleration of gravity. As far as this analysis goes it may be any function of the amplitude. The detailed solution would show that, for small amplitudes, f has approximately the constant value 2π.

Use in changing units. By writing out the dimensional formula of a quantity, a method is afforded of finding how the numerical magnitude changes when the size of the primary units of measurement changes. The method is so simple that no elaborate or formal exposition is necessary; two examples, in which the quantities are, respectively, speed and pressure, are given.

A given body is moving with a speed of 60 miles (mi) per hour (hr) and it is desired to convert this speed to feet (ft) per second (sec). One mile contains 5,280 feet and one hour contains 60 × 60 seconds: therefore, simple equalities hold (see 268). Or, a pressure of the atmosphere of 15 pounds per square inch may be expressed in kilograms per square centimetre as follows: One pound (lb) is equal to 0.4536 kilogram (kg) and one inch (in.) is equal to 2.54 centimetres (cm). Hence equalities in different dimensions follow (see 269). The method illustrated is obviously general. In fact, it may be extended to much more complicated sorts of examples in which the conversion is between systems of units different in general nature, as from a system with mass, length, and time as fundamental to one with force, velocity, and energy as fundamental. (P.W.B./L.I.S.)

(259) $$f(q_1, q_2, \cdots, q_n) = 0$$

(260) $$g(\pi_1, \pi_2, \cdots) = 0$$

(261) $$\pi_1 = q_1^a q_2^b q_3^c \cdots$$

(262) $$q_1^a = q_2^{-b} q_3^{-c} \cdots h(\pi_2, \pi_3, \cdots)$$

(263) $$q_1^a / (q_2^{-b} q_3^{-c} \cdots)$$

(264) $$\begin{cases} \alpha = 0 \\ \beta + \gamma = 0 \\ -2\gamma + \delta = 0 \end{cases}$$

(265) $$F[(l^{-1/2} g^{1/2} \tau)^\delta, \theta] = 0$$

(266) $$l^{-1/2} g^{1/2} \tau = f(\theta)$$

(267) $$\tau = f(\theta)\sqrt{l/g}$$

(268) $$60 \frac{\text{mi}}{\text{hr}} = 60 \frac{5{,}280 \text{ ft}}{3{,}600 \text{ sec}} = 88 \frac{\text{ft}}{\text{sec}}$$

(269) $$15 \frac{\text{lb}}{\text{in.}^2} = 15 \frac{0.4536 \text{ kg}}{(2.54 \text{ cm})^2} = 15 \frac{0.4536 \text{ kg}}{(2.54)^2 \text{ cm}^2} = 1.054 \text{ kg/cm}^2$$

BIBLIOGRAPHY. A nontechnical discussion of Newtonian mechanics may be found in BARRY M. CASPER and RICHARD J. NOER, *Revolutions in Physics* (1972). Elementary treatments are included in many introductory physics textbooks, such as DAVID HALLIDAY and ROBERT RESNICK, *Physics*, vol. 1 (1966). ROBERT B. LINDSAY and HENRY MARGENAU, *Foundations of Physics* (1936), discusses the logical foundations of mechanics and other areas of physics. KEITH R. SYMON, *Mechanics*, 3rd ed. (1971), is a good intermediate textbook; for an advanced treatment the reader is referred to HERBERT GOLDSTEIN, *Classical Mechanics* (1950). A classic work on fluid mechanics is HORACE LAMB, *A Treatise on the Mathematical Theory of the Motion of Fluids* (1879; 6th ed., *Hydrodynamics*, 1945). A modern treatment is found in LOUIS M. MILNE-THOMSON, *Theoretical Hydrodynamics*, 5th ed. (1968); and ARNOLD J.W. SOMMERFELD, *Mechanik der deformierbaren Medien*, 3rd ed. (1954; Eng. trans., *Mechanics of Deformable Bodies*, 1964), which is also useful for the mechanics of solids. Another reference is IVAN S. SOKOLNIKOFF, *Mathematical Theory of Elasticity*, 2nd ed. (1956), entirely devoted to the subject. A classic work on elasticity is A.E.H. LOVE, *Treatise on the Mathematical Theory of Elasticity*, 4th ed. (1944). A good introduction to statistical mechanics is found in FREDERICK REIF, *Statistical Physics* (1967). Another good textbook (intermediate level) by the same author is *Fundamentals of Statistical and Thermal Physics* (1965). Two advanced works are LEV D. LANDAU and E.M. LIFSHITZ, *Statistical Physics*, 2nd rev. ed. (1969; orig. pub. in Russian, 1951); and KERSON HUANG, *Statistical Mechanics* (1963). Elementary treatments of electromagnetic theory may be found in almost any introductory physics textbook; see HALLIDAY and RESNIK (*op. cit.*), vol. 2. Intermediate treatments include PAUL LORRAIN and DALE R. CORSON, *Electromagnetic Fields and Waves*, 2nd ed. (1970); and JERRY B. MARION, *Classical Electromagnetic Radiation* (1965). A standard advanced work is JOHN D. JACKSON, *Classical Electrodynamics* (1962). Many of the original papers on the theory of relativity are contained in ALBERT EINSTEIN *et al.*, *The Principle of Relativity* (1923). An introductory work is ANTHONY P. FRENCH, *Special Relativity* (1968). G.F.B. RIEMANN'S fundamental paper, "Über die Hypothesen welche

der Geometrie zu Grunde liegen," can be found in his *Gesammelte Werk*, 2nd ed. (1892), later published with commentaries by HERMANN WEYL, 3rd ed. (1923). For an introduction to Riemannian geometry, see LUTHER P. EISENHART, *Riemannian Geometry* (1926, reprinted 1949); TULLIO LEVI-CIVITA, *Lezioni di calcolo differenziale assoluto* (1925; Eng. trans., *The Absolute Differential Calculus*, 1927); and CHARLES WEATHERBURN, *An Introduction to Riemannian Geometry and the Tensor Calculus* (1938). For an advanced treatment of relativistic mechanics, see GOLDSTEIN (*op. cit.*), and for electrodynamics, JACKSON (*op. cit.*). For general relativity the reader might consult GEORGE C. MCVITTIE, *General Relativity and Cosmology*, 2nd ed. (1965). DAVID S. SAXON, *Elementary Quantum Mechanics* (1968), is a good introduction to its subject. More advanced works are LEONARD I. SCHIFF, *Quantum Mechanics*, 3rd ed. (1968); EUGEN MERZBACHER, *Quantum Mechanics*, 2nd ed. (1970); and ALBERT MESSIAH, *Mécanique quantique*, 2 vol. (1959; Eng. trans., *Quantum Mechanics*, 2 vol., 1961–62). Quantum field theory is the subject of many advanced works, such as JAMES D. BJORKEN and SIDNEY D. DRELL, *Relativistic Quantum Fields* (1965). There are many works available on dimensional analysis. An older work, still of value, is PERCY W. BRIDGMAN, *Dimensional Analysis*, rev. ed. (1931, reprinted 1963). See also H.E. HUNTLEY, *Dimensional Analysis* (1952); CHARLES M. FOCKEN, *Dimensional Methods and Their Applications* (1953); and L.I. SEDOV, *Similarity and Dimensional Methods in Mechanics* (1959; Eng. trans. from 4th Russian ed. of 1957).

(R.C.B./J.F.Br./R.Wk./J.E.Ma./A.Ei./S.S.C./R.Ro./P.W.B./L.I.S./Ed.)

Physics

Physics is the basic physical science. It deals with those fundamental questions on the structure of matter and the interactions of the elementary constituents of nature that are susceptible to experimental investigation and theoretical inquiry. Its goal is the formulation of comprehensive principles, or laws of physics, that summarize disparate phenomena in the most general possible way and that typically are expressed with economy and precision in the language of mathematics.

Aims of physics

Originally only a part of what was termed natural philosophy, physics now consists of what remains of basic physical science with the development of separate and autonomous scientific disciplines of astronomy, chemistry, geology, and the technological applications of fundamental science in engineering. Despite the fact that the content of physics changes continuously with the advance of scientific knowledge, its aim has remained the explication of basic observable physical phenomena without primary attention to possible applications, whereas, in such borderline fields as chemical physics, astrophysics, geophysics, and biophysics, the techniques and principles of physics are applied to the principal discipline.

Both experiment, the observation of phenomena in precise quantitative terms under strictly controlled conditions, and theory, the construction in mathematical terms of a unified conceptual framework, play essential and complementary roles in the advance of physics. Experiment discloses the facts of nature; theory attempts to make sense out of them. All physical inquiry can be reduced ultimately into the study of events involving matter in space and time, and measurable physical quantities may be expressed in terms of basic units for length, time, and mass. When a theoretical formulation has summarized the results of experiments with a reliability so great as to reflect apparently universal behaviour, it is said to be a law of physics, but always tentatively. For, if further experiment fails to confirm its predictions, it must be modified or discarded or the limited range of its applicability clearly recognized.

Perhaps the ultimate aim of physics is to comprehend the properties of the fundamental constituents of nature and their mutual interactions in a single master plan and from it to deduce the properties of aggregates of particles and all macroscopic phenomena. Certainly the search for such a grand scheme lies at the forefront of contemporary physics. Although that ultimate synthesis has not been achieved, the number of laws of physics now regarded as fundamental is remarkably small. Moreover, one segment of physics, known as classical physics, which is that segment treating the behaviour of objects that are large compared with an atom and move with speeds far less than that of light, is essentially complete within its limited applicability to the macroscopic domain. Whereas classical physics accounts for low-speed, large-scale phenomena, the quantum theory and the theory of relativity, both originating early in the 20th century and constituting the foundations of modern physics, are the basis for elucidating the submicroscopic, high-speed phenomena in atomic, nuclear, and elementary-particle physics.

THE SCOPE OF PHYSICS

The traditionally organized branches of classical and modern physics are described below.

Mechanics. Mechanics is the study of the motion of objects and their interaction as comprehended by basic physical principles. In this sense all of physics may be regarded as mechanics. Whereas classical mechanics, or Newtonian mechanics, deals with objects the size of which is large compared with that of the atom and that move at speeds far less than that of light, quantum mechanics, or wave mechanics, is the basis for understanding atomic and subatomic phenomena, and relativistic mechanics concerns high-speed phenomena.

Suborders of mechanics

Within classical mechanics, kinematics is concerned merely with the description of the motion of a particle, whereas dynamics is concerned with a particle's motion in relation to the forces acting on it. Statics, a special case of dynamics, concerns systems of objects in equilibrium. Particle dynamics treats objects as having vanishingly small dimensions and an ignorable substructure, whereas in rigid-body dynamics extended objects are imagined to be incapable of deformation. Elasticity is the mechanics of deformable solids; and hydrostatics and hydrodynamics treat, respectively, fluids at rest and in motion.

Newton's laws of motion. Isaac Newton's three laws of motion are the foundations of classical mechanics. The first law recognizes that motion in itself does not require explanation or a cause: an undisturbed object moves in a straight line at constant speed. Mechanics concentrates, then, not on motion as such but on changes in the state of motion of an object, arising from its interaction with one or more other objects, which thereby exert a force on it. Newton's second law relates the net force on an object to the rate of change of its momentum. Thus, knowing the forces acting on an object allows, at least in priniciple, its future history to be projected in detail; conversely, by studying an object's motion, one may deduce the forces to which it is subject. Newton's third law, that of action and reaction, states that when two particles interact, the forces on each are of equal magnitude but opposite direction, so that forces always appear in equal and opposite pairs.

Force of gravity. The study of the universal force of gravity is customarily placed within classical mechanics. Indeed, Newton's accounting in detail for the planetary orbits, as well as such subtle gravitational effects as the tides and the precession of the equinoxes (the slow cyclical change in direction of the Earth's axis of rotation), through this fundamental force represents the first triumph of classical mechanics. Nothing beyond it, apart from the technologies of rocketry and communications, is needed for the principal aspects of space flight.

Momentum and energy. The concepts of momentum, angular momentum, and energy and the conservation laws related to them (see below) also play a central role in classical mechanics. Indeed, they offer alternative, and often simpler, means of analyzing collections of two or more objects. Newton's laws of motion and the law of gravitation predict the motion of a planet about the Sun, giving its location at any instant; apart from any details of this motion, the total momentum, angular momentum, and energy of the planet–Sun system are each constant.

Thermodynamics and heat. The general topic of heat and the formal study of thermal systems in thermodynamics are characterized by the concept of temperature. The particular power of classical thermodynamics lies in its ability to analyze the most general types of thermal systems without regard for their detailed microstructure.

First law. The first law of thermodynamics is, in effect, the energy-conservation principle (for all changes in an isolated system, the energy remains constant) generalized to include heat, which is energy transferred by a temperature difference.

Second law. Most succinctly stated, the second law of thermodynamics is that the entropy of an isolated system increases with time, where entropy may be interpreted as a quantitative measure of the state of relative disorder of the particles comprising a system. Although one may, in principle, use Newtonian mechanics to trace out in detail the motions of the many particles in some fairly complex system (for example, the molecules of a gas) and from their detailed microscopic behaviour deduce the macroscopic properties of the gas, such a program is mathematically forbidding, practically impossible, but actually unnecessary.

Statistical mechanics. Instead, through the procedures of statistical mechanics, one concentrates not on the detailed motion of every particle but on the average behaviour of a large number, invoking the laws of probability or its physical counterpart, the assumption of molecular chaos. The basic premise is that all possible configurations of the particles are equally likely. So viewed, a gas-filled container will have essentially equal numbers of molecules in the left and right halves for the same reason that, of many tossed coins, half are heads and half tails. The disordered states, those with molecules distributed uniformly throughout the container, are much more probable than such a highly ordered state as that in which all molecules crowd together in one corner. For the enormous number of submicroscopic particles encountered in any system of reasonable size, the number of disordered states is so great compared with the ordered states that any undisturbed system will almost invariably evolve from a state of lesser disorder to one of increased disorder.

The rules of chance

Statistical mechanics interprets thermal energy as merely the energy of submicroscopic particles in disordered states, not visible macroscopically, and temperature as a quantitative measure of how energy is shared among the particles.

From a microscopic point of view, the laws of thermodynamics imply that, whereas the total quantity of energy of any system is constant, what might be called the quality of this energy is degraded as the system moves inexorably, by the laws of chance, to states of increasing disorder until it finally reaches the state of maximum disorder, that state in which all parts are at the same temperature and none of the system's energy may be employed usefully. This final condition of chaos is sometimes referred to as the "heat death."

Electricity and magnetism. First conceived of as distinctive phenomena, electricity and magnetism are subsumed under the term electromagnetism, which treats the interactions of particles having the physical property of electric charge. Charged particles at rest interact by an electric force; when in motion, an additional magnetic force comes into play. Indeed, all magnetic effects may, in fact, be traced to electrically charged particles in motion.

The field concept

The field concept plays a central role not merely in the classical formulation of electromagnetism but in all areas of classical and contemporary physics. An electrically charged particle may be thought to create an electric field in all regions of space surrounding it; and the electric force between a pair of charged particles may be described through the intermediary of the electric field as follows: particle *A* creates an electric field at the site of particle *B*, and particle *B*, finding itself in this field, responds to an electric force.

Remarkably, all of classical electromagnetism is summarized in four equations originally formulated in the 19th century by the British physicist James Clerk Maxwell; these relations embrace the interactions between charged particles and the electric and magnetic fields associated with them. Implicit in these relations is the existence of electromagnetic waves—electric and magnetic fields in space detached from the charges that create them, travelling at the speed of light, and endowed with such "mechanical" properties as energy, momentum, and angular momentum. The visible radiation, or light, to which the human eye is sensitive is but one small segment of an electromagnetic spectrum extending from long-wavelength radio waves to short-wavelength gamma rays.

Optics. Because light consists of electromagnetic waves, the subject of optics, which treats the propagation of light, may be regarded merely as applied electromagnetism. It is customary, however, to subdivide this branch of physics into geometrical optics, which is concerned merely with the tracing of light rays, and physical optics, which treats the distinctive wave phenomena of light.

While the formation of images in such optical devices as the microscope and telescope may be accounted for simply by ray optics—*i.e.*, by assuming that the passage of light may be represented by straight lines, or rays—subtle effects attributable to the wave character of visible light require the explanations of physical optics. The basic wave phenomenon is interference, whereby two waves meeting at one point in space may combine to yield an enhanced resultant wave, the crests of the component waves then adding together; or, at the other extreme, the two waves may annul, the crest of one wave being superimposed on the trough of the other. A related effect is diffraction, the interference from a large number of wave sources. Because of diffraction effects, light spreads into geometrical shadows; and the image produced by any optical device is intrinsically fuzzy to a degree controlled by the wavelength. The wave properties of light are exploited in such instruments as the interferometer and diffraction grating, which allow for the measurements to a small fraction of a wavelength of light.

Atomic physics. Classical mechanics and classical electromagnetic theory are found to be fundamentally defective when applied to problems in atomic physics. Although the relatively light electrons of an atom are attracted to its small massive nucleus, the atom is not merely a submicroscopic solar system with attraction by gravity replaced by attraction between electrically charged particles. Atomic structure is comprehensible only on the basis of quantum mechanics, and its subtler details require as well the use of relativistic mechanics (for both, see the section on them below).

Because of the minuteness of the atom, its properties must be inferred by indirect experimental techniques. Chief among these is spectroscopy, which is concerned with the measurement and interpretation of the electromagnetic radiation either emitted or absorbed by materials. The atoms of each chemical element radiate a spectrum with distinctive wavelengths, which reflect the atomic structure. Through the procedures of wave mechanics, it is found that the atomic energies and characteristic wavelengths may be computed from certain basic atomic constants, namely, the electron mass and charge, the speed of light, and the fundamental characteristic constant of the quantum theory, Planck's constant. Because the numerical predictions of atomic quantum theory are based on these fundamental constants, one motivation for increasingly high precision in their measurement is the more refined comparison between experimental observation of atomic properties and theoretical predictions.

Atomic constants

Solid-state physics. At first considered only a part of atomic physics, solid-state physics has developed into a separate branch of physics that treats the electrical, magnetic, optical, and elastic properties of solid materials and, by extension, those of materials in the condensed state. Primary attention is given to crystalline solids, because the arrangement of atoms in simple geometric array yields theoretically tractable many-particle systems for application of the quantum theory. Because the atoms in a solid are coupled together, procedures different from those used in studying the structure of single isolated atoms must be applied. Thus, whereas an electron may be associated with a single parent atom when the atom is isolated, in conducting materials the free electrons, responsible for the thermal- and electrical-conduction properties, must be regarded as belonging to the entire collection of atoms.

Nuclear physics. Nuclear physics deals with the questions of nuclear structure and the radiation from unstable nuclei. Smaller than the atom by a factor of about 10,000, the constituent particles of the nucleus attract one another so strongly (by the still incompletely understood nuclear force) that nuclear energies are approximately 1,000,000 times larger than atomic energies. The quantum theory is the principal tool for the explication of nuclear structure.

Unlike the unstable atoms, which emit an electromagnetic spectrum, unstable radioactive nuclei may also emit such particles as electrons and positrons (positively charged electron antiparticles; *i.e.*, identical to the electron in mass but opposite in magnetic and electric properties), massless neutrinos, and helium nuclei. A principal research tool for investigating nuclear properties is the scattering experiment, wherein energetic particles (having exceedingly high velocities) are used as projectiles directed against nuclear targets; and the energies, directions, and identities of the emergent particles are analyzed. In such collisions the chemical identity of the nuclear target may be altered in a so-called nuclear reaction (*i.e.*, the element of which the target consists is transmuted into a different element through changes in the nucleus). The fission and fusion nuclear reactions, in which unstable nuclei are, respectively, split into smaller or amalgamated into larger and more stable configurations are characterized by the release of energy far exceeding that in analogous chemical reactions, such as the bonding of atoms into molecules or the breaking of molecular bonds.

Particles with extremely high energies are found as well in the cosmic rays, consisting of particles of extraterrestrial origin, mostly protons, that rain continuously upon the Earth's atmosphere.

Particle physics. The frontier of contemporary physics lies in the study of the elementary particles, the fundamental constituents of all matter. This branch of physics, referred to as particle physics, or high-energy physics, searches for patterns of coherence in the relationships between the remarkably large number of entities that are thought not to be composites of still more fundamental entities. In addition to the familiar and altogether stable electron and proton, the elementary particles include highly unstable particles existing for as short a time as 10^{-23} second. High-energy particle accelerators are required to create unstable elementary particles, and their interactions are typically studied with devices such as bubble chambers (see below). Some systematic organization schemes and family relationships have been perceived among the fundamental particles, but it is still not known whether the present research in particle physics is converging toward a synthesis.

Search for coherence

Quantum mechanics. Although the several branches of physics differ in their subject matter, experimental methods, and theoretical techniques, certain general principles apply equally well to all. The forefront of contemporary advances in physics lies in the submicroscopic; and the most active research areas—atomic, nuclear, solid-state, and elementary-particle physics—are all based upon quantum theory and the theory of relativity, which form the theoretical foundations of modern physics.

Whereas classical mechanics is characterized by physical quantities that may vary continuously over a range of possible values and are causally interrelated, the distinctive characteristics of the quantum theory are discontinuity and intrinsic uncertainty.

According to the quantum theory (also known as quantum mechanics or wave mechanics) electromagnetic radiation can no longer always be considered to consist of continuous waves; instead, it must under some circumstances be viewed as a collection of particle-like photons, the energy and momentum of each photon being directly proportional to its frequency. Similarly, the quantum theory regards such a particle as an electron as endowed with wave properties, the particle's wavelength being inversely proportional to its momentum. In both instances the proportionality constant is the characteristic quantum, or Planck's constant, named after the originator of the quantum theory, the German physicist Max Planck. Because of the minute size of the quantum constant, the characteristic, and sometimes paradoxical, quantum effects are evident only at the level of microscopic or atomic phenomena. Quantum physics reduces to classical physics when applied to the macroscopic domain.

At least in principle, all of atomic structure, including the periodic table of chemical elements, can be accounted for by combining the electric force between electrically charged atomic constituents with the rules of the quantum theory and the symmetry properties of space. Roughly speaking, each atomic electron, regarded as a wave, must be fitted around the nucleus. Consequently, atomic systems are restricted to certain discrete, quantized energies. When an atom undergoes a discontinuous quantum transition, or quantum jump, its energy changes abruptly by a sharply defined amount, and a photon of corresponding energy is thereby created and emitted, or annihilated and absorbed. Although atomic energies are sharply defined, an electron's location is not. The quantum theory can yield only the probability of finding the particle at any given location.

The feature that distinguishes quantum theory from all other physics is the uncertainty principle, or principle of indeterminacy, of the German physicist Werner Heisenberg: specifying a particle's location with increasing precision necessarily increases the uncertainty in its momentum and conversely, the degree of intrinsic uncertainty being controlled by the magnitude of Planck's constant. Because both a particle's location and momentum must be known in detail to project its future history under the action of known forces, such a program is precluded by quantum theory in the micro-world. The complementary wave and particle aspects, or wave–particle duality, of electromagnetic radiation and of material particles are another manifestation of the uncertainty relation. If an electron exhibits wave characteristics, as in the phenomenon of electron diffraction, then it does not, and cannot, exhibit particle characteristics in this observation; similarly, if electromagnetic radiation interacts with matter, as in the photoelectric effect, the process is effectively a collision involving particle-like photons, and the wave nature of electromagnetic radiation is necessarily suppressed. In the principle of complementarity, the Danish physicist Niels Bohr asserts, in this spirit, that different experimental conditions require different but equally essential descriptions, whether in terms of waves or in terms of particles.

Uncertainty principle

Although it deals with probabilities and uncertainties, the quantum theory has been spectacularly successful in elucidating otherwise inexplicable atomic phenomena; and it has thus far met every experimental test.

Relativistic mechanics. In classical physics space is conceived as having an absolute character as the empty background against which events in nature unfold as time flows onward independently; events occurring simultaneously for one observer are presumed to be simultaneous for any other; mass is taken as indestructible; and particles are able to acquire an indefinitely large speed given sufficient energy. The theory of special relativity, primarily the creation of Albert Einstein in 1905 and now so adequately confirmed by experiment as to be termed the law of relativity, shows all these and still other apparently obvious assumptions to be false.

Assumptions of relativity. The specific and sometimes bizarre relativistic effects flow directly from two basic postulates: (1) All observers, whatever their state of motion relative to a light source, measure the same speed for light. (2) The laws of physics are the same in all inertial reference frames. The first postulate, the invariance of the speed of light, is an experimental fact. From it follow the distinctive relativistic phenomena of space contraction, time dilation, and the relativity of simultaneity: as measured by an observer at rest, an object in motion is contracted along the direction of motion; moving clocks run slow; and two spatially separated events that are simultaneous for a stationary observer occur in sequence for moving observers. Three-dimensional space intervals and time intervals are thus inex-

tricably related in what may be termed four-dimensional space-time.

The second postulate of relativity theory, the covariance of the laws of physics, is also basic to classical mechanics. It implies, for example, that table tennis played on a (nonaccelerating) moving train is just like table tennis played with the train at rest, the states of rest and motion being indistinguishable. Because such mechanical quantities as momentum and energy involve the speed of a particle, their relativistic forms differ from their classical counterparts. In particular, the total energy E of any object of mass m equals the product of the mass and of the square of the speed of light, c, or $E = mc^2$, so that mass and energy must be regarded as equivalent and interchangeable. Further, c represents the ultimate speed limit for a particle, its mass, energy, and momentum all approaching infinite values as the particle speed approaches that of light.

Equivalence of mass and energy

Special and general relativity. The special theory of relativity is restricted to reference frames at rest or in motion at constant velocity relative to the fixed stars (inertial frames); the general theory of relativity is an extension to accelerated frames of reference. Based on the principle of equivalence—for example, an occupant in a closed capsule finding himself pressed against one wall cannot tell whether the force on him is due to the gravitational pull of an object outside or to the acceleration of the capsule—the general theory of relativity geometrizes the phenomenon of gravitation, weaving it into the fabric of space-time. A consequence, confirmed by experiment, is that light itself interacts gravitationally with objects; thus, a beam of light is deflected as it passes close to so massive an object as the Sun, and the wavelength of light escaping from a star is increased (the gravitational red shift). Attempts to geometrize the fundamental force of electromagnetism into a unified field theory of general relativity have thus far been frustrated.

Relativity theory does not show, as some vulgarized statements of it may imply, that "everything is relative." Quite the contrary: although time and space intervals, as well as other physical quantities, are relative to an observer, the speed of light and the laws of physics are absolute.

The conservation laws. First recognized in the early stages of development of classical physics, the conservation laws have come to play an increasingly significant role in contemporary physics. Indeed, much of today's research in elementary-particle physics may be regarded as a quest for conservation laws and their relation to the fundamental forces and particles.

Every conservation law concerns a collection of two or more objects, isolated from external influence, for which some physical quantity remains unchanged in time, or is conserved, quite apart from the details of what goes on between the interacting particles within the system. The generality and simplicity of conservation laws lie in the fact that they apply and may be exploited even in situations for which one is ignorant of the details of internal mechanisms.

Generality of the laws

The conservation laws now believed to be universally valid are those for the physical quantities: (1) linear momentum; (2) mass–energy; (3) angular momentum; (4) electric charge; (5) electron number; (6) muon number; (7) baryon number. The classical counterparts of the first four laws are familiar in classical physics; the last three appear in the study of elementary particles. Roughly speaking, the linear-momentum conservation law implies that a particle cannot fly off in an explosion in one direction without a second particle flying off in the opposite direction. The mass-energy conservation law, recognizing the relativistic equivalence of mass and energy, combines the separate laws of mass and of energy conservation. The angular-momentum conservation law implies that, if one part of an object initially at rest is set spinning in one sense, another part must spin in the opposite sense. The electric-charge conservation law implies that an electrically neutral system may be comprised of equal amounts of positive and negative charge. For example, in the mutual annihilation of an electron with its antiparticle, a positron, their mass disappears; and two uncharged photons are created with the same total energy and fly off in opposite directions, thereby conserving mass–energy, linear momentum, and electric charge. The fundamental character of these four conservation laws is evident in the fact that each law may be deduced from or related to a symmetry property of space and time, together with some general requirements of relativistic electromagnetic theory.

The modern view of fundamental interactions among the elementary particles is that any conceivable process will occur, except as prohibited by a conservation law. Consequently, the failure to observe a process may imply the operation of still another conservation law and a related fundamental symmetry principle.

The conservation laws of electron number, muon number, and baryon number apply to groups of particles identified as belonging to the electron, muon, and baryon families. For example, the proton is perfectly stable; it never decays into less massive particles. This, and similar observations, identifies the proton as the lightest member of the baryon family of relatively massive elementary particles. Still other conservation laws have been identified that are not universally valid but are associated with special categories of particles.

The fundamental forces and fields. All of the forces of nature now known can be traced to four fundamental interactions: the gravitational force acting between any two particles with mass; the electromagnetic force acting between any two particles with electric charge; the strong interaction, or nuclear force, which acts between particles within the nucleus and in other interactions between massive particles; and the so-called weak interaction, which is illustrated in the decay of the neutron into a proton, electron, and antineutrino.

According to quantum field theory, each of the four fundamental interactions is mediated by the exchange of particles associated with the particular force. In classical electromagnetism a charged particle is thought to create a continuous electromagnetic field that acts on a second charged particle to transmit the force. According to quantum field theory, each charged particle continuously creates and reabsorbs field particles, consisting of virtual and unobservable photons; and the trading of such virtual photons between charged particles is the origin of the electromagnetic force. Thus, the field for each force is imagined to consist of corresponding field particles, and all processes occurring at the level of the elementary particles may be conceived to consist of the creation, annihilation, and interchange of particles. Although bizarre, this view of elementary microphysics is validated by its success in accounting for a large variety of processes and even in predicting the existence of heretofore unobserved particles.

Field particles

THE METHODOLOGY OF PHYSICS RESEARCH

Physics has evolved and continues to develop without any single strategy. Careful refinements in measurement to uncover anomalous behaviour, mathematical extrapolation of existing theory, critical re-examination of apparently obvious but untested presuppositions, argument by symmetry or analogy, aesthetic judgment, accident, and hunch all play a role. Thus, the quantum hypothesis of the German physicist Max Planck was based on the observed departures in blackbody (a body that absorbs all radiant energy, reflecting none) radiation from the predictions of classical electromagnetism; the English physicist Paul A.M. Dirac predicted the existence of the positron, and of antiparticles generally, by mathematically extending the quantum theory; the elusive neutrino, without mass or charge, was hypothesized by the Austrian-born physicist Wolfgang Pauli as an alternative to abandoning the conservation laws in the phenomena of beta-decay processes; James Clerk Maxwell conjectured that, if changing magnetic fields create electric fields, perhaps changing electric fields create magnetic fields; and Albert Einstein's formulation of special relativity theory was based upon a critical re-examination of the meaning of simultaneity, apparently without knowledge of the very

experiments on the speed of light that strongly supported his assumptions.

Whatever the detailed tactics, the physicist invariably attempts to render unsolved problems tractable by using successive approximations, first treating simplified or idealized situations to identify the principal effects and only later moving to finer details. Thus, in the kinetic-theory model of an ideal gas, the molecules are imagined to be particles like billiard balls with vanishing dimensions.

Correspondence principle

One general guiding principle, useful in extending theoretical formulations and first introduced by Niels Bohr in a more restricted context, is the correspondence principle: when a more general theory is applied to a situation in which a less general theory is known to apply, the more general theory must yield the same predictions as the less general theory; or, when theories correspond, the predictions must correspond. For example, when the wavelength of light is so small it can be ignored, the more comprehensive theory of physical optics must yield the same results as the more restricted theory of ray optics. Similarly, quantum mechanics must yield the same results as classical mechanics when the wave character of particles is small enough to be ignored or, equivalently, when the small value of the characteristic quantum constant (Planck's constant) may be taken to be zero. In like fashion, relativity physics coincides with classical physics for phenomena at such low speeds that the characteristic constant of relativistic physics (the speed of light) may be regarded as effectively infinite.

The fashion in which the experimentalist and theoretician attack and solve problems in physics may be illustrated by a few examples.

How are the elementary particles created and studied? High-energy accelerators, enormous in both size (up to miles) and cost (hundreds of millions of dollars), accelerate particles to such high energies (up to 500,000,000,000 electron volts) that, when they strike a target, unstable elementary particles are created in the conversion of energy into mass. The life history of such highly unstable, short-lived particles and the particles into which they decay may be studied in photographs taken of the tracks left by charged particles as they move through and create small bubbles in liquid hydrogen near its boiling point contained in a bubble chamber. Particles may also be detected one by one, and their energies measured, in high-resolution scintillation detectors. In such a device, a particle passing through a transparent material excites atoms to emit a weak pulse of light, which in turn dislodges an electron through the photoelectric effect. The electron current is amplified electronically, and the final output pulse is directly proportional to the energy of the initiating particle. When numerous particles are detected in rapid sequence, the output pulses from the detector may be fed to a multi-channel analyzer, which sorts and counts the pulses according to their size. What emerges finally is a record of the number of particles passing through the detector and their distribution in energy.

Use of mathematics

The theoretician uses mathematics first as a computational tool. Thus, Newton invented the calculus to solve the following problem, essential to his formulation of the universal law of gravitation: how does the Earth, imagined to consist of a large number of particles distributed throughout a sphere, attract a small object at its surface, assuming that the attractive force between any pair of particles is inversely proportional to the square of the distance separating them? The integral calculus, a procedure for summing many small contributions, yields the simple result that the Earth is equivalent to a single particle at its centre with a mass equal to that of the Earth.

The theoretical physicist also uses mathematics as a tool of logic—that is, as a means of translating one statement into another, possibly more useful, equivalent statement. Newton's work on gravity again provides an example. Astronomical observations showed that each planet, even when not maintaining a constant distance to the Sun, moves in an orbit such that an imaginary line joining the Sun to a planet sweeps out equal areas in equal time intervals. By strictly geometrical reasoning, this statement can be shown to be equivalent to the physically more significant statement that the force of the Sun on each planet always points exactly toward the Sun.

RELATIONSHIP OF PHYSICS TO OTHER DISCIPLINES AND TO SOCIETY

Impact of physics on related disciplines. Because physics elucidates the simplest fundamental questions in nature for which one may reach consensus, it is hardly surprising that it has had a profound impact on technology, other fields of science, and philosophy.

Whenever a branch of physics has reached such a degree of maturity that its basic elements are comprehended in general principles, it has moved from basic physics to applied physics, or technology. Thus, all of classical physics has become effectively applied physics, its contents forming the foundation of many branches of engineering. Discoveries in modern physics have been converted with increasing rapidity into technological innovations or analytical tools for bordering scientific fields. There are, for example, such newly created fields as nuclear engineering, solid-state electronics, quantum chemistry, and radio astronomy and such analytical tools as X-rays, radioisotopes, spectroscopy, and lasers, all originating from basic physics.

Apart from its applications, physics—especially Newtonian mechanics—has become the prototype of the method of scientific inquiry, its experimental methods and mathematical analysis sometimes being imitated in fields far removed from the physical sciences, for which physics has also served as a model.

Effect of physics on philosophy of science

The significant influence of physics on those branches of philosophy that treat the conceptual foundations of natural phenomena is indicated by its earlier designation as natural philosophy. Determinism, the philosophical doctrine that the universe is a vast machine operating on a strictly causal basis, with its future determined in detail by its present state, is rooted in the Newtonian model of mechanics, in which all future positions and velocities of a particle are determined completely by the forces on it. Moreover, the schools of Materialism, Naturalism, and Empiricism have considered physics as a model for philosophical inquiry. An extreme position is that of Logical Positivism, with its radical distrust of any elements in discourse not directly observable, whose advocates suggest that all significant statements must be formulated in the language of physics.

The uncertainty principle in the quantum theory has prompted a re-examination of the question of determinism; and the philosophical implications of this branch of physics, particularly the meaning of measurement, remain unresolved. It is fair to say that, although physicists agree that quantum mechanics works, they still differ as to what it means.

Reciprocal relationships

Impact of related disciplines on physics. The relationship of physics to its borderline disciplines is reciprocal. Just as technology feeds on fundamental science for new practical innovations, physics quickly appropriates the techniques and instrumentation of modern technology as tools for advancing basic physics. Thus, the experimental physicist utilizes increasingly refined and precise electronic devices; and he works closely with engineers in designing such machines as high-energy particle accelerators. Mathematics has always been the primary tool of the theoretical physicist, and even abstruse branches of mathematics such as group theory have become invaluable aids to the theoretician classifying elementary particles or investigating the symmetry characteristics of atoms and molecules. Much of contemporary research in physics simply could not be carried out without the high-speed computer. It allows the theoretician to undertake computations that are too tedious or complicated by conventional means; and it allows the experimentalist, by incorporating the computer "on line" into the output of his apparatus, to be provided nearly instantaneously with summarized data while an experiment is in progress.

Physics and education. Because physics is the foundation of the physical sciences and engineering, scientists

and engineers are trained in at least the elementary aspects of physics; moreover, the increasingly prominent role of quantum mechanics and nuclear physics in many branches of chemistry and engineering has resulted in further emphasis on these specialties. Professional physicists who make significant contributions in research or in applied physics typically have completed formal studies through the level of the Ph.D. degree, specializing in a subbranch of experimental or theoretical physics. They are employed chiefly in government and industrial research establishments and in academic institutions.

Professional organizations of secondary school and college teachers began in the late 1950s to give special attention to improving the quality of teaching of physics; among the efforts in secondary school curricular reform in physics in the United States are the Physical Science Study Committee and the Harvard Project Physics, programs made possible by substantial support from the federal government and foundations. The United Nations Educational, Scientific, and Cultural Organization (UNESCO), an international organization promoting the exchange of scientific and cultural information, has also been concerned with education in physics.

The physicist in society. Partly because of the esoteric character of much of contemporary physics and its remoteness from ordinary experience, physicists have sometimes appeared to be initiates in a latter-day secular priesthood who speak an arcane language and communicate the significance of their findings to laymen only with great difficulty. Yet the physicist has come to play an increasingly significant role in society, particularly since the massive efforts of physicists during World War II in developing nuclear weapons and radar. They have received substantial governmental support for research studies, even when not directly in the government employ, and the construction of such enormously costly installations as high-energy particle accelerators has been possible only through the creation of national and international laboratories, as illustrated, respectively, by the National Accelerator Laboratory at Batavia, Illinois, and the European Organization for Nuclear Research (CERN), comprised of 12 European member states and operating a large accelerator at the French–Swiss border.

The enterprise of basic physics is strongly dependent upon public support and funding; and with this development has come, albeit slowly and starting with the development of nuclear weapons, a growing recognition within the community of physicists of the social responsibilities of scientists for the consequences of their work and the more general problems of science and society.

Exchange of information in physics. The principal mode of exchange of information among physicists on scientific advances is through the publication of the results of their researches in professional journals. With the proliferation of specialized branches, the number of physics journals has reached several hundred, most published under the auspices of a comparable number of professional societies of physicists, the total international membership of which may be estimated to exceed 100,000. Among the largest groups, each comprised of member societies in specialized fields, are the European Physical Society and the American Institute of Physics. The latter organization has been active in devising improved schemes for abstracting, indexing, and making accessible information, and it maintains a Center for the History and Philosophy of Physics at New York City, the purpose of which is to document the history of recent physics and to collect archival materials.

The chief activity of the International Union of Pure and Applied Physics has been the defining of units in which physical quantities are measured and the promoting of uniformity in other matters of nomenclature.

BIBLIOGRAPHY

General introductory texts: H.Y. CARR and R.T. WEIDNER, *Physics from the Ground Up* (1971); ROBERT RESNICK and DAVID HALLIDAY, *Physics*, 2 vol. (1966); RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics*, 3 vol. (1963–65); R.T. WEIDNER and R.L. SELLS, *Elementary Classical Physics*, 2 vol. (1965) and *Elementary Modern Physics*, 2nd ed. (1968); L.W. TAYLOR, *Physics: The Pioneer Science* (1941); G. HOLTON and D.H.D. ROLLER, *Foundations of Modern Physical Science* (1958).

Relatively nontechnical expositions of topics in modern physics: K.W. FORD, *The World of Elementary Particles* (1963); MAX BORN, *The Restless Universe*, 2nd ed. rev. (1951); D.H. FRISCH and A.M. THORNDIKE, *Elementary Particles* (1964); *The World of the Atom*, 2 vol., ed. with commentaries by H.A. BOORSE and L. MOTZ (1966).

A classic exposition of relativity theory by its originator: ALBERT EINSTEIN and LEOPOLD INFELD, *The Evolution of Physics*, rev. ed. (1961).

General reference work with detailed articles on all branches of physics: *McGraw-Hill Encyclopedia of Science and Technology*, 3rd ed., 15 vol. (1971).

The strategy and philosophy of physics: RICHARD P. FEYNMAN, *The Character of Physical Law* (1965); A.B. ARONS and A.M. BORK (eds.), *Science and Ideas: Selected Readings* (1964); P.A. SCHILPP (ed.), *Albert Einstein: Philosopher-Scientist*, 2 vol. (1959); R.B. LINDSAY and H. MARGENAU, *Foundations of Physics* (1936); W. HEISENBERG, *Der Teil und das Ganze* (1969; Eng. trans., *Physics and Beyond*, 1971); P.W. BRIDGMAN in L.W. FRIEDRICH (ed.), *The Nature of Physical Knowledge* (1960); J.R. OPPENHEIMER, *The Open Mind* (1955). Reports on current developments in physics and the activities of physicists appear in the monthly journal, *Physics Today*.

(R.T.We.)

Physiographic Effects of Man

Man accomplishes some physiographic changes directly, for instance by quarrying rock or digging canals, and also contributes to the development of the Earth's surface features by influencing the performance of other agencies, as when he promotes soil erosion by clearing vegetation or initiates channel incision by building a dam. Such indirect effects are often unforeseen and may be far more powerful and long-lived than the force by which they were set in motion or directed.

No clear distinction can be drawn between artificial and natural landforms; indeed, the direct physiographic effects of man have invited comparison with the work of glaciers, volcanoes, and termites. Though not wholly felicitous, these parallels are collectively instructive in bringing out the variety of man's handiwork. In terms of scale, the features he constructs or excavates are dwarfed by comparable landforms fashioned by other agencies; and when viewed against the geological time scale, the cumulative effect of man's activities appears insignificant. On the other hand, the indirect agency of man is outstanding in its pervasiveness, no physiographic process being wholly immune from its influence.

The role of man in fashioning the landscape has attracted interest since antiquity. Detailed investigation began in the late 18th century and advanced concurrently with the marked acceleration in the growth of man's power to intervene. In the course of these two centuries, man has come to dominate the physical environment of restricted areas, and to affect the general environment to some degree over the entire globe. The combined pressure of technological advance and population increase has tended to enhance and accelerate the influence of man through time.

This article treats the emergence of man as a physiographic agency and summarizes some of the more important direct and indirect ways in which he influences and alters the Earth's surface features. Better understanding of this subject can be gained from knowledge of the other forces and processes that influence landforms. Hence, the interested reader should see LANDFORM EVOLUTION; PHYSIOGRAPHIC EFFECTS OF TECTONISM; FLUVIAL PROCESSES; SEDIMENT YIELD OF DRAINAGE SYSTEMS; COASTAL FEATURES; GLACIATION, LANDFORMS PRODUCED BY; WIND ACTION; and CLIMATIC CHANGE. For other articles treating man's works and projects, see also MINING AND QUARRYING; AGRICULTURAL SCIENCES; CANALS AND INLAND WATERWAYS.

RISE OF HUMAN AGENCY

Importance of fire

Man became an important influence on physiographic change with the help of fire. Repeated burning of the

vegetation cover alters its composition and hence its contribution to soil development; it can also lead to erosion by wind and water. To judge from the archaeological record, man's control of fire spans an estimated 300,000 years in Europe and Asia, over 50,000 years in southern Africa, and at least 12,000 years in North America. The firing of vegetation in driving game and improving the yield of edible plants has been practiced by hunting and collecting communities in many parts of the world and presumably played a part in Paleolithic economics. Its cumulative effects can only be conjectured. Nevertheless, there is botanical evidence to suggest that the world's principal grasslands and prairies, as well as some major woodlands (among them the pinewoods of the southeastern United States and the teak forests of Burma), are the product of repeated burning over millennia. Lightning fires are too infrequent to furnish an adequate explanation. Should the agency of man be confirmed, the soils associated with the tropical and temperate grasslands would rank among the most prominent incidental effects of human agency now in evidence.

The transition from hunting and gathering to stock-raising and cultivation was slow and piecemeal; extensive territories long remained subject to temporary exploitation. The outcome was mainly to concentrate and intensify the effects of man and the animals he needed on the vegetation cover: permanent clearings spread, grazing was increasingly controlled, and preferred plant species were introduced. The trend was accentuated by the rise of farming settlements such as those to be found in many parts of Southwest Asia by 6000 BC, for they made heavy demands on timber and fuel as well as the products of plant and animal husbandry. The introduction of tillage practices enabled man to modify soil structure directly; once animal power was diverted to this purpose, areas subject to prolonged cultivation were gradually sculptured into a stepped or furrowed configuration.

Although clearing, overgrazing, and cultivation were locally accompanied by soil erosion, many of the areas thus affected had undergone much more severe erosion during Paleolithic occupation. It is not clear how far burning contributed to a primarily climatic phenomenon, but there is little justification for assuming that, until the advent of farming, man invariably respected the equilibrium between soil-forming and erosive processes.

Ancient irrigation works

Where irrigation was called for and favoured by social and physical conditions, the manipulation of soil and water left a clear imprint on the landscape. Cultivation terraces, some of them more than 3,000 years old, occupy thousands of square kilometres in Asia, Europe, Africa, and the Andes area of South America. The dikes, canals, and embankments on which the earliest urban societies depended can be traced on the great alluvial plains of the Old World. Ridged fields created in pre-Columbian times to exploit land subject to seasonal flooding have recently been identified in many parts of tropical South America. In North Africa, the control of ephemeral floods in the coastal uplands awaited Carthaginian settlement and reached its apogee in Roman times; elsewhere, it was not until the Middle Ages or later that the action of running water was significantly modified by engineering works.

Such devices had implications for landscape development that went beyond their immediate purpose. Terracing reduced the sediment yield of river basins, and the soils they held developed distinctive characteristics, especially when subject to waterlogging. In parts of Mesopotamia irrigation was not matched by adequate drainage and soon led to the accumulation of noxious salts at the surface. The embanking of streams promoted silt deposition within their channels, which, in turn, rendered them unstable and increased the incidence of flooding; the Huang Ho is a case in point. In the Nile Valley, millennia of flood irrigation have contributed to the processes tending to build up the plain; in the Tafilalt oasis of Morocco, medieval stream diversion provoked the accumulation of alluvium, reaching a maximum depth of five metres (16 feet).

Where runoff had been brought under control by damming, as in Roman Libya, or by underground drainage, as in Etruscan Italy, valley erosion was checked and the topography now contrasts sharply with that of river basins not thus protected. Conversely, irrigation made it possible to cultivate marginal areas where the role of vegetation in shielding the soil from water and wind action was especially important; the Romans, at other times hailed as supreme practitioners of conservation, have on these grounds been held responsible for the decay that ultimately overtook large parts of their empire.

Materials for the pyramids and other works

The building of cities, engineering works, harbours, monuments, and defenses had as its counterpart the extraction of large volumes of rock and clay. The Sadd-el-Kafara, a dam built south of Cairo between 2950 and 2750 BC, alone embodied 100,000 tons of material; the Great Pyramid of Khufu at Giza consists of 2,300,000 blocks with an average weight of 2.5 tons. Most of the material used in construction was obtained from quarries and floodplains.

Shafts had been employed in Neolithic flint mines throughout western Europe, notably at Spiennes, Belgium, where the resulting pitted landscape extends over two hectares (five acres). By the early fourth millennium BC, shafts were in use in the copper mines of Armenia and the Caucasus. Shafts also made possible the digging of tunnels called *qanats* or *foggara* that were as long as 20 kilometres (12 miles) in order to tap the underground water held by alluvial deposits; the aligned crater-like depressions by which they are marked are a prominent feature of the Iranian desert, where the technique has been known for 3,000 years. The scars of mining and of other forms of excavation spread with the adoption of new minerals and fuels and the dispersion of technical knowledge; *qanats* are to be found in other parts of Asia, in Europe, North Africa, and the Americas.

Aerial photography has played an important part in filling in the details of the ancient man-made landscape. It has revealed the grid laid out by Roman surveyors in Tunisia and other parts of the Mediterranean, the puzzling Nazca designs of the Peruvian desert, and many mounds and dikes that fulfilled military and religious functions. The picture that emerges is one of gradation between areas subject to long-standing, though varied, direct and indirect modification, and others where man's handiwork betrays shortlived, opportunist use of the land.

EFFECTS OF MODERN MAN

In reviewing the effects of modern man on the landscape, it is convenient to classify his activities in terms of the geological processes to which they most closely correspond. The issue of motives, and any assessment of the outcome in economic or aesthetic terms, are not central to this article.

Direct effects. The technological and demographic upsurge that characterized the 18th and 19th centuries was marked by a rapid increase in the power of direct human intervention; by 1922 it was possible to claim that, in the densely populated British Isles, man's capacity for removing and comminuting rock exceeded that of all the atmospheric forces combined.

Land cultivation and mining. The uneven spread of this power is illustrated by the distribution of cultivation methods. An estimated 11 percent of the world's land area is under cultivation, of which one-tenth is irrigated. The level of technology employed ranges from digging sticks to mechanical devices that wholly transform the existing soil structure, and from the occasional diversion of spates to the application of specified volumes of water at controlled rates. Soil chemistry may be modified by the abstraction of nutrients by plants until their exhaustion, or it may be manipulated by rotation and chemical fertilization. Other weathering agents are influenced: frost action may be encouraged to improve tilth, or checked if injurious to the crop.

The boundaries of the sown areas continue to fluctuate. In New England, large areas have reverted to forest in the course of the last century; conversely, drainage has added over 600,000 hectares (1,500,000 acres) to the cultivable

territory of Finland since 1941. The margins of arid areas are rendered especially unstable by rainfall variability.

The relief produced by plowing and the excavation of drainage and irrigation canals is limited by the stability of the soil material as well as by the available power and equipment. In this respect it conforms in vertical scale to the products of water erosion, however much it may depart from the fluvial landscape in its geometrical pattern. In contrast, the depths attained by open-pit mining are determined primarily by the economics of mining. The copper pit at Bingham, Utah, is over 800 metres (2,600 feet) deep and has not yet reached the limits of rewarding excavation. Laterally, the scope is technically unlimited, although the largest quarries, such as the Hull-Rush-Mahoning iron-ore pit in Minnesota, remain far smaller than the Aso caldera of Japan and other volcanic features.

Quarries and excavations

Open-pit mining accounts for 70 percent of the world's output of minerals and rocks. The excavations range from small borrow pits, such as the 300,000 marl pits that dot East Anglia, England, to vast stepped amphitheatres. The strip-mining of coal produces its own distinctive topography, as does the working of placer deposits by hydraulic methods or by dredging. The terrain thus affected is rapidly expanding; in the United States, for example, the production of natural aggregates increases by 5 percent each year. Because sand, gravel, and rock suitable for crushing are low in value, their extraction tends to be concentrated near urban and industrial sites. This tendency is countered, however, by the needs of highway and aqueduct construction, and it may be largely ignored where more precious minerals are involved, as in the case of the great copper open-pit mine of Chuquicamata in the Atacama Desert of Chile. Military and scientific activities, such as the explosion of atomic devices, and the craters produced by bombing, have further dispersed man's erosional activities.

Man-made pits are often bordered by waste tips and mounds. The proportion of usable to unusable solids ranges from less than 20 percent of the output in sands and gravels to as much as 14,000 percent in the case of slate. Contributing to the process are such underground workings as the gold mines of the Witwatersrand, South Africa, which yield some 60,000 tons of waste each year. The potential instability of some of these features is brought out by the collapse of tips at Espenhain, East Germany, in 1959 and Aberfan, Wales, in 1966. The unconsolidated material of which they are composed has constructional value, and, together with urban and industrial waste, it is employed to level and create building land in response to economic and social pressures.

Artificial channels and coastal works. Surface waters are more widely affected. The density of the drainage network is increased by canals, ditches, and other artificial channels. Surface drains may triple the length of channel that would otherwise characterize a particular area; in parts of Scotland the increase has been twenty-fold. Conversely, subsurface drainage can lead to the local elimination of organized runoff. Canals, including the Panama cut, exploit existing watercourses and also disregard watersheds; the channel that links Abidjan, Ivory Coast, with the sea bears little relationship to the natural drainage. The geometry of the channels and their floodplains is often modified to improve the efficiency of drainage; dikes, embankments, and auxiliary canals remain a prominent feature of the fertile alluvial valleys of the Old World and have transformed the valley of the Mississippi River. Lake Kariba, over 280 kilometres (170 miles) long, on the Rhodesia–Zambia border, is the outstanding example of further diversification produced by artificial lakes and reservoirs.

Harbours, breakwaters, and dikes

Coastal modification similarly ranges from the minor adjustment of existing features to the creation of novel landforms. Most of the major harbours of the present day, including those of Sydney, San Francisco, and New York, occupy propitious sites; a few are largely artificial, such as those of Takoradi, Ghana; Sète, France; and Algiers, Algeria. The port of Ischia, Italy, lies on a crater lake that has been opened to the sea. Offshore islands are also exploited in the building of breakwaters; artificial islands—for instance, Rincon Island in the Gulf of Mexico—may be the solution to the problems posed by offshore drilling.

Protective works range from those designed to preserve the existing situation, such as sea walls, to groins and other devices that work in conjunction with the transport by waves and currents. The creation of land at the expense of the sea has been most successfully pursued in areas already subject to coastal sedimentation. In England, the area around the Wash has gained 16,000 hectares (40,000 acres) of land since the 18th century. In the Netherlands, over 740,000 hectares (1,800,000 acres) have been won in this way since the early 13th century; the dike that blocks the Ijsselmeer lies at a maximum distance of 85 kilometres (53 miles) from the former coast and has reduced the length of the coastline by about 300 kilometres (200 miles). The material yielded by the dredging of existing facilities is sometimes used for reclamation rather than being dumped at sea: a combination of dredging and filling was employed in developing San Juan Bay, Puerto Rico.

Man's direct intervention is seen to differ from other physiographic processes chiefly in its versatility. It is not unique in effecting both chemical and physical changes and transcending the distinction between the organic and inorganic worlds. Tree roots, for example, break up rocks by mechanical means and also influence microbial activity. But human agency is not confined to a specified range of environments or sources of energy. Glaciers and rivers are creatures of topography and climate; volcanoes are restricted to well-defined zones; and even the action of wind is strongly localized. Man, however, is potentially free of such constraints.

Indirect effects. The incidental effects of man's activities on the landscape have long attracted attention. Plato, for instance, ascribed the barrenness of Attica to soil erosion consequent upon deforestation; Pausanias pointed out that silting at the mouths of the Achelous and Maeander rivers was accelerated by cultivation within their catchments. Later writers came to regard such phenomena as indicative of the tendency by modern man to disturb the equilibrium attained by nature, a viewpoint that was powerfully endorsed by the soil erosion that afflicted the U.S. Midwest in the 1930s.

The difficulties to be faced in measuring such unforeseen and often undesirable developments have long been emphasized. The need remains to rely on historical evidence and on deduction based on present-day parallels, and many of the effects are still poorly understood. Nonetheless, advances in the study of soil mechanics, meteorology, geophysics, and allied fields have thrown light on some of the mechanisms by which indirect human agency operates and have demonstrated that the consequences are sometimes out of all proportion to the energy involved in man's intervention.

Land subsidence and slope failure. In many cases the role of man is self-evident and easily evaluated. The underground extraction of solids and fluids, for example, can lead to subsidence, and its prediction presents few problems. The salt fields of Cheshire, England, are widely affected by such activity, and flooding of the depressions has formed lakes known locally as meres or flashes. The coal-mining areas of Pennsylvania provide instances both of subsidence and of collapse. Where water, oil, or gas is withdrawn from an underground stratum that is poorly consolidated, the loss of supporting pressure provided by the fluid can result in compaction. Mexico City has been subject to subsidence since pre-Columbian times because of drainage by surface channels; the process was greatly accelerated by pumping to supply urban needs, and sinking at the rate of 15 centimetres (six inches) per year was observed in places. Despite restrictions placed on pumping, parts of the city are still settling by about three centimetres each year. In the Wilmington, California, oil field, the abstraction of petroleum and gas has produced depression of the surface over an area of about 56 square kilometres (22 square miles), to a maximum depth of nine metres (30 feet). The application of large surface

Underground fluids

loads also leads to depression; at the site occupied by Lake Mead, Nevada, for example, a maximum of 175 millimetres (seven inches) of subsidence has occurred in the course of 15 years, a figure that comes close to the predicted elastic yield of the Earth's crust.

Analogous topographic changes can arise through drainage or irrigation of the soil. The drainage of organic soils promotes their shrinkage through the decomposition and oxidation of organic matter. Where the superficial deposits consist of loose, dry material of low density, such as wind-lain particles, the application of water sometimes produces the phenomenon of hydrocompaction. Over 500 square kilometres (200 square miles) have been thus affected in the San Joaquin Valley of California, and in places depression amounts to more than five metres.

Slope failure, like subsidence, occurs when gravitational stresses overcome the resistance of the rock material (see EARTH MOVEMENTS ON SLOPES). Man may bring this about by undercutting and steepening slopes, or by changing the mositure content and hence the cohesion of deposits in which they are developed. The former is illustrated by the extensive landslides that took place during construction of the Panama Canal, one of which involved 400,000 cubic metres (14,000,000 cubic feet) of material. Infiltration from gardens and cesspools allegedly triggered an extensive landslide in the Palos Verdes Hills of Los Angeles in 1956; in three years movement exceeded 20 metres (70 feet). The Vaiont Dam in Italy was destroyed when a landslide comprising over 240,000,000 cubic metres (8,500,000,000 cubic feet) slid into the reservoir; by raising the water table, the dam itself had contributed to slope instability. In areas with a permanently frozen subsoil (permafrost), thawing may be encouraged by the disturbance or removal of vegetation or superficial sediments and soils, since these act as insulators. Highway and railroad construction in Alaska has been accompanied by landslides, slumps, and rock flows, and road surfaces are distorted by frost heaving. The vibration produced by passing traffic may also trigger movement in unstable slopes.

Techniques have been developed to preserve or restore slope stability, for example by creating an insulating layer over frozen ground or by controlled drainage of terrain weakened by seepage. The Folkestone and Warren landslips of Kent, England have been controlled by drainage combined with coastal protection, because undercutting by wave action is an important contributory factor. In California, waterlogged clay slopes have been dried by circulating warm air through tunnels. The use of vegetation to bind the surface of cuttings and embankments is commonplace.

Erosion and sedimentation. Any change in the geometry, water discharge, or sediment load of a stream will disturb existing equilibrium conditions, and is followed by readjustment that may involve erosion, deposition, or both. The silt that accumulates behind a dam often forms a delta; in the case of the Furnish Dam, Oregon, the entire basin was filled in 16 years. By trapping sediment and altering the pattern of flow, dams often lower the river bed in the valley below the dam. The channel of the Red River of Oklahoma and Texas has in this way become both wider and deeper for a distance of 160 kilometres (100 miles) below the Denison Dam. River straightening and dredging sometimes lead to further incision by accelerating the velocity of flow, as in the Rhine, which has deepened its bed at Duisburg by two metres (seven feet) since 1900. On the other hand, man may promote deposition by introducing surplus sediment, as in the Sierra Nevada of California and other areas subject to hydraulic mining.

The incidental effects of stream modification extend to lakes and coasts, whether as changes in the water balance or in terms of the sediment supplied. The progressive fall that has characterized the Aral and Caspian seas in recent years is attributed largely to dam construction on their affluents (inflowing streams); it has exposed 38,000 square kilometres (15,000 square miles) of land in the former and 25,000 square kilometres (10,000 square miles) in the latter. The area of San Francisco Bay has contracted by over 700 hectares (1,700 acres) since 1850, largely because of the import of sediment by rivers affected by hydraulic mining of placer gold; in contrast, Ventura Beach, California, has been deprived of much of its sand supply by the damming of the Ventura River.

Urbanization and hydrologic change

Urbanization is a further source of hydrological change. By paving the Earth, man armours it against weathering and erosion; he also renders it impermeable and thus produces a far higher proportion of runoff (as opposed to infiltration) than would otherwise prevail. A little less than one percent of the world is occupied by buildings, roads, and kindred structures, but their uneven distribution is brought out by the corresponding figures of 10 percent for England and Wales and of over 50 percent for Brooklyn, New York. Within these units there are large areas that are entirely covered. Measurements taken in the United States have shown that, during construction, building sites release sediment at rates up to 40,000 times greater than those of corresponding rural areas under grass or woodland; for urbanized and developing zones as a whole, the factor ranges between ten and 100. Channel flow gains little from groundwater and is therefore closely related to the incidence of precipitation; runoff is rapid, and the proportion lost to sewers is quickly restored to surface drainage. Hence, discharges are relatively higher and more short-lived than in areas where infiltration plays a prominent role in the hydrologic cycle and, in turn, contribute to stream load by producing channel enlargement. In brief, the progress of urbanization enhances stream activity, always providing that some erodible material has been left exposed to the action of water.

Like river basins, stretches of shoreline operate as physiographic units that adjust to man-made modification by changes in geometry or in process. Where breakwaters, jetties, or groins are erected to trap material, further erosion may take place along the coast if the beach depends for its equilibrium on the supply of sand or shingle by longshore movement of sediment. Much agricultural land has been lost in this way near Durban and Madras, and a breakwater erected to form a harbour at Santa Barbara, California, caused silting in the harbour and erosion along more than 15 kilometres (nine miles) of coast. The extraction of shingle can be equally deleterious by robbing the beach of its protection against wave action; it has aggravated erosion along the Holderness coast of England, whose susceptibility to erosion is indicated by the loss since Roman times of a strip averaging four kilometres (two miles) in width.

At times man intervenes to restrain erosion, as when he fixes coastal dunes with the help of vegetation, or seeks to establish or restore a balance between coastal erosion and deposition. In England, trucks have been employed to carry shingle from the beaches of Rye to those from which it had been derived at Winchelsea, and a similar function is fulfilled by pumping sand across the erosive channel that leads to Palm Beach, Florida. Where the problem is to maintain or create a deep navigable channel, submarine erosive processes are turned to advantage by means of training walls or by judicious dredging. Liverpool Docks and the New York Harbor both benefit from this stratagem.

Soil erosion

Man's indirect influence is displayed most prominently in the varied physiographic changes that are grouped under the heading of soil erosion and that affect two-thirds of the Earth's land surface. In the United States, for example, 5 percent of the land suitable for cultivation or forestry has been entirely ruined, 15 percent has lost three-quarters of its topsoil, and 41 percent is subject to moderate erosion. In addition to its local implications, soil erosion affects other landscape elements and the performance of allied processes. The incidence of violence of flooding is increased by rapid runoff and low infiltration; stream channels are choked by silt; and material removed by wind erosion accumulates in drifts.

Many factors help to determine the stability of soils, and the scope for human intervention is correspondingly large. Differences in the character and density of the

vegetation cover generally account in large measure for areal variations in the severity of soil erosion; hence, the importance of fire and burning, as previously discussed. In the United States it has been found that land used for row cropping (for instance, under cotton or tobacco) may erode at rates that are 100 times greater than for woodland or pasture; in Cyprus, the sediment yields from vineyards are more than six times those on comparable terrain under forest cover. In either case, slight modification of the vegetation cover will produce disproportionate changes in the rate of soil loss.

Soil structure provides further opportunities for inadvertent triggering or acceleration of erosional activity. Vegetation and plant litter protect the soil against the impact of raindrops and the action of running water and wind. A soil structure that is resistant to these forces may be destroyed by compaction, whether by hooves or vehicle tires. Paving can give rise to new water-courses as well as favouring the activity of those already in existence. In many badly eroded areas, gullies are fed by runoff from roads that have not been provided with adequate gutters. Once linear erosion is initiated, soil moisture is locally depleted; the benefits of good soil structure are overridden by downcutting and bank collapse.

Unlike vegetative cover and soil structure, whose reinstatement is hampered and ultimately precluded by erosion, the factor of topography is amenable to correction; hence the emphasis on terracing and contour plowing in many soil conservation schemes.

Climate and air pollution. Rainfall intensity and the duration of individual storms are important factors influencing flood flow and the protective value of the vegetative mat, particularly in semiarid areas. Although large-scale weather modification (*q.v.*) by man is still in its exploratory stages, there is little doubt about man's capacity to stimulate precipitation by seeding suitable clouds with condensation nuclei. Hail suppression by this means is already practiced, and the mitigation of destructive downpours appears feasible. The climatic changes associated with urbanization are not wholly irrelevant to erosion, because they may involve increases in the total annual precipitation of up to 18 percent, a value that will be enhanced by the impermeability of the terrain on which it falls.

Air pollution, which is an important element of urban climates (*q.v.*), may be influential in rural areas if it is virulent enough to weaken the vegetation cover. The emission of sulphur dioxide by the copper smelters of Ducktown, Tennessee, has had this effect and thereby led to severe gullying; a similar chain of events has been reconstructed for the Swansea Valley of Wales in the 18th and 19th centuries. It is also suggested that overgrazing in parts of the Rājasthān desert of India has increased the local dust cover and that this, in turn, leads to greater aridity and to desert formation.

Air pollution is suspected of influencing world climate as a whole. The carbon dioxide content of the atmosphere has increased by an estimated 11 percent over its value in 1850, and the associated warming (greenhouse effect) is put at 0.6 percent between 1880 and 1940. It was formerly held that the combustion of fossil fuels was the primary source of the additional carbon dioxide; the oxidation of humus, peat, and other organic material is now regarded as a major contributor, but man retains some responsibility because the process is encouraged by forest clearing, cultivation, and the drainage of bogs. At the same time, the addition of particulate matter to the atmosphere by mankind as a whole is believed to have had a cooling effect amounting to about 0.05° C (0.09° F) each decade. In brief, the extent of human influence on long-term temperature changes remains uncertain, but the question deserves to be pursued in view of the implications such changes have for the vegetation cover and the pattern of precipitation.

Earthquakes. Forces acting within the Earth are not immune from the influence of man. It has been observed that the incidence of earthquake shocks has increased near the sites of large reservoirs, among them lakes Mead, U.S.; Kariba, Rhodesia; Koyna, India; and the Arapuni River, New Zealand. On the other hand, there is some evidence to suggest that Lake Mead has reduced earthquake activity, as the seasonal variations in its level during 1939–51 show some correlation with the number of shocks within a given radius. Underground nuclear explosions are followed by aftershocks, and some believe that they may trigger impending earthquakes in the vicinity. This has given rise to proposals for the controlled release of strains within the crust by means of explosions, in order to avert major earthquakes. The finding that the leakage of fluids into fault zones appears to affect their activity has similarly encouraged attempts to lubricate lines of crustal fracture subject to intermittent but large-scale displacement.

Interaction of human and natural forces. Other environmental changes that have been attributed wholly or in part to human agency can be explained in different ways. Many of the erosional features that are regarded as characteristic of accelerated erosion are duplicated in the geological record for periods apparently immune from human depredation. The sediment yields and erosional rhythms they represent are in some cases found greatly to exceed those now prevailing in the area. In New Mexico, for example, ancient arroyos appear to have formed at a rate that is more than nine times greater than those of the immediate prehistoric and historic period; and in some valleys in Iowa the deposition of alluvium in prehistoric times took place at over twice the modern rate. These findings are stimulating enough in challenging the view that a placid physiographic regime was disrupted by abuse of the land, although the possibility that abuse of the land is a long-standing attribute of man has already been raised. An alternative approach is to postulate changes in the climatic conditions governing physiographic processes. In the Mediterranean area, for example, erosional features often attributed to medieval deforestation and overgrazing can be interpreted in terms of a slight change in the seasonal incidence of rainfall. The onset of gullying in the southwestern United States is attributed by some to the arrival of white settlers armed with herds of cattle and metal plows, and by others to a climatic change that led to a weakening of the protective grass cover and perhaps also to a higher incidence of erosive floods.

In many such cases the answer probably lies in the combined effect of human and climatic influences. Man's potential for influencing other physiographic processes is understood; the difficulty lies in isolating his contribution to the landscape. There is great scope for work designed to reduce the area of doubt, if only because the evaluation of man's role is a prerequisite for effective control of deleterious physiographic change.

BIBLIOGRAPHY. GEORGE PERKINS MARSH, *Man and Nature* (1864), is the first general study on the subject, gloomy but not wholly pessimistic. R.L. SHERLOCK, *Man As a Geological Agent* (1922), deals with the British Isles in unusual detail. Physiographic change is one of the many topics treated in an international symposium on *Man's Role in Changing the Face of the Earth*, ed. by WILLIAM L. THOMAS, JR. (1956). ROBERT F. LEGGETT, *Geology and Engineering*, 2nd ed. (1962), considers many of the problems that confront the civil engineer. The use of air photographs in tracing prehistoric, classical, and medieval man-made features is illustrated in JOHN BRADFORD, *Ancient Landscapes* (1957).

(C.V.-F.)

Physiographic Effects of Tectonism

It has long been known that tectonism has profoundly influenced the physiography of the Earth. The Russian scientist M.V. Lomonosov stated in 1763 that the Earth's relief is formed by shaking of the Earth (earthquakes) or long-term sinking and raising of its surface. The German geographer Alexander von Humboldt suggested that mountain ridges were formed by upwelling of molten lava (volcanism) from the Earth's interior. This idea was derived from observation of volcanic activity in Central America. The French scientist Élie de Beaumont, however, thought that the Earth's relief features were principally attributable to gradual compression or shrinking of the Earth's crust because of slow cooling since its

formation. In this view, tangential pressures in the crust caused folding and general tectonic movements.

The Austrian geologist Eduard Suess noted that rock deformation was particularly intensive in mountainous areas, and he came to the conclusion that a close link existed between mountain ranges and crustal uplift and between oceanic and other basins and tectonic sinking or depression. In general, scientists today discriminate between tectonic movements that are geographically restricted and involved in mountain building (orogenic movements) and those movements that are broader, less localized, and associated with the uplift or depression of plateaus and basins (epeirogenic movements). Both types of tectonism have occurred throughout geological time and are responsible for the Earth's present relief features or physiography. Modern investigations have led to the suggestion that all tectonism is produced by the movement of great blocks of the Earth's crust relative to each other. This theory, called plate tectonics, is described in the articles SEA-FLOOR SPREADING; OCEANIC RIDGES; CONTINENTAL DRIFT; and MOUNTAIN-BUILDING PROCESSES.

This article deals with the effects of tectonism on the Earth's surface features, including the ocean floor. For further information on the several theories of formation of these features or landforms, see LANDFORM EVOLUTION.

Recent tectonic movements. *Seismic, or rapid, movements.* Direct observations and theoretical generalizations suggest a close connection between earthquakes, which periodically occur in different parts of the globe, and recent tectonic movements. Such movements are considered to be mechanical dislocations or faults within the crust. When faulting occurs, earthquakes are generated. The detection, recording, and analysis of the resulting seismic waves provides information on the intensity of an earthquake and its location—both where it occurred on the surface of the Earth and at what depth.

Tectonic movements of a seismic character are often called rapid because the physical consequences occur swiftly. The reasons for these tectonic movements cannot be directly observed because they occur at depths of hundreds of kilometres, but strong earthquakes generally manifest themselves in repeated disturbances of the Earth's crust through time and thus produce discernible physiographic effects at the surface.

Rapid tectonic movements frequently occur in the Soviet Union, where intensive earthquakes of destructive character take place several times every year; less intensive earthquakes are even more common. The earthquake epicentres are geographically localized and this provides the opportunity to carry out regional seismic investigations and to forecast the locations and the intensity of possible earthquakes. The greatest frequency of intensive, rapid tectonic movements has been established as characteristic of mountain regions—from North Africa, the Alps, Balkans, and Caucasus to the Tien Shan, Pamirs, and the Himalayas and around the margins of the Pacific Ocean (Alaska, the Aleutians, Coast Ranges, Andes, mountains of New Zealand and Indonesia, and the Japanese islands, the Kurils, and Kamchatka). There are reasons to suppose that the main sources of rapid tectonic movements are the zones of deep faults that extend from the crust into the upper mantle (zone beneath the crust) of the Earth.

Mountain regions of intensive seismic activity

Slow tectonic movements. Slow tectonic movements of the Earth's surface were detected first in Scandinavia (the Baltic or Fennoscandian Shield, an area of ancient crystalline rocks) and subsequently in the Canadian Shield. In both regions dome-shaped, elevated, ancient shorelines and deposits of former lakes and seas were discovered. These phenomena have been explained in terms of a rather broad upward arching during the last 10,000 to 12,000 years because of constant melting of thick glacial ice cover and consequent rising caused by lightening of weight. The maximum uplift of Fennoscandia for this period was about 250–275 metres (820–900 feet) in the northern part of the Gulf of Bothnia, according to some authorities; uplift in the Hudson Bay region of Canada has attained nearly 300 metres (1,000 feet). Simple calculation shows that average rate of such tectonic movements reaches two to three centimetres (0.8 to 1.2 inches) per year in the tops of arches, gradually decreasing toward their peripheral regions.

Slow tectonic uplift and subsidence with such characteristics are common almost everywhere. These slow tectonic movements were detected by precise geodetic measurements, and their average velocities ranged from tenths of millimetres to several millimetres per year for plains areas and ten millimetres (0.4 inch) or more per year in mountain regions. Unfortunately, data are available for no more than 100 years. For this reason it is not yet possible to define the general trends (or periodicity) of such movements, as is possible in regions of postglacial isostatic uplifts, where there are records or available data. Nevertheless, the main conclusion about recent tectonic mobility, supported by seismic phenomena, may at present be considered proved by numerous geodetic, oceanographic, and geomorphological investigations.

Recent horizontal movements of separate parts of the Earth's surface also have been detected with the help of precise geodetic measurements and geomorphic observations. The features of horizontal displacement along the Talass-Fergana Fault in Soviet Central Asia or more complex movements of separate points along the San Andreas Fault in California are examples. In both regions single horizontal displacements of several metres have been observed immediately after strong earthquakes.

Recent tectonic mobility must be considered proved, although not all aspects are yet understood. The evidence consists of orogenic zones that coincide with seismic phenomena, shield areas in which isostatic movements occur, and regional lines of horizontal dislocations. It is not yet possible, however, to make an evaluation of tectonic mobility through time and, in particular, to define the periodicity, if any, that is manifested. Further measurements and observations are needed for a full understanding of these phenomena.

Effects on the Earth's surface. *Uplift and denudation.* Modern concepts of the origins of the Earth's surface features are based upon acceptance of the existence of a dynamic equilibrium between vertical tectonic movements (uplift and subsidence) and denudational processes, which encompass the erosion of rocks and accumulation of the resulting sedimentary debris.

Rates of continental denudation

The German geographer Albrecht Penck considered data on the volume of transported river sediments and concluded in 1894 that a layer 0.8 millimetre (0.03 inch) thick is annually transported to the sea by rivers. More recent calculations have shown that the solid sediment transported by rivers is alone sufficient to account for reduction of the Earth's surface by 0.09 millimetre (0.004 inch) per year. Other values, given by various scientists, range from 0.01 to 0.02 millimetre (0.0004 to 0.0008 inch) in lowlands to 0.6 to 0.8 millimetre (0.02 to 0.03 inch) per year in mountains. Values as great as 0.16 millimetre per year have been given for the United States.

It should be noted that the sediments transported by rivers make up only part of the general volume of all sediments that are subject to continental denudation. Some sediment is not carried to the sea at all but is retained within the area being eroded—on terraces, hillslopes, in local closed depressions, and elsewhere. This material, which is not transported to the sea, is not included in the calculations on which the foregoing data are based. Hence, values of continental denudation are greater than stated.

The general correlation between recent tectonic movements and continental denudation may be considered an indication of dynamic stability of the Earth's surface. The stability is characterized by the following process: slow tectonic uplift of certain areas stimulates erosion and denudation, whereas other parts of the Earth's surface undergo tectonic subsidence and the eroded material tends to accumulate in place, although part of it is transported to the sea in one form or another.

Data on the rate of development of a normal soil profile (sequence of vertical layers) under a natural vegetation cover are instructive in this regard. The use of radiomet-

ric measurements for defining the absolute age of humus in the recent soils has shown that a time period ranging from several hundred to 1,000 to 3,000 years (depending on the genetic type of a soil) is required for the formation of a normally developed soil. This means that the natural "growth" of soil thickness ranges from one to ten millimetres per year. It shows that under the conditions of natural denudation, a well-regulated dynamic system exists. The three basic processes of the system—tectonic movements, soil formation, and continental denudation—tend essentially toward a state of balance. The middle component of the system, the soil-vegetation cover, serves as its main regulator.

Mudflows as an example of natural denudation

Equilibrium is lost when the natural soil-vegetation cover is disturbed, whether by man's activities or natural causes. Rather widespread phenomena of natural denudation are evident in mudflows, which form in mountains during rainstorms when there is an exceedingly high surface runoff. They are characteristic of many high mountain zones, where thick accumulations of loose sediments are present, and in arid regions, where the vegetation cover is sparse. Catastrophic displacements of ground masses on slopes (avalanches, landslides, and mudflows) are quite common in these areas. They are usually caused by earthquakes; and many interruptions of the soil-vegetation cover on mountain slopes are characteristic elements of the natural landscapes and must be considered a reflection of disturbed equilibrium.

Relief features. The state of equilibrium between tectonic movements and the processes of denudation should provide for a stable, level character of the Earth's surface; it is clear, however, that such is not the case. Aside from extraordinary disturbance of the equilibrium by catastrophic movements of sediment, more gradual tectonic movements and denudation are not balanced in the strict sense and may be progressively altered and directed over time. Consideration of this balance—or lack of balance—has led to the distinction among three main categories of elements in the structure of the Earth's surface. The largest category consists of first-order relief features, which are elements of the so-called geotecture. They include crystalline massifs, continental platforms, oceanic hollows, and large mountain systems.

Positive and negative elements of lesser magnitude make up the second-order relief features that complicate the surface of continents and ocean floors; these elements are called the morphostructure. This group includes plateaus, uplands, lowlands, ridges, and similar features. Morphostructures may be described chiefly as large forms of relief that emerge as a result of the interaction of tectonic and denudational forces. Tectonic movements play the leading role in this interaction.

Small elements of relief of the third order include such features as river valleys, lake basins, and dune or karst (cavern) forms. They are sometimes called elements of morphosculpture. The origin of these morphosculptural elements is largely dependent upon denudational processes.

Neotectonics as the basis of present-day relief

It has been suggested that all tectonic movements that are expressed in present-day relief should be called neotectonics. In papers by Soviet scientists this connotation emerges from considerations of the general amplitude and character of the most recent tectonic movements. A review map of these movements in the Soviet Union served as a model for international compilations of this type. The main basis for the work was study of the thickness of Pliocene–Quaternary deposits (younger than 7,000,000 years) that accumulated in areas of tectonic subsidence (depressions and basins) and of the amplitude of uplift of ancient surfaces in mountain ridges. In these instances, the obtained values of uplift and deposition amounted to hundreds of metres of relief over distances of several kilometres. The role of neotectonics in forming the present relief features is acknowledged by all authorities.

Tectonism of the ocean floor. It formerly was believed that the ocean floor was tectonically stable, but this idea was contradicted by the occurrence of tsunamis, large surface waves that are generated by submarine earthquakes. Because tsunamis were observed only on the coasts of the Pacific Ocean, they were thought to be related to the seismic focuses of the Pacific margin, sometimes called a ring of fire because of recurrent volcanism and tectonic movements. The most destructive tsunamis are now known to originate from earthquakes centred on the continental slope, however, where huge masses of sedimentary bottom deposits shift and slide toward lower levels.

The relief of the ocean floor is characterized by great complexity and diversity. Many landforms are similar to those on the continents, but some large and peculiar submarine geomorphic elements have no analogs on the continents. A high degree of tectonic mobility characterizes the ocean floor, particularly in the vicinity of oceanic ridges (*q.v.*). This mobility has produced complicated tectonic dislocations (transform faults), new forms of bottom relief, and the graded bedding so typical of underwater deposits resulting from density currents (*q.v.*).

The main elements of the ocean-floor relief consist of (1) underwater margins of continents, or the continental shelves; (2) transitional zones between continental shelves and the oceanic floor, namely, continental slopes; and (3) the ocean bed and midoceanic ridge systems. The latter form a single global system of mountain structures that continues through all the oceans. Each oceanic ridge is a complicated system exhibiting parallel crests with one or several groove valleys along their lengths. These structures coincide with a high degree of recent tectonic activity and volcanism and the development of a distinctive system of huge transverse breaks, which express themselves as large terraces, narrow troughs, and asymmetric ridges on the ocean floor. In a number of cases such breaks not only cross the ridges, dividing them into separate big blocks, but also extend over immense distances across the abyssal (deep sea) plains of the oceanic bed.

The unclear nature of the abyssal "plain"

The morphostructural character of the abyssal plains is less clear. The term plain, for example, is not fitting in many instances. The flattest elements of the oceanic floor occupy a very small part of the oceanic bed, not more than 10 percent. Moreover, the covering layer of bottom deposits often hides a complexly dissected underlying tectonic relief. Volcanic relief in the form of hills and seamounts occupies vast areas. A continuous generation of underwater relief probably occurs in the ancient hollows of the oceans. It is clear that if tectonic movements of the Earth's crust provide the basis of the main elements of land-surface relief, the tectonic factor must be given even greater priority when relief of the ocean floor is considered. At the great oceanic depths involved the forms of relief directly created by tectonics are probably preserved without significant change because of the absence of denudational processes and burial by bottom deposits. The relationship between recent tectonism and relief features of the Earth's surface is thus clearly evident on the ocean floor. Further study may lead to resolution of the gap in different approaches to the problem of the physiographic effects of recent and ancient tectonism.

BIBLIOGRAPHY. The following books give general but moderately technical treatment to the physiographic effects of tectonism: V.V. BELOUSOV, *Basic Problems in Geotectonics* (1962), is an English translation of a treatment of the fundamental laws and basic problems of general tectonics, with special reference to the Soviet Union and more general discussion of world tectonics. L.C. KING, *The Morphology of the Earth*, 2nd ed. (1967), provides a detailed discussion of the Earth's surface features and the fundamental forces affecting their development. The *Encyclopedia of Geomorphology*, ed. by R.W. FAIRBRIDGE (1968), includes articles on all major aspects of the physiographic effects of tectonism. For a more modern but somewhat technical account of all physical attributes of the Earth that are related to tectonism, see P.J. HART (ed.), *The Earth's Crust and Mantle* (1969).

(I.P.G.)

Physiology

The word physiology was first used by the Greeks around 600 BC to describe a philosophical inquiry into the nature of things. The use of the term with specific reference to vital activities of healthy humans, which began in the

16th century, also is applicable to many current aspects of physiology. In the 19th century, curiosity, medical necessity, and economic interest stimulated research concerning the physiology of all living organisms. Discoveries of unity of structure and functions common to all living things resulted in the development of the concept of general physiology, in which general principles and concepts applicable to all living things are sought. Since the mid-19th century, therefore, the word physiology has implied the utilization of experimental methods, as well as techniques and concepts of the physical sciences to investigate causes and mechanisms of the activities of all living things.

HISTORICAL BACKGROUND

The philosophical natural history that comprised the physiology of the Greeks has little in common with modern physiology. Many ideas important in the development of physiology, however, were formulated in the books of the Hippocratic school of medicine (before 350 BC), especially the humoral theory of disease in the treatise *De natura hominis* ("On the Nature of Man"). Other contributions were made by Aristotle (*Lykaion*, about 325 BC) and Galenos or Galen of Pergamon (*c.* AD 130-*c.* 200). Significant in the history of physiology was the teleology of Aristotle, who assumed that every part of the body is formed for a purpose and that function, therefore, can be deduced from structure. The work of Aristotle was the basis for Galen's *De usu partium* ("On the Use of Parts") and a source for many early misconceptions in physiology. The tidal concept of blood flow, the humoral theory of disease, and Aristotle's teleology, for example, led Galen into a basic misunderstanding of the movements of blood that was not corrected until William Harvey's work on blood circulation in the 17th century.

The publication in 1628 of Harvey's *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* (An Anatomical Dissertation Upon the Movement of the Heart and Blood in Animals) usually is identified as the beginning of modern experimental physiology. Harvey's study was based only on anatomical experiments; despite increased knowledge in physics and chemistry during the 17th century, physiology remained closely tied to anatomy and medicine. In 1747 in Berne, Switzerland, Albrecht von Haller, eminent as anatomist, physiologist, and botanist, published the first manual for physiology. Between 1757 and 1766 he published eight volumes entitled *Elementa physiologiae corporis humani* (*Elements of Human Physiology*); all were in Latin and characterized his definition of physiology as anatomy in motion. At the end of the 18th century, Antoine Lavoisier wrote about the physiological problems of respiration and the production of heat by animals in a series of memoirs that still serve as a foundation for understanding these subjects.

Progress in the 19th century

Physiology as a distinct discipline utilizing chemical, physical, and anatomical methods began to develop in the 19th century. Claude Bernard (1813–78) in France; Johannes Müller (1801–58), Justus von Liebig (1803–73), and Carl Ludwig (1816–95) in Germany; and Sir Michael Foster (1836–1907) in England may be numbered among the founders of physiology as it now is known. At the beginning of the 19th century, German physiology was under the influence of the romantic school of *Naturphilosophie*. In France, on the other hand, romantic elements were opposed by rational and skeptical viewpoints. Bernard's teacher, François Magendie (1783–1855), the pioneer of experimental physiology, was one of the first men to perform experiments on living animals. Both Müller and Bernard, however, recognized that the results of observations and experiments must be incorporated into a body of scientific knowledge, and that the theories of natural philosophers must be tested by experimentation. Many important ideas in physiology were investigated experimentally by Bernard, who also wrote books on the subject. He recognized cells as functional units of life and developed the concept of blood and body fluids as the internal environment (*milieu intérieur*) in which cells carry out their activities. This concept of physiological regulation of the internal environment occupies an important position in physiology and medicine; Bernard's work had a profound influence on succeeding generations of physiologists in France, Russia, Italy, England, and the U.S.

Müller's interests were anatomical and zoological, whereas Bernard's were chemical and medical, but both men sought a broad biological viewpoint in physiology rather than one limited to human functions. Although Müller did not perform many experiments, his textbook *Handbuch der Physiologie des Menschen für Vorlesungen* and his personal influence determined the course of animal biology in Germany during the 19th century.

Methods

It has been said that, if Müller provided the enthusiasm and Bernard the ideas for modern physiology, Carl Ludwig provided the methods. During his medical studies at the University of Marburg in Germany, Ludwig applied new ideas and methods of the physical sciences to physiology. In 1847 he invented the kymograph, a cylindrical drum that still is used to record muscular motion, changes in blood pressure, and other physiological phenomena. He also made significant contributions to the physiology of circulation and urine secretion. His textbook of physiology, published in two volumes in 1852 and 1856, was the first to stress physical instead of anatomical orientation in physiology. In 1869 at Leipzig, Ludwig founded the Physiological Institute (*neue physiologische Anstalt*), which served as a model for research institutes in medical schools all over the world. The chemical approach to physiological problems, developed first in France by Lavoisier, was expanded in Germany by Justus von Liebig, whose books on *Organic Chemistry and its Applications to Agriculture and Physiology* (1840) and *Animal Chemistry* (1842) created new areas of study both in medical physiology and agriculture. German schools devoted to the study of physiological chemistry evolved from Liebig's laboratory at Giessen.

The British tradition of physiology is distinct from that of the continental schools. Around 1869 Sir Michael Foster became Professor of Practical Physiology at University College in London, where he taught the first laboratory course ever offered as a regular part of instruction in medicine. The pattern Foster established still is followed in medical schools in Great Britain and the U.S. In 1870 Foster transferred his activities to Trinity College at Cambridge, England, and a postgraduate medical school emerged from his physiology laboratory there. Although Foster did not distinguish himself in research, his laboratory produced many of the leading physiologists of the late 19th century in Great Britain and the U.S. In 1877 Foster wrote a major book (*Textbook of Physiology*), which passed through seven editions and was translated into German, Italian, and Russian. He also published *Lectures on the History of Physiology* (1901). In 1876, partly in response to increased opposition in England to experimentation with animals, Foster was instrumental in founding the Physiological Society, the first organization of professional physiologists. In 1878, again due largely to Foster's activities, the *Journal of Physiology*, which was the first journal devoted exclusively to the publication of research results in physiology, was initiated.

Founding of the Physiological Society

Foster's teaching methods in physiology and a new evolutionary approach to zoology were transferred to the U.S. in 1876 by Henry Newell Martin, a professor of biology at Johns Hopkins University in Baltimore, Md. The American tradition drew also on the continental schools. S. Weir Mitchell, who studied under Claude Bernard, and Henry P. Bowditch, who worked with Carl Ludwig, joined Martin to organize the American Physiological Society in 1887, and in 1898 the society sponsored publication of the *American Journal of Physiology*. In 1868 Eduard Pflüger, professor at the Institute of Physiology at Bonn, founded the *Archiv für die gesammte Physiologie*, which became the most important journal of physiology in Germany.

Physiological chemistry followed a course partly independent of physiology. Müller and Liebig provided a stronger relationship between physical and chemical approaches to physiology in Germany than prevailed else-

where. Felix Hoppe-Seyler, who founded his *Zeitschrift für physiologische Chemie* in 1877, gave identity to the chemical approach to physiology. The American tradition in physiological chemistry initially followed that in Germany; in England, however, it developed from a Cambridge laboratory founded in 1898 to complement the physical approach started earlier by Foster.

SCOPE

Intradisciplinary nature. The anatomical and medical origins of physiology still are reflected in university courses and textbooks that concentrate on functional organ systems of animals; *e.g.*, frog, dog, cat, and rat. The trend in physiology, however, is to emphasize function rather than structure; *i.e.*, comprehensive functional specializations such as nutrition, transport, metabolism, and information have replaced earlier structural studies of organ systems. This trend can be explained in part by the fact that the analysis of an organ system typically involves studies at the levels of cells and molecules, and functional emphasis accommodates such studies better than the organ-system approach.

Functional and comparative emphasis

Early in the 20th century, the emphasis on cells as units of function resulted in a view that all physiology is essentially cell physiology and that all teaching therefore should pivot around the properties of cells. In later years successful analyses of cellular mechanisms involving synthesis, control, and inheritance led to similar emphasis at a new and more fundamental level, the molecules that comprise cells. The study of physiology now encompasses molecules, cells, organs, and many types of animals, including man. The comparisons resulting from such studies not only strengthen human physiology but also generate new problems that extend into evolution and ecology. Much of the impetus for comparative physiology has resulted from the economic or medical importance to man of parasites, insects, and fishes.

Most of the physiology of micro-organisms and plants developed independently of animal physiology. The concept of comparative biochemistry provided the foundations for a physiology of micro-organisms that extended beyond parasitic forms of medical importance and resulted in recognition of the fundamental roles of micro-organisms in the biosphere. Botanists and agriculturists explore the physiology of higher plants, but fundamental differences in the modes of life of animals and plants leave little common ground above the molecular and cellular levels. In a little-known textbook, Claude Bernard stated that there is only one way to live, only one physiology of all living things. The goal of general physiology is to abstract this single physiology from the physiologies of all types of organisms. Although common or general features usually are found at the cellular and molecular levels of organization, multicellular structures also are studied. Processes that underlie cell function are emphasized in an approach based on analyses in terms of physical and chemical principles.

Topics of study. *Metabolism.* In the late 19th century the principle of conservation of energy was derived in part from observations that fermentation and muscle contraction are essentially problems in energetics. Biological energetics began with studies that established the basic equation of respiration as:

$$\text{Fuel} + \text{oxygen} \rightarrow \text{carbon dioxide} + \text{water} + \text{heat}.$$

It was realized that the heat produced in fermentation and the work performed during muscle contraction must originate in similar processes, and that fuel in the equation above is a source of potential energy. Early in the 20th century studies of animal calorimetry verified these concepts in man and other animals. Calorimetry studies showed that the energy produced by the metabolism of foodstuffs in an animal equals that produced by the combustion of these foodstuffs outside the body. After these studies, measurement of the basal metabolic rate (BMR) was used in the diagnosis of certain diseases, and data relating the composition of foodstuffs to their value as sources of metabolic energy were obtained.

Early in the 20th century it was established that measurable amounts of the carbohydrate glycogen are converted to lactic acid in frog muscles contracting in the absence of oxygen. This observation and studies of alcoholic fermentation confirmed that the energy for fermentation or muscle contraction depends on a series of reactions now known as glycolysis. In order to show that the conversion of glycogen to lactic acid could provide the necessary energy for muscular contraction, extremely delicate measurements of the heat produced by contracting muscles were required. As a result of glycolysis studies, adenosine triphosphate (ATP) was recognized as an important molecule in cellular energy transfer and utilization; *e.g.*, movement, generation of electricity, transport of materials across cell membranes, and production of light by cells. Soon it was discovered that a muscle protein called myosin acts as an enzyme (organic catalyst) by liberating the energy stored in ATP and that ATP in turn can modify the physical properties of myosin molecules. In addition, it was shown that a muscle fibre has an elaborate and ordered structure, which is based on a precise arrangement of myosin and another muscle protein called actin.

Glycolysis is an anaerobic process (*i.e.*, it does not require oxygen) and may represent one of the oldest mechanisms for cellular energy transfer, since the process could have evolved before there was free oxygen in the earth's atmosphere. Most cells, however, derive their energy from a series of reactions involving oxygen and called the Krebs tricarboxylic acid cycle. The enzymes for the cycle are part of the structure of a mitochondrion, which is an elaborate cellular component filled with membranes and often shaped like a very small bean. In the course of the oxidation, three molecules of energy-rich ATP are generated for each oxygen atom used to form a molecule of water. The mitochondrion, therefore, is the cellular site of respiratory combustion first clearly demonstrated in whole animals by Lavoisier.

The ultimate source of foodstuffs used by animals is plants. Early 19th-century studies of photosynthesis were closely related to those of respiration, and began with Joseph Priestley's demonstration that plants could restore the air used during respiration or combustion. The most important equations for living things therefore, are mutually inverse. In respiration:

$$\underset{\text{carbo-hydrate}}{(CH_2O)_n} + \underset{\text{oxygen}}{nO_2} \rightarrow \underset{\text{carbon dioxide}}{nCO_2} + \underset{\text{water}}{nH_2O} + \text{heat}.$$

In photosynthesis:

$$nCO_2 + nH_2O + \text{light} \rightarrow (CH_2O)_n + nO_2.$$

In the 1930s, it was shown that photosynthesis involves splitting hydrogen from water and that the oxygen liberated in photosynthesis comes from water. During the light reactions, light energy is captured by a green pigment called chlorophyll and used to generate reactive hydrogen and ATP that are used during dark reactions in which carbohydrates and other cell constituents are synthesized.

Integration of classical, chemical, and comparative approaches

The classical fields of organ-system physiology have a role subsidiary to that of cellular metabolism. Feeding and digestion, for example, become a means for the enzyme-catalyzed breakdown of organic compounds into relatively small molecules that can be transported readily; nutrition, therefore, is a way to supply animals with sufficient sources of energy and specific substances that they cannot synthesize. Comparative animal studies, which were of practical importance in the discovery of some vitamins, led also to the general observation that the specific nutrient requirements of animals are consequences of a slow evolutionary deterioration in which synthetic abilities are lost through changes or mutations in hereditary material.

Nutrition and digestion, however, also have been important in obtaining information at the cellular and molecular levels. It was through studies of digestion, for example, that the existence and nature of enzymes were first disclosed clearly. In addition, early recognition of similarities between digestion and fermentation foreshadowed knowledge of the important role of fermentation in cellu-

lar metabolism. Finally, the study of vitamin nutrition was closely integrated with that of cellular oxidation, in which certain vitamins play an essential catalytic role.

In intact organisms, the chemical activities of individual cells do not interfere with the functions of the organism. Much of the study of physiology now is concerned with the ways by which cells obtain their nutrients and dispose of their waste products. Knowledge of the mechanism of protein synthesis and its connections with inheritance and cellular control mechanisms have initiated new inquiries into functions at all levels; *i.e.*, cells, organs, and organisms.

Circulation of blood

Transport. Many important advances in surgery and medicine have been based on the physiology of circulation, which was first studied in 1628. The measurement of blood pressure, for example, was introduced on a practicable basis late in the 19th century and has become an important part of medical diagnosis. The physiology of circulation is concerned with the origin of blood pressure in the force of the heartbeat and the regulation of heart rate, blood pressure, and the flow of blood.

Variations in heart rate that led Aristotle to consider the heart as the seat of the emotions—a myth that persists even now—were among the phenomena whose explanation revealed the existence of the autonomic nervous system. Variations in heart rate are less important to the circulatory system, however, than is the ability of the heart to adjust the strength of its beat to meet certain demands of the body.

The peripheral control of blood pressure and blood flow depends upon a maze of interacting control mechanisms, most significant of which are direct control of the diameter of small arterial branches that enlarge or dilate in response to chemical products formed during metabolism. Increased metabolic activity of tissues such as muscles or the intestine, therefore, automatically induces increased blood flow through the dilated vessels. This action, which could result in a fall in blood pressure, is offset by central-reflex controls that constrict arterial branches not dilated as a result of local chemical effects. Certain regions of the skin and the intestines serve as reservoirs for blood that may be diverted to muscles or the brain if necessary. Peripheral control may break down if excessive demands are made upon it in hot weather (heat stroke), during vigorous exercise after meals (muscle cramp), and after extensive loss of blood or tissue damage (wound shock) or extreme emotion with consequent activation of the autonomic nervous system (emotional shock). A remarkable adaptation occurs in air-breathing vertebrates—reptiles, birds, and mammals—which dive for food or protection. During a dive, the flow of blood to all parts of the body except the brain and the heart is reduced substantially. The energy for muscle contraction is provided by the anaerobic process of glycolysis because the oxygen in the blood goes to the brain and heart, which cannot function without a constant supply of oxygen.

Comparative studies have disclosed two major patterns in circulatory systems. Among vertebrates and a few invertebrates—notably annelid worms and cephalopod mollusks—the blood flows entirely in closed channels or vessels, never coming into direct contact with cells and tissues; blood pressure and the velocity of flow are high and relatively constant, and the volume of blood is small. In many invertebrates—especially arthropods and mollusks other than cephalopods—the blood flows for part of its course in large sinuses or lacunae and comes directly into contact with the tissues. Blood pressure and the velocity of flow are low and variable in these invertebrates, and the large volume of blood is comparable to the total volume of all body fluids in vertebrates.

Consideration of the blood as a transport system has centred especially on the transport of oxygen and carbon dioxide. The colour of blood changes as it passes through the lungs; venous blood is dark purple and arterial blood is bright red because of the properties of a blood pigment called hemoglobin. The complete structure of hemoglobin now has been determined, and minute variations in this structure have enabled man to study fundamental questions of heredity at the molecular level. The development of blood-banks and the techniques involved in blood transfusions depend on knowledge of the physical, chemical, and biological properties of blood. These properties include a remarkable diversity of hemoglobin, both among individuals and species and also within an individual during development. In many instances variations in protein composition better adapt a species to its circumstances.

Transport through cell membranes

Studies of membrane transport at the cellular level are an important part of general physiology. Although quantitative theories of diffusion and osmosis that developed around 1900 were applied to cell physiology, a number of phenomena (*e.g.*, movement through membranes of certain ions and other compounds of biological importance) did not behave according to established physical principles. As a result of studies of osmotic and ionic regulation in freshwater animals, the concept of active transport was formulated. Crucial to the acceptance of this concept were studies with frog skin, which can transport sodium ions against chemical and electrical forces; the transport, specific for sodium ions, is dependent on a continuing input of metabolic energy. Efforts now are directed toward establishing a molecular mechanism that may involve an enzyme found in surface membranes of cells. This enzyme breaks down ATP and releases the energy in the molecule only if sodium and potassium ions are present.

Information transfer. The physiology of animals differs from that of plants in the rapid response of animals to stimuli. René Descartes, responsible for the concept of the reflex that dominated neurophysiology for most of its history, thought a sensory impulse was "reflected" from the brain to produce a reaction in muscles. Later studies of the effects of ions on nerves suggested that a nerve must be surrounded by a membrane and that a nerve impulse results from a change in the ability of the membrane to allow passage of potassium ions. When it was shown that nerves are made up of thousands of tiny fibres, which are processes that extend from cells located in the brain or spinal cord, the nerve impulse hypothesis was applied to individual nerve fibres rather than to whole nerves. Electronic technology provided the techniques and giant nerve fibres of squids provided the experimental material that enabled two Nobel prize winners for physiology, Alan Lloyd Hodgkin and Andrew Fielding Huxley, to extend this hypothesis into a theory of the excitation of nerve cells in which sodium ions and potassium ions play principal roles.

Reflex concept

The reflex concept, however, was not dependent on understanding the molecular basis of excitation, conduction, and transmission. Early in the 20th century the role of interaction of nervous centres in controlling muscle contractions was established. The reflex now is conceived as a unit in which nerve impulses initiated in sensory neurons or nerve cells are conducted to a centre in the brain or spinal cord. In the centre impulses initiated in motor neurons are conducted to muscles and induce a reflex response. Two processes can occur in the centre; one is associated with central excitatory states, the other with central inhibitory states. The net effect of any stimulus or group of stimuli, therefore, can be interpreted as an interaction of these opposing states in the centre.

After the demonstration that the effects of the vagus nerve in slowing the heart are mediated by a chemical substance, subsequently identified as acetylcholine, the concept of chemical transmission of nervous impulses was extended to the central nervous system. Typically, transmission of excitation from cell to cell is accomplished by the liberation of a chemical transmitter from a nerve ending.

The reflex concept gave rise to premature attempts to develop a psychology based on reflexes. These attempts (behaviourism) were advanced by the Russian I.P. Pavlov's discovery of conditioned responses. Originally known as conditioned reflexes, these responses have been found in most animals with central nervous systems. More complex than simple reflexes, their mechanism has not yet been established with certainty.

Sensory functions at cellular level

The analysis of sensory functions also extends to the cellular level. Sense organs are diverse in structure and sensitivity to specific stimuli. It may be that the common molecular basis for the differences in sensitivity is a change in permeability of a special region of the membrane surrounding a sensory cell. This change in permeability could allow a nerve fibre to become excited and initiate a nerve impulse. Neurophysiology has borrowed from, and contributed to, the information theory used in communications engineering. The function of sense organs is to gather information both from the environment and the organism. The central nervous system integrates this information and translates it into a program of response involving the entire organism. In addition, the brain can store information previously received (memory) and has the ability to initiate actions without obvious external stimulation (spontaneity). Some aspects of memory and integrative function have been modelled in electronic computers; in fact the development of computers was closely connected with the development of ideas about the functions of the central nervous system.

The analytical interpretation of central nervous function remains, however, a complex and difficult field, even though recent progress has brought closer together the study of behaviour in terms of nerve function and behavioral models. Considerable effort now is directed to the localization of brain function. Although specific centres for reception of sensory information and integration of motor programs are known, the integrative functions that tie them together, as well as the functions of memory, are not so well established.

Regulation. The concept of internal regulations is attributed to Claude Bernard, who thought of blood as an internal environment in which cells function; according to Bernard, maintenance of the internal environment at a constant level was a major responsibility of all body functions. Bernard showed in studies of the formation and breakdown of glycogen in the liver that internal organs can secrete materials into the blood. Other investigators demonstrated such a secretion and used the word hormone to describe the substance. One classical study concerned control of the secretion of digestive fluids by the pancreas; an active substance secretin was purified, as have been a number of similar materials from the digestive tract. The field of endocrinology now is a major part of physiology.

Complementary roles of hormones and nerves

The endocrine system complements the nervous system in control and coordination. Hormones, liberated into blood and other body fluids by endocrine glands and transported throughout the body, usually act either on specific target organs or on certain activities of many organs. Nervous coordination most often is concerned with rapid responses of short duration; endocrine coordination, however, usually is involved in slower responses of longer duration. Stationary-state regulation or homeostasis depends on the action of hormones at many points. The hormones insulin and glucagon, both formed in specialized endocrine tissue in the pancreas, control the level of sugar in blood. Vasopressin from the pituitary gland at the base of the brain and aldosterone from the adrenal glands near the kidneys control salt and water balance of the blood. Hormonal regulation, however, is not confined to homeostasis. The cyclic events of the female reproductive cycles in mammals, for example, are determined by a complex sequence of endocrine interactions involving hormones of the pituitary gland and the ovary.

The pervasive regulatory action of hormones is part of a large system of interactions to which the term feedback generally is applied. Hormones involved in homeostatic regulation, for example, influence their own secretion. The secretion of certain steroid hormones, which have a significant action on the conversion of amino acids to glycogen, is controlled by another hormone called the adrenocorticotropic hormone (ACTH), which is formed in the anterior pituitary gland. In turn the secretion of ACTH is controlled by a releasing factor formed in the midbrain and liberated from the stalk of the pituitary gland. ACTH liberation normally is controlled by the concentration of steroids in the blood, so that an increase in steroid concentration inhibits ACTH secretion; this negative feedback, however, may be overcome in certain conditions of intense nervous stimulation.

A similar pattern of releasing factors, by which the nervous system interacts with the endocrine system, also is known for other anterior pituitary hormones; *e.g.*, those involved in the reproductive cycle and in responses of the thyroid gland to temperature changes. In addition, neurosecretory cells—nerve cells specialized for endocrine function—liberate hormones (*e.g.*, vasopressin) that act directly on a specific target. Comparative studies show that neurosecretory cells are important in developmental and regulatory functions of most animals. Discrete endocrine glands, however, occur less frequently; in insects and crustaceans, cycles of growth, molting (shedding of the cuticle), and development are controlled by hormones. The identification of insect hormones may be useful in controlling pests through specific interference with processes of growth and development.

SOCIETIES AND PUBLICATIONS

The British and American societies of physiologists founded late in the 19th century were followed by similar organizations in other countries. These societies function primarily as organs of communication among physiologists, through the publication of scholarly journals and through meetings at which the results of research are presented and discussed. The need for international meetings was recognized in 1881, when the first International Congress of Medicine was held in England, and the Physiological Society invited foreign physiologists. In 1889 the first International Congress of Physiology was held in Basel; since then international congresses have met in various countries, usually every three years. Organization of an international congress and promotion of international cooperation among physiologists have become responsibilities of the International Union of Physiological Sciences, which was founded in 1953 and now includes societies in 40 countries.

Membership in physiological societies, long regarded as a distinction, usually is based on the publication of original scholarly research in physiology. The orientation of the societies is strongly medical, since most professional physiologists are associated with medical schools; consequently, physiologists with other interests have organized to represent fields of physiology less closely related to medicine. The Society of General Physiologists and the Biophysical Society in the U.S. are examples. French physiologists have not had quite so distinct an identity. In France and in the French cultural spheres of South America, North Africa, and Asia, the Société de Biologie has combined a variety of approaches to biology and medicine including physiology. In 1913, U.S. societies concerned with medically related fields such as physiology, pharmacology, biochemistry, and pathology joined to form the Federation of American Societies for Experimental Biology, so that meetings of related groups could be held simultaneously; societies concerned with nutrition and with immunology subsequently joined the Federation.

PROSPECTS

Physiology now is a mature science; during a century of growth, physiology has been the parent of numerous related disciplines, of which biochemistry, biophysics, general physiology, and molecular biology are the most vigorous. Physiology, however, retains an important position among the functional sciences closely related to medicine. Although many research areas, especially in mammalian physiology, have been fully exploited from a classical-organ and organ-system point of view, comparative studies in physiology may be expected to continue. Solution of the major unsolved problems of physiology will require technical and expensive research by teams of specialized investigators. Unsolved problems include the unravelling of the ultimate bases of the phenomena of life. Research in physiology also is aimed at integration of the varied activities of cells, tissues, and organs at the

level of the intact organism. Both analytical and integrative approaches uncover new problems that must be solved. In many instances, the solution is of practical value in medicine or helps to improve man's understanding of himself and other animals.

BIBLIOGRAPHY

History: K.J. FRANKLIN, *A Short History of Physiology*, 2nd ed. (1949), through the 19th century; J.F. FULTON (ed.), *Selected Readings in the History of Physiology* (1930), excerpts from classical publications; T.S. HALL, *Ideas of Life and Matter* (1969), on concepts of physiology and related fields; K.E. ROTHSCHUH, *Geschichte der Physiologie* (1953), a detailed account in German, with emphasis on German physiologists.

Classics: J. BARCROFT, *Features in the Architecture of Physiological Function* (1934), broad interpretations of physiology; W.B. CANNON, *The Wisdom of the Body* (1932), on regulatory physiology; A.V. HILL, *Adventures in Biophysics* (1931), studies on the biophysics of muscle; M. FOSTER, *Lectures on the History of Physiology During the 16th, 17th, and 18th Centuries* (1901, reissued 1970), a classic by one of the founders of modern physiology; C.S. SHERRINGTON, *Man on his Nature*, 2nd ed. (1951), a summary of and reflections on classical work in neurophysiology.

Textbooks and publications in specialized fields: E.F. ADOLPH, *Origins of Physiological Regulation* (1968), a broad survey of regulatory physiology; AMERICAN PHYSIOLOGICAL SOCIETY, *Handbook of Physiology* (1959–), a series of volumes summarizing in detail the status of research in specialized areas of physiology; H. DAVSON, *A Textbook of General Physiology*, 3rd ed. (1964), a revision of a classical English text by SIR WILLIAM BAYLISS; A.C. GIESE, *Cell Physiology*, 3rd ed. (1968), a standard text; B.T. SCHEER, (ed.), *Comparative Physiology* (1968), a collection of recent research papers.

(B.T.S.)

Piauí

Piauí is a state (*estado*) of northeastern Brazil, bordered by the states of Ceará, Pernambuco, and Bahia on the east and south and by Maranhão on the west and north. While it covers less than 3 percent of Brazil's total area, it is about equal in size to the Federal Republic of Germany, with 96,886 square miles (250,934 square kilometres). Its population early in the 1970s was over 1,700,000. The state capital is Teresina, located at the confluence of the Parnaíba and Poti rivers. The state has a small Atlantic coastline of about 40 miles.

Early settlers

The settlement of Piauí in the 17th century came with the expansion of cattle ranching in the backlands of the northeast, setting it apart from the sugarcane agriculture of the coastal lowlands. The settlers arrived from the east, moving up the valley of the Rio São Francisco in Pernambuco and onward in a westward direction into Piauí. One of their leaders was Francisco Dias d'Avilla, who fought bloody battles with the Indians. He and his lieutenants established ranches along the Parnaíba and Gurguéia rivers in the 1670s and 1680s.

The Rio Parnaíba runs along the western boundary of the state, linking the Atlantic port of Parnaíba with the inland cities of Teresina and Floriano. As one moves away from the river and the coastal area toward the south and east, the land rises gradually in a series of plateaus edged by cliffs. On the border of the State of Ceará, the plateau is broken by a gap through which the Rio Poti runs.

Temperatures show little variation, averaging about 79° F (26° C) in the northern part of the state and a few degrees less in some of the higher altitudes in the south. Annual rainfall is about 59 inches in the north, but, in the drier east and southeast, it averages about 20 inches a year. The dry months are in winter and spring.

The vegetation of the south and east is that of Brazil's semi-arid northeastern backlands: a thorny, deciduous scrub woodland known as *caatinga*. This gives way in the north to deciduous forests interspersed with areas of palms and cocoa trees.

Piauí is sparsely populated (about 18 persons per square mile). Most of its inhabitants are of mixed Indian and European ancestry, although a Negroid stock is found in the northeastern coastal zones, where the colonial sugar plantations once employed large numbers of African slaves. The largest city is Teresina, the capital, with an urban population of about 180,000 in 1970. Other towns include Parnaíba (70,000), near the Atlantic coast (with Luíz Correia as its port); Floriano, Campo Maior, Picos, and Oeiras. The principal economic activities are the raising of livestock and the production of palm seeds, palm oil, and palm wax from the carnauba and babassu palm trees.

Standards of living are among the lowest in Brazil. Infant mortality rates are very high, largely because of infectious or parasitic diseases. In 1969 there was one hospital bed for every 923 inhabitants. Hospital clinics reached about one-quarter of the population. The national government, however, was engaged in programs to control malaria and other endemic diseases.

Piauí had about 2,700 primary schools in 1969, with about 194,000 students. In the lower level secondary schools there were about 35,000 students. The state also had eight newspapers, nine radio stations, one theatre, and 13 cinemas.

The federal government was building a dam and hydroelectric plant on the Rio Parnaíba near Guadalupe. In 1970 the first part of the Boa Esperança project was completed, providing 108,000 kilowatts of power out of a projected 216,000 kilowatts.

Piauí's isolation

Piauí is a somewhat isolated state, lacking good communications even with neighbouring states of the Northeast. The Rio Parnaíba has not become the artery of commerce that geography suggests it should have. There is no railroad network in the backlands. Teresina, the capital, has rail connections across the State of Maranhão on the west to the port of São Luis; another railroad runs from Teresina northward through Campo Maior to the state's port of Luís Correia; but no railroad line crosses the state from east to west. One good all-weather federal highway runs from Teresina northeastward to Sobral in the State of Ceará. Of Piauí's 14,000 miles of roads at the start of the present decade, however, less than 200 miles were paved. The better roads were concentrated in the north, with Teresina as their hub, and in the central part of the state around Picos and Oeiras. Several important new highways were under construction by the federal government, including one from Teresina to Picos, another from Fortaleza in Ceará to Picos and southward toward Brasília and an east–west trans-Amazonian road that will pass through Picos.

BIBLIOGRAPHY. The few works dealing with Piauí in any depth are written in Portuguese. For its history, see R.N. MONTEIRO DE SANTANA, *Evolução Histórica da Economia Piauìense* (1964) and *Perspectiva Histórica do Piauí* (1965).

(M.A.de L.)

Picasso, Pablo

Unique for its volume (over 20,000 works), for its development in time (75 years of continuous creation), its diversity of style and technique—painting, drawing, engraving, and collage; sculpture and constructions; pottery and other mediums—and for its intensity, Picasso's *oeuvre* is both "autobiographical" (as an art dealer, Daniel Henry Kahnweiler, termed it) and politically committed. Of fundamental aesthetic importance, it launched, exploited, and perhaps surpassed the artistic revolution of the 20th century. Like the man himself, the work is unpredictable and complex; it lives through contrasts and disguises, though its vital essence is never hidden.

LIFE AND WORKS

Early years. Pablo Picasso, son of José Ruiz Blasco, a Castilian art teacher in Málaga (later in La Coruña and Barcelona), and of Maria Picasso, an Andalusian of Majorcan origin, was born in Málaga on October 25, 1881. At the age of 15, having brilliantly passed the entrance examination, he was admitted to La Lonja, the Barcelona school of fine arts, and in the following year he was accepted at the Real Academia de Bellas Artes de San Fernando in Madrid. But, impatient under formal training, he had scarcely shown his extraordinary precocity there before he reurned to Barcelona, where he frequent-



Picasso, 1931.
Cecil Beaton

ed the café Els Quatre Gats (the Four Cats), a meeting place for artists fashioned after the Chat Noir (Black Cat) in Paris. Although among this young Catalan avant-garde the influence of Munich and the Jugendstil (late-19th- and early-20th-century decorative style) outweighed that of Paris, it was to Paris that the young Picasso went in October 1900. There, though constantly near starvation, he became infatuated with the street life of Montmartre, Paris' bohemian district, and made many studies of the city's poor. More importantly, he discovered the posters of the 19th-century artist Henri de Toulouse-Lautrec, which inspired some of his pastels and a "Moulin de la Galette." Returning three months later to Spain, he co-founded in Madrid the short-lived magazine *Arte Joven* (first issue March 31, 1901; "Young Art"). On a second trip to Paris, in the summer of 1901, he exhibited his works at Ambroise Vollard's fashionable gallery in the rue Lafitte and became good friends with the avant-garde poet Max Jacob. He discovered Vincent van Gogh and, influenced by him, painted "The Absinthe Drinker" (1901, William Jaffe Collection, New York City) in strident colours and "The Dwarf Dancer." Suddenly, the 20-year-old painter, who now signed himself "Picasso," the name of his mother, moved toward a symbolism of pathos and misery inspired by the French painter Maurice Denis and the Spanish painters Isidro Nonell y Monturiol and El Greco. This was his Blue Period, so called because most of these paintings were dominated by various shades of blue, heightening their mood of isolation and despair. The "Evocation" ("The Burial of Casagemas," Musée d'Art Moderne de la Ville de Paris) of 1901 (Casagemas was a bohemian friend who had committed suicide because of a broken romance); the "Maternites" (1903, Museo de Arte Moderno, Barcelona); "The Tragedy" (1903, National Gallery of Art, Washington, D.C.); "Celestine" (1903, collection of artist); "La Vie"; an etching, "Frugal Repast," of 1904; "The Old Guitarist" (1903, Art Institute of Chicago); the "Woman with a Crow" (1904, Toledo Museum of Art, Ohio); and "Woman Ironing" (1904, Justin K. Thannhauser Collection, New York) display famished figures whose forms are extended vertically in immobile and tragic poses.

Blue Period

At first dividing these financially difficult years (1902–03) between Paris and Barcelona, Picasso, in April 1904, settled for good in Paris in the so-called "Bateau-Lavoir," a dilapidated building in Montmartre, where he soon met Fernande Olivier. She has described him at the age of 23:

> small, dark, stocky, restless, disquieting, with eyes somber, deep, piercing, strange, almost staring. Clumsy gestures, womanish hands, badly dressed, rather messy. Thick hair, black and lustrous, slashing across his intelligent and stubborn forehead. Half bohemian, half worker in his appearance. . . .

Picasso's happiness with his newfound mistress may have been reflected in his work. For soon after they met in 1904, he entered his Rose Period, when his works were filled with delicate pinks and the figures, while still somewhat melancholic, were not desolate as the subjects of the Blue Period paintings had been; figures became more robust, and family groups replaced the lonely prostitutes and beggars of earlier works. Picasso, who liked to attend the Médrano Circus with his friends, painted, according to the writer Guillaume Apollinaire, in a palette with the "flatness of frescoes" his "taciturn and androgynous harlequins": the "Family of Saltimbanques," "Girl on a Ball" (1905, State Pushkin Museum of Fine Arts, Moscow), and "Woman with a Fan," an unsmiling Gioconda who raises an enigmatic hand as though bidding farewell to the works of Picasso's youth. In the course of a stay at Gósol, in Spain, in the summer of 1906, he began to paint solid, subtly distorted female nudes at their toilettes, seen in "The Coiffure" (Metropolitan Museum of Art, New York), "Nude on Red Background" (Louvre, Paris), and "La Toilette." Picasso's increased concern with formal pictorial problems is seen in two celebrated portraits, "Gertrude Stein" and "Self-Portrait," in which all psychological interpretation is excluded in favour of a purely sculptural expression.

This rough affirmation of form led suddenly, between the end of 1906 and the spring of 1907, to "Les Demoiselles d'Avignon," a revolutionary and uncompleted work, aggressive and discordant in structure. The joint influence of the Postimpressionist painter Paul Cézanne and (in the central figures) of Greco-Iberian art is combined with that of African art (in the two figures on the right). The treatment of form in this key work of the 20th century represented such a rupture with traditional Impressionism that it was not understood at the time either by Picasso's fellow iconoclast painters André Derain and Georges Braque or by the anarchist critic Félix Fénéon. Only the youthful art dealer Daniel Henry Kahnweiler recognized the work's importance. He bought all of its preliminary studies, but Picasso refused to sell the painting itself. Instead, he rolled it up and put it away out of public view for several years.

Significance of "Les Demoiselles d'Avignon"

After a brief "Negro Period" ("Woman with a Fan," 1908; Hermitage, Leningrad), Picasso painted landscapes and still lifes (at La Rue-des-Bois, where he spent the summer of 1908) marked by the influence of Cézanne. It was simultaneously the "Demoiselles," these landscapes, and those of Braque that originated the style that, in 1908, led the critic Louis Vauxcelles, following Henri Matisse, to speak of "little cubes" and shortly thereafter of "cubism."

Picasso had met Matisse at Gertrude Stein's in 1906, and he was to keep up a friendly rivalry with the older man, so different from himself. His circle at the time included the poets Apollinaire, Max Jacob, André Salmon, Maurice Raynal, and Pierre Reverdy and the painters Derain, Kees van Dongen, Juan Gris, and Henri Rousseau.

Early Cubism

Friendship with Braque. In the winter of 1908, Picasso and Braque began to work along parallel courses—and in close friendship. In their search for an impersonal and universal pictorial language, they for a long time signed their canvases only on the back. From reality they strove above all to extract volumes and space, which they expressed by a play of planes and facets. After a stay at Horta de Ebro in the summer of 1909, Picasso reduced his palette to a subtle range of ochres and greenish grays ("The Reservoir," Gertrude Stein Collection, Paris; "The Factory at Horta," Moscow; "Nude in an Armchair," Douglas Cooper Collection, New York). Each of the four years that followed his return to Paris was to see a change in his method of constructing an object, and the method mattered more than the object itself, whether the latter be café table, guitar, violin, or woman with guitar. After the remarkable portraits "Ambroise Vollard" (1909–10, State Pushkin Museum of Fine Arts, Moscow), "Wilhelm Uhde" (1910, Roland Penrose Collec-

tion, London), and "Daniel Henry Kahnweiler" (1910, Art Institute of Chicago), Picasso came close to geometric abstraction in the course of the summer he spent at Cadaqués, in Catalonia, in 1910. The following summer at Céret (Roussillon), his forms crumbled more and more, and typographical characters appeared, calling the viewer's attention to the painted surface itself. In the same year, Picasso's works were exhibited at the Photo-Secession Gallery in New York.

In 1911, when Fernande Olivier left him, he met Marcelle Humbert, the mistress of the painter Louis Marcoussis. She soon moved in with Picasso, who fondly called her "Eva," signifying that she was the first woman in his affections, and began to paint her name into several still lifes ("Ma Jolie," 1911–12, Museum of Modern Art, New York City; "J'aime Eva," 1912, Columbus Gallery of Fine Arts, Ohio).

Use of Collage

Spending the summer of 1911 with Braque at Céret and Sorgues (Vaucluse), Picasso followed his friend's example in using the tecnhique of "pasted papers": commercial paper imitating wood or marble, newspapers, matchboxes, tobacco labels, and so on. This technique marked a transition from the Analytical Cubism of 1910–11, still linked to Cézanne, to the Synthetic Cubism of the years 1912–14, which replaced traditional, recognizable images of reality by signs whose raw sculptural effect increases the expressive while decreasing the representational value. Stressing the impact of these signs was the sculpture called "scraps constructions"—guitars and still lifes of rudely painted sheet iron or wood, for example. This "conceptual" (as opposed to purely visual) art contrasts with the refined empirical realism of classical Western tradition in order to rediscover the intense style of primitive art, but it is also "materialistic" long before the artist turned to abstraction and to mixing his pigments with sand. "When I want to paint a cup," Picasso said, "I will show you that it is round, but it may be that the general rhythm and construction of the picture will oblige me to show that roundness as a square." This formal coherence confers on the work an independent existence comparable to that of a poem.

But just at the moment when Picasso's Cubism could boast most clearly of colour and imagination (the scattering of dots and pieces of wallpaper of "rococo" Cubism), circumstances created a climate unfavourable for his work. The outbreak of war in 1914 dispersed the artists, separating him from his friends. Neither Picasso nor Braque, whose works could be seen at Kahnweiler's small gallery, had participated in group exhibitions of the "minor" Cubists (Albert Gleizes and Jean Metzinger, for example), who had made cubism into a fashionable dogma and whose superficial and composite works were then better known to the public.

Set designs for the ballet

Despite the climate of chauvinism produced by World War I, the young writer Jean Cocteau tried to bridge the gap between the painters of Montparnasse and the worldly cosmopolitan European intelligentsia. In 1917 he persuaded Picasso, depressed by the sudden death of Marcelle Humbert, to leave the Cubist milieu of Paris and travel to Rome, Naples, Pompeii, and Florence in order to collaborate with Sergey Diaghilev's Ballets Russes. Picasso designed the sets for the ballet *Parade*, with book by Cocteau and choreography by Léonide Massine, in which Cubist "stage managers" were brought together with the impudent music of Erik Satie. The opening-night audience hissed the ballet but applauded Picasso's painted curtain.

Picasso, who had lived in Montmartre and then in Montparnasse (1912–16) and Montrouge (1917), now moved to an elegant quarter. In 1918 he married Olga Koklova (whom he divorced in 1935); she was a dancer with the Ballets Russes and daughter of a Russian colonel. A social climber accustomed to luxury, Olga soon changed Picasso's bohemian life-style, using his ever-increasing wealth and prestige to gain entrance to Paris' most exclusive circles. They went to fashionable resorts: Biarritz (1918), where he decorated a rich villa, Saint-Raphael (1919), Juan-les-Pins (1920). Max Jacob called this his "duchess period." His work reflected a new tendency: henceforth, instead of displaying the austere unity of his Cubist period, Picasso's art moved simultaneously in opposite directions. From 1915 on, coexisting with the audacious still lifes of the Cubist years are figures of a subtly detached classicism that has been called "Ingrism" (after the 19th-century painter Jean-Auguste-Dominique Ingres): portraits of "Ambroise Vollard" (1915); "Max Jacob" (1915); "Jean Cocteau"; Igor Stravinsky" (1917); "André Derain"; "Sergey Diaghilev" (1919); "Olga in an Armchair" (1917); "Paul as Harlequen" (1924, collection of artist); and "Paul as Pierrot" (1925). He continued his collaboration with the Ballets Russes, on Manuel de Falla's *Three-Cornered Hat* (1919), Stravinsky's *Pulcinella* (1920), de Falla's *Cuadro Flamenco* (1921), and *Le Train bleu* by Cocteau and Darius Milhaud (1924). While dancers, harlequins, and pierrots invaded his work, gigantesque Neoclassical figures also appeared, in the painting "Three Women at the Spring" (1921, Museum of Modern Art, New York) and on the curtain for the ballet *Le Train bleu*. A masterly series of large still lifes before the window (1919; Saint-Raphael) carried Cubism up to his two versions of the "Three Musicians" (1921, Museum of Modern Art, New York; Philadelphia Museum of Art). Picasso then freed Cubism of all austerity, substituting the curve for the straight line ("Still Life with Antique Head," 1925; Musée National d'Art Moderne, Paris) or literally creating a style, as on the curtain for Erik Satie's ballet *Mercure* (1924; Musée National d'Art Moderne, Paris), where the Cubist experience is combined with the sinuous and somewhat morbid elegance of the postwar period. In 1923, in the magazine *The Arts*, Picasso declared: "I do not develop, I am . . . I can hardly understand the importance given to the word research. . . . to find is the thing."

Association with the Surrealists. The year 1925 marked a sudden break. Although he had rightly stated that he was not a Surrealist, Picasso took part in the first Surrealist exhibition (at Galerie Pierre Loeb, Paris). His large canvas "The Three Dancers" (1925, Tate Gallery, London) was said to be marked by "convulsive beauty," and he was to do a collage for the cover of the first issue of *Minotaure* (1933). He was also friendly with André Breton, the author of Surrealist manifestos; the Surrealist poet Paul Éluard; and the writer Michel Leiris. One may suppose that it was under the widespread influence of Surrealism that he created his famous "metamorphoses" in which he makes cruel sport of anatomy. The anxiety of the Surrealists, which was more intellectual than Existential, seems to have engendered in him a more personal and authentic anguish. The monstrously expressive "Bathers" of the "Dinard period" (1927–30) and certain cruelly violent female heads led to the "Seated Bather," a skeleton cut from some unreal mineral substance and deposited on a beach beyond the world. Later he painted "Bathers with a Toy Boat" (1937, Peggy Guggenheim Collection, New York), portraying almost-engaging female monsters whose arbitrary obesity is enveloped by a poetic eroticism. At the same time, his new mistress, the blonde Marie-Thérèse Walter (who in 1935 gave birth to their daughter Maia), inspired in him the tender interlude of "Nude Woman in a Red Armchair" (1932, Tate Gallery, London), "The Dream" (1932, Victor Ganz Collection, New York), and "Girl Before a Mirror" (1932, Museum of Modern Art, New York), brilliant canvases, all in luminous curves whose vegetal suppleness bespeaks the artist's emergence from a period of restlessness and anxious invention.

Interest in sculpture

After buying the château of Boisgeloup, near Gisors in Normandy (1930), Picasso, who had done no sculpture since "Woman's Head" of 1909 (if one excludes his Cubist "constructions" and the famous "Glass of Absinthe" of 1914), set to work in this fundamental area, soon mastering techniques that he borrowed or invented and producing such works as "Heads" in welded sheet iron, 1931; slender "women" shaped from the uprights of canvas stretchers, 1934; a large "Head of a Woman" in gilded bronze, 1932; "Woman with Foliage," 1934; and "Woman with Apple," 1934). "Sculpture," he has said,

"is the best commentary that a painter can make on his painting." At the same time, he exercised his imagination in the illustration of literary works, which represented a kind of holiday for his genius. He produced no fewer than 35 works, including Honoré de Balzac's *Chef-d'oeuvre inconnu* ("Unknown Masterpiece"), published by Vollard, 1931; Ovid's *Metamorphoses*, 1931; the 18th-century *Histoire naturelle*, by Georges-Louis Leclerc, comte de Buffon, 1942; and *Gongora*, 1948.

After a pilgrimage in 1934 to the Spanish sources of his inspiration—San Sebastián, Madrid, Toledo, and Barcelona—Picasso's darker emotions found expression in the themes of the bull and the minotaur, notably in his 1935 etching "Minotaurochy." "Picasso's *Minotaur*, who carouses, loves and suffers," writes Kahnweiler, "is Picasso himself." Thus, the story of his life blends more and more with his work. In 1936, he began to live with a beautiful Yugoslav named Dora Maar (Markovich), and his work took on the heightened emotional intensity inspired by the new liaison. Moreover, at this time, the Spanish Civil War, which had begun in 1936, aroused in him a political passion that led to his greatest masterpiece, "Guernica." He had just set up a new studio and there engraved "The Dream and Lie of Franco," when on April 28, 1937, the small Basque town of Guernica with its civilian population was destroyed by Hitler's bombers. As early as May 1, Picasso had produced the first sketch for "Guernica," a large composition commissioned for the pavilion of the Spanish Republic at the Paris World Exhibition of 1937, and by May 11 the whole immense canvas (11 feet, 6 inches by 25 feet, 8 inches [3.50 by 7.82 metres]) was outlined. If Picasso never explained this work of anger, violence, and death, this depiction of war through the tragedy of the bullfight, it is perhaps because its full meaning cannot be articulated; for, in one critic's words, it "overflows with fullness and presence, with signs and cries."

"Guernica"

Parallel to the series "Weeping Woman" (Roland Penrose Collection, London), Picasso created for the first time, in charming portraits of Dora Maar and Nush Éluard, the wife of the French poet, faces combining a frontal and profile view in order to avoid the fixed, immobile expression characteristic of the traditional portrait. On the eve of World War II, "Night Fishing at Antibes" emerged from the memory of a walk in the port, an insignificant event thus becoming the point of departure for a work of art.

World War II and postwar period. During the war years, Picasso was forbidden to exhibit his work after the German occupation of France; under the hardships of wartime, he experienced severe hunger and want for the first time since his early days in Paris. At Royan and then in Paris, Picasso's art became harsher and more sombre, distorting the human face more than ever, as in "Seated Women" of 1937–44, which, with its cruel and discordant colours, seems to be composed of enormous braided ropes or networks of cobwebs. Picasso reached the depths of despair in a large, ghastly nude "Woman Dressing Her Hair" (1940; collection of artist); in the "Still Life with Ox Skull" (1942, Dusseldorf, West Germany), painted after the death of his friend the sculptor Julio Gonzáles; and in "Aubade" (1942, Musée National d'Art Moderne, Paris), in which immobile monsters seem carved out in airless spaces without light. "If I am Picasso," he has said, "it is because I created monsters." But the anguish that underlies these works did not exclude either a vitality or a plastic coherence that places other works, such as the famous "Bull's Head" made from the handlebars and seat of a bicycle, very far from the detached dream world of Surrealism. Turning away from the funereal "Skull" of 1941, Picasso in 1944 modelled the "Man with a Sheep," a figure of peace and hope. A bronze of this work has adorned a square in Vallauris since 1950.

Shortly after the liberation of Paris on August 25, 1944, Picasso joined the Communist Party and, also, for the first time exhibited at the Salon d'Automne. Opinion was divided, as it was to be later, in 1945, over a large Matisse–Picasso exhibition in London, which included "Guernica"; but in retrospect "Guernica" was to appear the very model of politically committed art. "One does not paint in order to decorate apartments. Painting is an instrument of offensive and defensive war against the enemy," Picasso declared in the enthusiasm of his political faith, while, as in the time of Cubism and "Guernica," retiring to spiritual solitude. Writers and artists such as Jean-Paul Sartre, Albert Camus, Éluard, Breton, Reverdy, Cocteau, Leiris, and the photographer Brassaï, Paris society, and foreign visitors invaded his studio in the rue des Grands Augustins, which was guarded, not too successfully, by his faithful friend Jaime Sabartés. But fame did not prevent this restless and tireless artist from continuing his work. A late riser, he worked far into the night under artificial light, observed a regimen of exercise, and received visitors, to whom he showed himself charming or aggressive, taciturn or ironical, but always fascinating, unpredictable, and elusive. Interested in all techniques, he resumed lithography in the workshop of Fernand Mourlot, executing 200 large lithographs between 1945 and 1949. After meeting Françoise Gilot, in 1946, he painted, during the four months spent at the Musée Grimaldi in Antibes, perhaps his happiest works—the "Femme-fleur" (1946) and a large pastoral scene, "La Joie de vivre" ("The Joy of Life," 1946; Musée Picasso, Antibes). Fauns, pipe players, nymphs dancing or sleeping, centaurs—a whole Mediterranean mythology comes to life beside the landscapes of Vallauris and the happy scenes ("Maternity with Orange," 1951, collection of artist) in which his children Claude (born in 1947) and Paloma (1949) appear.

Political commitment

Beginning in the autumn of 1947, Picasso took up pottery making in the Madoura factory belonging to the Ramié family in Vallauris, and in 1948 he exhibited 150 pieces (representing only a tenth of his production) at the Maison de la Pensée Française in Paris. Under his fingers an amphora (an ancient Greek jar) becomes a woman or a condor (1947), a gazelle or a centaur (1948); plates are transformed into the faces of fauns or swarm with battling centaurs, bullfights (1948), doves (1949), and owls. "If the monster does nothing but smile," said Picasso in connection with the exhibition, "people are disappointed." In the same year he attended the World Peace Congress in Wrocław, Poland, and in 1949 the Peace Congress in Paris used his famous dove on its poster. Picasso gained international fame, and visitors from all over the world flocked to Vallauris. Since October 1948 he had been living at the villa called La Galloise, and there he painted "Massacre in Korea" (1951, collection of artist) and the portrait of "Hélène Parmelin," wife of the painter Édouard Pignon, whom he invited to come and live "a painter's life" near himself. He also painted "War and Peace" (1952), which consists of two large panels each measuring 15 feet, 5 inches by 33 feet, 7 inches (4.70 by 10.24 metres) and now adorning the Temple of Peace attached to an old chapel in Vallauris. These are politically motivated works, more colourful and narrative in spirit than "Guernica" and deliberately optimistic.

Pottery in Vallauris

All this activity did not prevent Picasso from taking a daily swim with his family at Golfe Juan—or from renewing the language of sculpture. By using such materials or fragments as potsherds, basketwork, and bits of metal scraps, he created unexpected masterpieces: the "Goat," the "Pregnant Woman," "Woman with Perambulator" (1950), "Goat's Skull and Bottle" (1951), the "Crane" (1951–52, G. David Thompson Collection, Pittsburgh), and "Baboon and Young," which attains its powerful expression of the animal's motherhood by the superposition of two small toy automobiles to represent the head. "Angry Owl" bristles with old bent nails. No less effective in their construction are the "Women's Heads" of 1954, models for gigantic, multicoloured "signals" that 10 years later were to be executed in concrete by the Norwegian sculptor Carl Nesjar. Six "Bathers" (1956) and the "Female Bather Playing" (1958), assemblages of packing-case wood, display a geometric humanity of disconcerting humour, a kind of 20th-century "savage art."

Later life and further innovations. The year 1953 saw a break with Françoise Gilot, who left for the United States with their two children. Picasso, now 72, transposed the event into vengeful pictures on the theme of the ravisher minotaur and a series of 180 drawings (November 1953–February 1954) of incisive and dazzling virtuosity. These drawings, an indiscreet and ironic confession on the theme of the young woman and the old painter (or old clown), were a kind of spontaneous self-analysis that freed the artist from his obsessions. Elsewhere, in the "Romancero du Picador" (89 wash drawings, July 1959–June 1960), he depicts the doomed figure of the bullfighter passing from the mocking cries of the arena to an orgy with prostitutes, while a duenna hovers in the shadows.

Break with Françoise Gilot

After a charming interlude of 12 portraits of "Sylvette" (1954) in Vallauris, the painter was to produce still further surprises. Interior scenes, often enlivened by the presence of his new love Jacqueline Roque in a Turkish costume, reveal in effect "La Californie," the Belle Époque villa at Cannes that he bought in 1955. In 1958 he married Jacqueline and bought the austere château of Vauvenargues in the Sainte-Victoire chain of mountains near Aix. The location suggested to him the solemnity of Spain, but it also brought him back to subtle contrasts of vivid colours (obtained by the use of industrial lacquers, as in "Dining-Room at Vauvenargues," [1959–60, Galerie Louise Leiris, Paris]; and "Jacqueline de Vauvenargues" (collection of the artist).

Transpositions of museum masterpieces

A further innovation consisted in transforming masterpieces of the past into Picassos, "museum copies" that became original creations. He had already transposed some of El Greco's works, Nicolas Poussin's "Triumph of Pan" into "Bacchanal" (collection of artist), in 1944, and Gustave Courbet's "Les Demoiselles des bords de la Seine" into "Girls on the Banks of the Seine, a Paraphrase of Courbet" (1950, Kunstmuseum, Basel, Switzerland). In 1955 he composed 15 variations on Eugène Delacroix's "Women of Algiers," freely Cubist in a riot of oriental colours (the 15th version is in the Victor Ganz Collection, New York). "Las Meninas" ("Maids of Honour") by Diego Velázquez (of whom Picasso said, "Here is the true painter of reality!") was to serve as the theme for 44 variations executed in five months in 1957. The "Déjeuners," after Édouard Manet's "Déjeuner sur l'herbe" ("Luncheon on the Grass"), developed into a series of 27 paintings and 138 drawings (1960–61) and was completed at Mougins, where in 1961 Picasso settled in the rustic house Notre-Dame-de-Vie.

This violent and archaeological approach is meant to show, as Picasso explained, that "in art there is no past or future" and that a valid work is "capable of living always in the present." It is thus "in the present" that Picasso means to create, as when, in his 1958 mural for the UNESCO building in Paris, he treated the classical theme of the fall of Icarus in totally unexpected images. Similarly, he subjected traditional themes to an audaciously simplified technique in large canvases of 1963–65, such as "The Painter and His Model" (June 1963; Berggruen, Paris) and "Reclining Woman with Cat" (March 1964; two versions, Galerie Louise Leiris, Paris). For him this was a way of escaping all historical considerations and the tyranny of beauty in order to produce new and immediate pictorial forms.

In his pottery, on the other hand, Picasso continued to play with Mediterranean culture and antiquity: vase-women, the amphora with "Two Fauns Playing the Flute" (1950), and a fine series of owls and doves (1953). He makes profiles out of broken bricks, and, on Roman or clay tiles, he draws portraits of Jacqueline and of "musketeers" (or even "Balzac scenes" that suggest bourgeois photography). His sculpture from this time includes works in sheet iron, cut, bent, and painted to form the large woman with outspread arms, women's heads, and bearded men.

Late drawings and prints

Between the ages of 85 and 90 Picasso the prodigious draftsman produced three series of drawings of unprecedented freshness. They include mythological scenes, circus scenes, drinking and carousing scenes, and amorous scenes (82 drawings and washes, 1966–67) mingling Penelopes and Ledas, charioteers, and musketeers "*à l'espagnole.*" The same themes are treated more realistically in 347 untitled engravings (March–October 1968) showing Spanish silhouettes, nocturnal pursuits, abductions, procuresses, and audacious alcove scenes in which Pope Julius II witnesses the free and easy frolicking of Raphael and his Roman mistress. Here he had recourse with consumate artistry to all the techniques of engraving (etching and aquatint, drypoint, burin, and so on). A third series (1969–71) returns to drawing and, reintroducing a theme that had been absent since 1925, closes with the self-portraits as harlequin—these last follow the "Mousquetaires," paintings and large lithographs of 1969. In October 1971, Paris paid homage to Picasso on the occasion of his 90th birthday by displaying eight of his works in the Louvre, in the place formerly occupied by da Vinci's "Mona Lisa." He died on April 8, 1973, at Mougins, France.

Evaluation. The man was infinitely changeable, impossible to grasp. Gertrude Stein wrote in 1933 that, because he was Spanish, life for him was tragic, bitter, and painful. Others, however, have seen him as "simple and cheerful, so sure of himself," and had been struck by his extraordinary intelligence in conversation and manners, his sense of living, and the total absence of any "artistic" pose. For the critic Jean Cassou, he was above all an artist in the fullest sense of the word," a man exclusively possessed by the thing that is born and takes shape in his hands, the thing he invents or, as he says, finds." Faithful to "a fundamental and impatient temperament . . . virtuoso and actor . . . he imperturbably exercises his powers of metamorphosis." But, in his anxious desire for coherence, he could not rest until he had transformed this temperament into the strictness of form.

His work is literally invaded by man: first of all by the man himself, Picasso, through the many self-portraits—acknowledged or disguised—and mythical symbols: the melancholy harlequin, the minotaur hungry for the fruits of the Earth, the triumphant sculptor, and time and again the painter and his model; next by humanity, or rather by that image of mankind offered by the culture of the Mediterranean, which his work sometimes exalts and at other times opposes and destroys. Thus, it is not surprising that, apart from his austere Cubist phase, the human face is omnipresent and the female body constantly represented —naked, in abduction or rape, playing, sleeping, or dancing. Also evident are themes of man's struggles and leisure pastimes as well as death, culture, and art.

Anguished, complex, ironic, always intense, seldom serene, Picasso's work is the story of a life governed by art, sometimes exalting, sometimes overcoming an exceptional virtuosity. Devoid of illusions ("I have never tried to please the public"; "People don't buy my pictures, they buy my signature"), Picasso had entered into an unacknowledged competition with his contemporaries (Braque, Matisse, Fernand Léger), with the great artists of the past (Velázquez, Ingres, Delacroix, Manet), and with himself: "Modern art should be killed. . . . That means that one should kill himself." Again, "In working, one must always be against. . . . If you know in advance what you're going to do, what good is that? It's better to do something else."

A realist and therefore a resolutely figurative artist, he had nevertheless renounced, since his Cubist days, the allure of the imitative image, in order to invent new forms that live more intensely than reality itself. Nevertheless, no artist so relentlessly revolutionary has ever better expressed the nostalgia that the modern era feels for a past that it itself challenges. Picasso has been called a nihilist, an anarchist. On the contrary, one should weigh the energy, strength of affirmation, and paradoxical "humanity" of his work, in which the darkest anguish carries on a dialogue with a savage *joie de vivre* and the cruellest violence confronts an imperceptible smile.

MAJOR WORKS

PAINTING (EARLY YEARS): "Le Moulin de la Galette" (1900; Justin K. Thannhauser Collection, New York). (BLUE PERIOD): "The Mourners" (1901; Edward G. Robinson Collec-

tion, Beverly Hills, California); "The Dwarf Dancer" (1901; Museo de Arte Moderno, Barcelona); "The Blind Man's Meal" (1903; Metropolitan Museum of Art, New York); "La Vie" (1903; Cleveland Museum of Art). (SALTIMBANQUE AND ROSE PERIOD): "Woman in a Chemise" (*c.* 1905; Tate Gallery, London); "Woman with a Fan" (1905; Averell Harriman Collection, New York); "Family of Saltimbanques" (1905; National Gallery of Art, Washington, D.C.); "La Toilette" (1906; Albright-Knox Art Gallery, Buffalo, New York); "Self-Portrait with a Palette" (1906; Philadelphia Museum of Art); "Gertrude Stein" (1906; Metropolitan Museum of Art, New York). (THE NEGRO PERIOD): "Les Demoiselles d'Avignon" (1906–07; Museum of Modern Art, New York); "Fruit Dish" (1908–09; Museum of Modern Art, New York); "Bust of a Woman" (1909; Tate Gallery, London). (ANALYTICAL CUBISM): "Seated Nude" (1909–10; Tate Gallery, London); "Girl with a Mandolin" (1910; private collection, New York). (SYNTHETIC CUBISM): "Accordionist (Pierrot)" (1911; Solomon R. Guggenheim Museum, New York); "Still Life with Chair Caning" (1911–12; collection of artist); "The Aficionado" (1912; Kunstmuseum, Basel, Switzerland); "Harlequin" (1913; Municipal Museum, The Hague); "Harlequin" (1915; Museum of Modern Art, New York); (CLASSICAL PERIOD): "Harlequin" (1917; Museo de Arte Moderno, Barcelona); "Guitar" (1919; Rijksmuseum Kröller-Müller, Otterlo, The Netherlands); "Two Seated Women" (1920; Walter P. Chrysler, Jr. Collection, New York); "The Rape" (1920; Museum of Modern Art, New York); "Three Musicians" (1921; Museum of Modern Art, New York); "Pipes of Pan" (1923; collection of artist); "Seated Women" (1923; Tate Gallery, London). (LATER WORKS): "Seated Bather" (1930; Museum of Modern Art, New York); "Guernica" (1937; Museum of Modern Art, New York); "Night Fishing at Antibes" (1939; Museum of Modern Art, New York); "Les Premiers Pas" ("First Steps," 1943; Yale University Art Gallery, New Haven, Connecticut); "The Charnel House" (1945; Walter P. Chrysler, Jr. Collection, New York); "Las Meninas," 58 canvasses (1957; collection of artist).

SCULPTURE: "The Jester" (bronze, 1905; Phillips Collection, Washington, D.C.); "Woman's Head" (bronze, *c.* 1909; Museum of Modern Art, New York); "Glass of Absinthe" (painted bronze and silver spoon, 1914; Museum of Modern Art, New York); "Still Life" (painted wood with upholstery fringe, 1914; private collection, London); "Guitar" (floorcloth, string, nails, and newspaper, 1926; collection of artist); "Sculpture" (iron wire, 1928; collection of artist); "Sculpture" (bronze, 1931; collection of artist); "Head of D.M." (bronze, 1941; collection of artist); "Bull's Head" (bicycle saddle and handlebars, 1943, collection of artist); "Man with a Sheep" (bronze, 1944); "Goat" (plaster, later cast in bronze, 1950; collection of artist); "Baboon with Young" (bronze, 1951); "Steel Sculpture" (1967; Civic Center Plaza, Chicago).

PRINTS AND DRAWINGS: "Salomé" (drypoint, 1905); "Max Jacob" (pencil, 1915; Mlle Dora Maar Collection); "The Bathers" (pencil, 1918; Fogg Art Museum, Cambridge, Massachusetts); "Portrait of Stravinsky" (pencil, 1920); "The Sculptor's Studio" (a series of 46 etchings, 1933–34); "Minotauromachia" (1935); and a series of minotaur etchings (1933–35); studies for "Guernica" (1937, in various media); "Self-Portrait" (pen and ink from "Desire Caught by the Tail," 1943); "Portrait of D.M. As a Bird" (pen, ink and wash, 1943); "Françoise" (lithograph, 1946); "Model and Monkey Painter" (ink and wash, 1954).

CERAMICS: Among many examples, "Woman Vase" (1948); "Three Doves" (1953); "Plate with Bullfight" (*c.* 1957).

THEATRE DESIGN: Decor for the ballets *Parade* (1917); *The Three-Cornered Hat* (1919); *Pulcinella* (1920); *Cuadro Flamenco* (1921); *Mercure* (1924).

BIBLIOGRAPHY

Biography and Criticism: ALFRED H. BARR (ed.), *Picasso: Fifty Years of His Art* (1946), a fundamental work; WILHELM BOECK and JAIME SABARTES, *Picasso* (1955), with 596 illustrations; MAURICE RAYNAL, *Picasso* (1953), for the period to 1950; PIERRE DUFOUR, *Picasso 1950–1968* (1969, in French) —all the above contain good bibliographies. Two critical biographies are ROLAND PENROSE, *Portrait of Picasso* (1956; reprinted as *Picasso*, 1971); and ANTONINA VALLENTIN, *Pablo Picasso* (1957), in French.

Catalogs: CHRISTIAN ZERVOS, *Pablo Picasso*, 23 vol. (1932–71), a basic work reproducing all the works in black and white; BERNHARD GEISER, *Picasso, peintre-graveur*, 2 vol. (1968); GEORGES BLOCH, *Pablo Picasso: Catalogue de l'oeuvre gravé et lithographié, 1904–1967* (1968); D.H. KAHNWEILER, *Les Sculptures de Picasso* (1948; Eng. trans., *The Sculptures of Picasso*, 1949); DOUGLAS COOPER, *Picasso Théâtre* (1967; Eng. trans., 1968); MAURICE JARDOT, *Pablo Picasso: Drawings* (1959); *Hommage à Pablo Picasso*, introd. by JEAN LEYMARIE (1966).

Contemporary accounts: GUILLAUME APPOLLINAIRE, *Les Peintres cubistes* (1913; Eng. trans., 1944); JEAN COCTEAU, *Le Rappel à l'ordre* (1926); D.H. KAHNWEILER, *Der Weg Zum Kubismus* (1920; Eng. trans., *The Rise of Cubism*, 1949); FERNANDE OLIVIER, *Picasso et ses amis* (1933); JAIME SABARTES, *Picasso: Portraits et souvenirs* (1946); GERTRUDE STEIN, *The Autobiography of Alice B. Toklas* (1933) and *Pablo Picasso* (1938); FRANCOISE GILOT and CARLTON LAKE, *Life with Picasso* (1964); ANDRE BRETON, *Le Surréalisme et la peinture* (1965); MARIO DE MICHELI, *Scritti di Picasso* (1964); HENRI CLOUZOT, *Le Mystére Picasso* (1955), a film.

(P.Du.)

Piccard Family

A family of Swiss scholars, the Piccards have gained international renown for their original contributions to science and public affairs. Their most illustrious members were the twins Auguste and Jean, who explored both the upper stratosphere and the depths of the sea in vehicles of their own design.

Horst Tappe—EB Inc.



Jacques, 1967 (left), and Auguste Piccard, 1961.

Rodolphe (1807–88) was the first to inaugurate the family's international vocation by obtaining an appointment as painter to the court of Russia, in St. Petersburg (now Leningrad) in the 19th century. His brother, Jules-François, began the family's scientific orientation by preparing many plates for the official map of Switzerland. Two of Jules-François's sons followed in his footsteps: Jules was a professor of chemistry at the University of Basel; and Paul, an engineer from Geneva, designed and sold the first electrical turbine at Niagara Falls.

Jule's sons Auguste and Jean, identical twins, were born in Basel on January 28, 1884. They enrolled together at the Swiss Federal Institute of Technology, in Zürich, where they followed parallel careers, one in physics, the other in chemistry. When they became doctors of science, both decided to teach in universities; Jean, the chemist, going first to Munich, then to Lausanne, then to the United States; and Auguste, the physicist, staying on at the Institute. Both married foreign students: Jean married Jeannette Ridlon, an American student of his at Chicago, in 1919, and Auguste married the daughter of a French historian at the Sorbonne.

Auguste became interested in balloon ascents, not as a sport but as a means of making experiments. Oriented toward pure physics, he participated in many important research studies. When the University of Brussels in 1922 created a chair for applied physics, Auguste, who was also a mechanic and an engineer, readily accepted the post. Maintaining the youthful zest for adventure that had been inspired by his studies, Auguste, who later explored both the sky and the depths of the ocean, amply fulfilled his dreams, pursuing them with the will of an adult and the rigour of science. Having studied cosmic rays, he conceived of an experiment for observing them at 16,000 or more metres (over 50,000 feet) in the upper atmosphere. Previous ascents had shown that the stratosphere could be fatal and that to penetrate the iso-

Flights to the stratosphere

thermal layer, with its low pressure, a revolutionary balloon would be necessary. He built such a balloon in 1930, with Belgian financing. Its main innovative feature was an airtight cabin, equipped with pressurized air; today this technique has become commonplace on airplanes. Another innovation was the design of a very large balloon having sufficient ascent strength so that, on departure, it need not be completely filled. On May 27, 1931, Auguste Piccard and Paul Kipfer reached an altitude of 15,781 metres (51,762 feet), where the atmospheric pressure is about 1/10 that at sea level. A jammed valve made descent impossible in daylight, and they were obliged to wait for the cold night air to contract the balloon. They landed on an Austrian glacier. Interest in the flight was enormous: because no communication had been received from the two scientist-adventurers, they were believed to be dead. After they were found alive, they were received triumphantly, first in Zürich, then in Brussels; the tall silhouette of "Professor Piccard" became world famous.

In 1932, in a new cabin, this time equipped with a radio, Auguste Piccard was able to reach an altitude of 16,940 metres (55,563 feet). In 1933, using the same technique but with bigger balloons, other balloonists rose to 18,500 metres (60,700 feet) in the U.S.S.R. and 18,665 metres (61,221 feet) in the United States.

In 1926 Jean Piccard established residence in his wife's native land, the United States, while he was a research instructor at the Massachusetts Institute of Technology. The lives of the twins converged again in 1934, however, when, with his wife as pilot, Jean followed his brother's example by making an ascent into the stratosphere, attaining an altitude of 17,550 metres (57,564 feet). In 1936 he was appointed professor of aeronautical engineering at the University of Minnesota.

It was known that in order to go higher bigger balloons would be necessary and that they would be difficult to build. Jean Piccard attempted a launch in a cluster of large balloons in 1937; after a successful flight, the equipment was accidentally destroyed by fire. He also devised a plastic envelope that would enclose the cords supporting the nacelle, or balloon gondola. Under his direction, the University of Minnesota played a paramount role in the conquest of the stratosphere.

As a child, Auguste had been fascinated by accounts of marine fish and thought that man should also descend into the depths. As a student, he had conceived of engines capable of withstanding great pressures, yet still manoeuvrable. Now, after his aeronautical successes, he wanted to build a device capable of resisting the pressures of the ocean depths, the bathyscaphe.

Depth-resistant cabins are, of necessity, heavier than water. Up to that time around 1948 they had been suspended from a cable, but, at great depths, this procedure was not dependable. Piccard revolutionized the dive by the principle of the balloon. Just as a lighter-than-air balloon carried the nacelle, a lighter-than-water float would support the cabin. Just as the balloon required a release of ballast to rise, the bathyscaphe would release weight in order to ascend. Fundamentally lighter than water, it would be safe. Air, because it is too easily compressed, was not used in the floats; Piccard chose gasoline.

The bathyscaphe

World War II interrupted construction of the bathyscaphe, which was not completed until 1948. Dives at great depths were conducted off the coast of Senegal, where the French government had stationed a ship with a military team under the command of Jacques Cousteau. In October 1948 an unpiloted trial dive, with the bathyscaphe F.N.R.S. 2 was conducted successfully. The cabin withstood the 1,400-metre (4,600-foot) pressure perfectly, but the float was severely damaged by a heavy swell of water encountered after the dive. The float had to be abandoned, and more solid structures were built around the cabin to permit towing at sea. Although collaboration with the French navy had been secured, it was a difficult job, both administratively and practically; the sailors found it disagreeable to work with a foreigner, a civilian, and a landlubber. Tensions mounted until Jacques Piccard, Auguste's son, intervened.

Jacques, born on July 28, 1922, the only son among Auguste's five children, had become an assistant in economics at the University of Geneva and so had not followed the family tradition in science. Destiny, however, was to bring him back into the fold. As an economist, he had already conducted the negotiations with the French government. Then, while in Trieste, Italy, for the purpose of preparing a study on that port, he received an unexpected offer from local industry to build a new bathyscaphe. Thus, in August 1953, two bathyscaphes competed in the Mediterranean, at Toulon, France, and at Naples: the French craft F.N.R.S. 3 descended to 2,100 metres (6,900 feet), and the Piccards' "Trieste" went down to 3,150 metres (10,330 feet). At the age of 69, August Piccard had realized his dream. His son, abandoning economics, followed in his father's footsteps and collaborated in the construction of the bathyscaphe. It was the year of Jacques's marriage to Marie-Claude Maillard, the daughter of a minister. In 1954 Auguste Piccard retired and left Brussels for Switzerland.

The "Trieste"

France recaptured the diving record from the "Trieste" by going to a depth of 4,050 metres (13,280 feet). The "Trieste" continued to dive in 1954 and 1956; in 1957, the U.S. Navy financed 26 dives off Naples and then bought the engine. Jacques Piccard, hired as a scientific consultant, went to live in San Diego, California, and in 1959 the bathyscaphe was transported to the island of Guam, where it registered dives of 5,530 metres (18,150 feet), and then 7,000 metres (23,000 feet), and on January 23, 1960, a dive of 10,916 metres (35,800 feet).

Auguste Piccard died of a heart attack on March 24, 1962 in Lausanne; Jean, deeply affected by his brother's death, survived him only a few months (died January 28, 1963, in Minneapolis, Minnesota). Jean's wife, Jeannette, became an upper-atmosphere expert for the U.S. National Aeronautics and Space Administration. His three sons, born in 1920, 1924, and 1926, became, respectively, an engineer; a professor of political philosophy; and a manufacturer of hot-air balloons.

After completing his American contract, Jacques Piccard returned to Lausanne. He designed and built the "Auguste Piccard," a mesoscaphe, or sight-seeing submarine, the star attraction of the Swiss National Exhibition held in Lausanne in 1964 when, during the summer, it carried 33,000 tourists to a depth of 100 metres (300 feet) in the lake. Later, the vessel became the property of a United States-Canadian diving firm.

Jacques continued to develop new projects, one of which interested a large American firm: a submarine, which, by balancing in water, could be carried along by subsurface currents in the Gulf Stream; called the "Ben Franklin," it was built in Switzerland. The project involved another long stay by Jacques in the United States, this time in Palm Beach, Florida. From July 14 to August 14, 1969, the "Ben Franklin," with Jacques Piccard and five observers aboard, drifted in the Gulf Stream along the U.S. coast. As scheduled, the experiment lasted one month without the submarine's being propelled; the water was continuously monitored for several characteristics, and underwater soundings were taken.

Gulf Stream research

Jacques returned to Lausanne after completing this contract and began work on new projects, including a submarine for studying pollution in Swiss lakes. He also established a foundation for environmental protection. Though dispersed throughout the world, the family has maintained ties with the little village of Lutry, on Lake Geneva, for Swiss tradition holds that one is forever bound to the cradle of his ancestors.

BIBLIOGRAPHY. No book-length biographical study of the Piccard family exists in English; in French, see PIERRE DE LATIL and JEAN RIVOIRE, *Le Professeur Auguste Piccard* (1962). The same authors' *A la recherche du monde marin* (1954; Eng. trans., *Man and the Underwater World,* 1956), deals with their explorations. English translations of the writings of the Piccards are: *Earth, Sky and Sea* and *In Balloon and Bathyscaph* (both 1956), by AUGUSTE PICCARD; and *Seven Miles Down* (1961) and *The Sun Beneath the Sea* (1971), by JACQUES PICCARD.

(P. de L.)

Piciformes

The bird order Piciformes includes the familiar woodpeckers, their relatives the piculets and wrynecks (collectively, the family Picidae), and the exotic tropical jacamars (Galbulidae), puffbirds (Bucconidae), barbets (Capitonidae), honey guides (Indicatoridae), and toucans (Ramphastidae). This arboreal group of about 370 species is distributed on all continents except Australia and Antarctica, but only the woodpecker family is widespread outside the tropics. Although the six families comprise the order, over half of the species represent one subfamily (Picinae, true woodpeckers) of the family Picidae. The order includes such familiar birds as the European great spotted woodpecker (*Dendrocopos major*) and green woodpecker (*Picus viridis*) and the American flicker (*Colaptes auratus*) and downy woodpecker (*Dendrocopos pubescens*). Piciforms are economically important because they include many insect-eating species. A few species are harmful in eating fruit (toucans) or damaging trees (sapsuckers, genus *Sphyrapicus*), but even these species eat some insects and other animals, and they of course serve as natural balances within the ecological systems of which they are a part.

Drawing by R. Keane

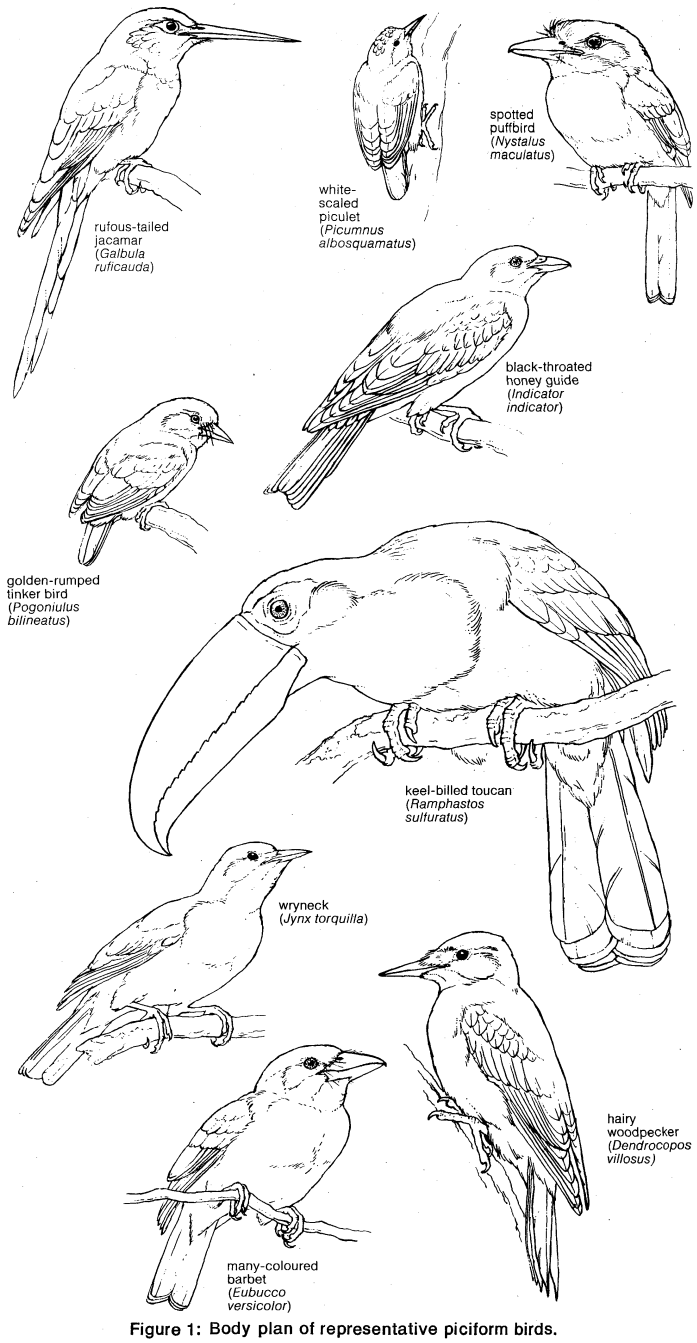


Figure 1: Body plan of representative piciform birds.

The tropical distribution of most piciform families suggests that these represent specialized remnants of a once more numerous and diverse array of arboreal birds. Perhaps many elements of the order became extinct during the geologically recent burgeoning of the diverse, ubiquitous, and highly successful order Passeriformes, which the piciforms resemble in many ways. It has even been suggested that both groups be merged into one order. Distribution of the order

Piciform species vary in size from about nine to more than 60 centimetres (3½ to 24 inches) in overall length. They vary greatly in the structure of their beaks and only slightly less in the rest of their morphology, some, such as the huge-billed toucans and the sturdy woodpeckers being very specialized. The specialized habits of the wax-eating honey guides are unique among birds. The most numerous and widely distributed groups, the barbets (70 species) and the woodpeckers (about 203 species), excavate their own nesting cavities, thereby avoiding competition with other birds and, incidentally, providing homes for many other species of vertebrates.

IMPORTANCE TO MAN

Most piciform birds consume insects, some foraging in places (within bark) reached by no other birds. Thus, they are valuable in the control of insects, even helping to prevent the spread of tree diseases, such as Dutch elm disease, by destroying insect carriers. Fruit-eating barbets and toucans spread seeds of certain tropical trees, the seeds being passed through their bodies at distances away from the parent tree. Honey guides lead men, as well as other mammals, to sources of honey. Colourful toucans, barbets, and jacamars, and even the dull-coloured puffbirds delight natives and ever-increasing visitors to the tropics. Woodpeckers are conspicuous and often colourful species of forests, and their actions have long elicited interest and admiration from man—their winter visits to suet feeders in the Northern Hemisphere are anticipated and enjoyed by millions of onlookers. The excavation of nesting and roosting holes by woodpeckers and barbets provides homes for countless birds of other species, which utilize the holes after the excavators are finished with them. Foresters in Europe frequently compensate for the lack of old woodpecker holes needed by valued insect-eating species by placing artificial nest boxes in their cultivated forests.

Fruit eating by barbets and toucans sometimes causes damage to cultivated fruit plants, especially near villages and towns. The few species of sapsuckers in North America and southern Asia occasionally destroy trees by creating openings through which disease agents may enter, although wounds in healthy trees rapidly heal. Piciform birds that excavate their nests usually choose dead or dying trees, but some excavate in live trees or cacti. The area around the excavation may be sealed off by scar tissue in some trees, especially cacti, resulting in no damage but diminishing the tree's value. Some woodpeckers damage fence posts and utility poles, particularly in open country, in well-kept woodlots, and near cultivated forests where dead trees are scarce.

NATURAL HISTORY

Distribution and ecology. Piciform birds are varied in their habits, although certain facets of behaviour relating to hole nesting, such as the snakelike hissing of young birds in the nest, are shared with some other hole-nesting birds. Jacamars and puffbirds, the fly-catching members of the order, are stolid, unsuspicious birds, allowing close approach by humans. Woodpeckers, toucans, and barbets are more wary, but the smaller species of piculets and woodpeckers sometimes can be approached closely.

Jacamars and puffbirds inhabit various portions of tropical American forests. Jacamars sit quietly on perches in openings and edges, flying infrequently, but often flicking their tails. Large insects, such as butterflies, are captured, taken to a perch, and usually beaten against it. The wings of the insects are removed if the prey are to be fed to the young.

Puffbirds are often found in farm areas, where they may be seen along the road looking something like dull-coloured forest kingfishers. They have a rapid, darting flight, and some species, such as the swallow-winged puffbird (*Chelidoptera tenebrosa*), spend much time making sorties after insects high in the trees. Others forage low in bushes, obtaining prey on the ground as well as in undergrowth. Tail-jerking movements are observed in some puffbirds (*e.g.*, the two-banded puffbird, *Hypnelus bicinctus*), as in jacamars.

The barbets are weak-flying, essentially resident birds inhabiting woodland and savannas, occasionally the fringes of deserts. They feed on insects, fruit, and even flowers and buds. The tiny African tinkerbirds (*Pogoniulus*), so named for their repetitive vocalizations, feed mainly or entirely on insects, foraging rapidly over bark and occasionally tapping with the beak, very much like piculets of the woodpecker family. Some barbets have serrated, or "toothed," edges of the bill, in the manner of toucans; these are apparently used for tearing open fruit.

Guiding by honey guides

Honey guides are nonmigratory birds, dwelling in pairs or alone in forests and savanna. They feed on insects, fruit, and beeswax. Only a few honey guides of the genus *Indicator* actually are known to "guide" certain mammals, including man, to bees' nests. Most honey guides are able to enter bees' nests without assistance, there to eat larvae and, primarily, wax, which is digested with the help of special bacteria within the birds' digestive tracts. Honey guides that "guide" use a peculiar chattering to attract the attention of an animal likely to follow. The greater honey guide (*Indicator indicator*) spreads its tail conspicuously and uses a peculiar undulating flight in attracting attention. It perches and calls again, then moves further on in another short flight, as the interested animal draws near. Upon reaching the bees' nest, the honey guide waits quietly in the vicinity until the animal has taken the honey and moved off. It then flies to the devastated nest and feeds on scattered pieces of wax and exposed larvae.

Toucans are strictly New World tropical-forest birds, although they may forage about farms. One group, the hill toucans of the genus *Andigena*, occurs in the Andes up into temperate forests. No toucans are geographically migratory, but some species make local movements up and down mountain slopes. Toucans eat much fruit, perhaps aided by their serrated beaks. Smaller pieces or small fruits are first held in the tip of the bill and then thrown back into the mouth. They also eat larger insects, birds' eggs, and even young birds. The colour patterns of the bill are distinct in each species and probably serve a behavioral function. Toucans have a peculiar habit of cocking the tail forward over the head, particularly while sleeping in a cavity; D'Arnaud's barbet (*Trachyphonus darnaudii*) also does this during displays.

The family Picidae is comprised of three distinct subfamilies. One, the Jynginae, contains two closely related species of wrynecks (*Jynx*), found in Eurasia and central Africa. Wrynecks, which have soft, unspecialized tails, forage for ants in trees and on the ground.

The piculets (Picumninae) comprise a second subfamily of soft-tailed, woodpecker-like birds. All but four of about 30 species live in forests and dry woodlands of the New World tropics. The major genus *Picumnus* is peculiarly distributed in having about 24 species in the New World and one species (*Picumnus innominatus*, the speckled piculet) in southern Asia, a pattern of distribution matched by the woodpecker genus *Celeus*. All the piculets are tiny, the largest—the peculiar Haitian piculet (*Nesoctites micromegas*)—being only sparrow sized. They feed in a woodpecker-like manner, tapping and probing usually on small branchlets and vines, and they allow close approach, continuing to forage under the eyes of an observer. Their food is composed mainly or entirely of insects.

Feeding behaviour of woodpeckers

The true woodpeckers (Picinae), with their stiffened tails, usually perch lengthwise on tree branches or trunks. Among the relatively few species that go to the ground, all hop, except for two species of flickers (campo flicker, *Colaptes campestris*, and Andean flicker, *Colaptes rupicola*) that walk. Foraging is by diverse modes. True excavating, which penetrates the substrate of the bark, is not as common as one would expect on viewing the woodpecker bill. Most woodpeckers forage mainly by gleaning items from the surface of a tree, by probing into cracks, and by occasional to frequent tapping, sufficient to break barely through the surface. Only the heavy-billed, so-called ivorybills (genus *Campephilus*) of the New World, some of their relatives, other similarly specialized Asian groups, and a few of the smaller woodpeckers (genus *Picoides*; some species of *Dendrocopos*) regularly chisel deep into trees. The food of woodpeckers is varied, mainly consisting of insects, but some species (especially of the genus *Melanerpes*) frequently eat fruits and berries. Ants are a favourite food, especially in tropical regions—fully half of the numerous tropical species feed largely or entirely on ants. Three woodpeckers (ground woodpecker, *Geocolaptes olivaceus* of Africa, the campo flicker, and the Andean flicker) are ground dwellers, feeding mostly on ants, while seven or eight other species frequently forage on the ground. The flight of woodpeckers is strong, with definite up–down undulations.

Most woodpeckers are permanent residents, the few exceptions including the North American yellow-bellied sapsucker (*Sphyrapicus varius*) and flicker, which are highly migratory.

Reproduction. The ability of most piciform birds to excavate their own nest holes is probably a major factor responsible for their success. With relatively diverse types of bills, the jacamars, puffbirds, and barbets excavate nesting cavities, as do the sturdier billed woodpeckers, although the former utilize a softer substrate such as earth or dead trees. Woodpeckers alone within the order have biologically "capitalized" on their adaptations for excavating nests, by various further specializations for nesting (even in hard, living trees) and for foraging on the surface and within the bark of trees. They have thus exploited niches unavailable to other birds, a fact that accounts for their success relative to other members of the order. Yet because of their specialized habits, only a limited number of kinds of woodpeckers can exist together, usually no more than five to eight species (more in some tropical environments), differing somewhat in size and habits.

Difficulties in the study of hole-nesting birds

Because of the concealing nature of the nest cavity, the behaviour of hole-nesting birds is more difficult to study than that of species building exposed nests. The nesting behaviour of most piciform families must be inferred from observations on a few examples, the woodpeckers being the only group that has been extensively studied.

Jacamars are usually resident wherever they occur and nest at diverse times of the year depending on the locality. Nests are excavated in banks or in steep hills. Both sexes dig with the beak, kicking out dirt with their feet. The tunnel is comparatively short, usually less than one metre (three feet) in depth. No lining is placed in the nest chamber at the end of the tunnel. Three or four white eggs are incubated by both sexes, with the female incubating at night. The incubation period of several species of jacamars is known to be from 19 to 21 days. The young are downy and hatch with their eyes closed. They remain for 20 to 26 days in their nest before leaving. Most of the food consists of large insects brought to the young by both parents. The parents do not clean debris and fecal material of the young from the nest, which becomes littered.

Puffbirds excavate their nests in the ground or in termite nests on the ground or in trees. The tunnel is usually deeper than that of a jacamar, often reaching a metre or more into the ground. Several pairs may nest close together, semicolonially. Both sexes excavate the burrow. Some species leave the chamber at the end of the tunnel unlined, while others may line it with grass or dead leaves. Several species, including the white-whiskered puffbird (*Malacoptila panamensis*), pile sticks and leaves at the mouth of the nest tunnel, effectively concealing it. Both sexes incubate the two (occasionally three) white eggs. The newly hatched young are naked and are fed by both parents until old enough to leave the nest.

Social nesting and roosting

Although the majority of barbets nest in isolated pairs, there is a marked trend toward sociality. Large numbers of one or several species may gather to feed at a fruiting tree. Some species are found in family groups, and displaying groups of several individuals may be seen during the onset of breeding. The most social of all are the plain-coloured barbets (genus *Gymnobucco*) of Africa, the brown barbet (*Calorhamphus fuliginosus*) of southeastern Asia, and some species of the genus *Megalaima* of southern Asia. These nest colonially and feed in gatherings of up to 50 or more birds (*Gymnobucco*). They are aggressive, noisy birds and may take over an entire dead tree for nesting, driving away other bird species. In some barbets of the genera *Semnornis* and *Megalaima*, communal roosting occurs, with numbers of birds spending the night together in the same hole. Nests are usually excavated in trees, but a few species burrow into the ground or into termite nests. Both sexes excavate the cavity, which typically has a nearly round entrance, small for the size of the bird. The same cavity may be used for several nesting seasons. No lining material is placed in the nest cavity. The shape of the barbet's bill, which is not at all chisel-tipped like that of a woodpecker, makes the excavation of cavities in trees a task requring great effort, and it appears that only dead, well-rotted trees are used. Two to five white eggs are laid in the cavity; both parents then incubate the eggs for 13 to 15 days or more. Both also may spend the night in the nest. The young develop slowly, spending 30 or more days in the nest. The adults clean the nest, removing excreta. Often the young and adults continue to roost together after the young have left the nest, occupying a separate "dormitory" hole excavated for that purpose. On occasions, more than one pair of birds may use the same nest (white-eared tinkerbird, *Pogoniulus leucotus*).

Most, if not all, of the honey guides are brood parasites, laying their eggs in the nests of other species, especially those that nest in holes or in deep, covered nests. The host's eggs often are punctured by the female honey guide when she lays her own egg. Only a single egg is laid in each host's nest. Because of the parasitic habit, the incubation periods of honey guides are not well known, but periods of 12 to 16 days have been found in species of *Indicator*. The young birds spend a long time in the nest, 35 to 40 days or more. When newly hatched, they have hooks on the tips of their bills, which are used to injure and kill the young of the host species. These hooks later fall off.

Toucans nest in natural cavities or in old woodpecker nests. Smaller species may usurp a nest freshly constructed by woodpeckers. Old, well-rotted cavities may be enlarged slightly. One to four eggs are laid on the floor of the nest, which may be unlined or lined with a few leaves. About 16 days are required for incubation, performed by both sexes. The birds are shy about the nest and hence difficult to observe. The young, naked when hatched, bear large pads on their heels, on which they perch in the nest. The nestling period is generally long, up to seven weeks or more, varying with the size of the birds. Toucans, except the toucanets (genus *Aulacorhynchus*), do not maintain clean nests. Family groups may remain together for a long time, for small flocks are often seen throughout the year.

Wrynecks usually nest in natural crevices, old woodpecker holes, or old swallow holes in banks. They lay two to 14 eggs, and both sexes incubate. Like the American flicker and probably other woodpeckers, the wryneck is an indeterminate layer; *i.e.*, the bird produces eggs until a certain number have accumulated.

The nests of piculets are constructed in small trees, stubs, fence posts, and the like. The entrance is tiny and rounded, resembling a barbet nest entrance. Two to four eggs are laid and incubated for about two weeks. The initially blind and naked young grow for three weeks or more before leaving the nest for the first time. Both parents incubate and feed the young, and in at least some species they both remain in the nest at night. After fledging, the young may continue to roost in the nest at night with their parents for several months. Like the wrynecks and most other birds and unlike woodpeckers, the piculets often perch crosswise on branches.

Displays of woodpeckers

The reproductive behaviour of woodpeckers is better known than that of most other Piciformes. The sexes are clearly marked by simple pattern differences, most commonly involving the presence or absence of red on the head. Observations at bird-feeding stations suggest that sexual recognition occurs throughout the year, with males dominant over females. Aggressive displays are conspicuous and numerous in woodpeckers, including various head postures centring on the bill, bill pointing, head turning, swinging of the head and body from side to side, bobbing and bowing of the head, wing flicking and spreading, tail spreading and turning, and even aerial displays. These displays exhibit the complex head markings, as well as the body, tail, and wing patterns found in most species. Many vocal and instrumental signals are used, including the widely prevalent tapping or drumming of the bill against a tree or, in cities, on tin roofing.

Most woodpeckers actively defend territories, often throughout the year. This perhaps accounts for their strongly developed aggressive behaviour. Few species are markedly social, but members of the New World genus *Melanerpes* especially tend to be, even while nesting. Up to 11 different adults of the yellow-tufted woodpecker (*Melanerpes cruentatus*) were observed feeding young in three different nests in eastern Peru. Some of the adults fed young in two and even in all three nests. Andean flickers nest in loose colonies in banks and may be seen in groups of ten or 12 birds. Otherwise, social gatherings of woodpeckers generally involve family groups during and sometimes after the breeding season. Most woodpeckers are aggressive toward others of their own species, even to some extent toward their mates and young. Aggressive encounters are most frequent between members of the same sex. When an intruding woodpecker invades the territory of a pair during the breeding season, it usually is attacked by the member of the pair that is of the same sex as the intruder. Conflicts between birds after the breeding period are more intense when the woodpeckers involved are of the same sex. Experiments with the North American flicker show that even paired birds recognize each other primarily by their sexual colour markings, for a female given an artificial black moustache stripe (characteristic of the male) was attacked and pursued unceasingly by her mate until the experimenter rescued the exhausted female and removed the "moustache." She was once again fully accepted by her mate.

Nesting of woodpeckers

Nests are excavated by the birds themselves, although a few species will accept artificial nest boxes. Two species, the ground woodpecker and the Andean flicker, always nest in the ground, digging nests in banks, and the campo flicker of South America often does so. Other woodpeckers chisel holes in trees, fence poles, cacti, or similar sites. Usually both sexes construct the nest, but either sex may do most of the work, depending on the species. The nest cavity varies with the size of the bird. A flicker (30 centimetres, or 12 inches, long) was found incubating eggs in a tree stub only 12 centimetres (five inches) thick; so less space is required for nesting than one might expect. Nevertheless, the size of the stub or tree necessary for nesting is proportional to the size of the woodpecker. Also, the size and abundance of dead or dying trees influence the abundance and types of woodpeckers that can occupy an area.

From two to nine white eggs are laid on the bare floor of the woodpecker's nest cavity. Most species probably are indeterminate layers. By removing the freshly laid eggs as they appeared daily, one experimenter induced a female flicker to lay a total of 71 eggs. The incubation period of woodpeckers varies from 11 to 20 or more days. Both adults incubate the eggs, and typically the male alone occupies the nest during the night. The young are naked and helpless when hatched and spend up to 36 days in the nest. Both parents feed the young and keep the nest clean by removing excreta frequently. In some species the young are driven off by the adults soon after they can feed themselves; in others the family remains together longer, sometimes even until the next breeding season.

Vocalizations. The voices of Piciformes are rarely melodious and are often harsh or strident. The vocalizations of jacamars are squeaky, the notes sometimes being run together into a trill. Whistles or trills may be alternated or mixed, forming a simple song. Puffbirds are relatively quiet, producing thin whistles, peeps, and twitters. The vocalizations of toucans are loud and often harsh, especially those of the larger species, such as those in the genus *Ramphastos*. The calls of barbets are monotonous and repetitive, but some are bell-like and pleasing to the ear, especially those rendered antiphonally (*i.e.*, the birds alternating) or simultaneously by members of a pair. Displays may accompany this singing. With some colour patterns like their relatives the woodpeckers, barbets have similar displays, such as head swinging, head bobbing, bill pointing, and others.

Although the vocalizations of woodpeckers are individually less complicated than the songs of passeriform birds (the order that includes the songbirds), up to eight or nine different and varying calls may be used by some species (*e.g.*, the ladder-backed woodpecker, *Dendrocopos scalaris*). These function as "advertising songs" (usually repetitive, often harsh, "rattle" or "wick" calls), as aggressive or submissive calls associated with displays, and in other ways.

FORM AND FUNCTION

Adapted for diverse modes of life, piciform birds vary strikingly in their morphology and in their habits. The interpretation of differences is difficult in many cases, especially considering that different structural means may achieve the same purpose. Thus, the fly-catching jacamars are exceptionally long-tailed, but the puffbirds, which feed similarly, are generally short-tailed.

Characteristic of all members of the order is the yoke-toed (zygodactylous) foot in which the outer toe (toe IV) as well as the hind toe (I) points to the rear. This was long considered an adaptation to climbing or perching on the vertical trunks of trees, but recent studies have shown that this type of toe arrangement may really denote a generalized perching foot. Most of the strongly wood-pecking species of the Picidae that are four-toed spread the toes laterally and anteriorly—*i.e.*, up the tree that they are climbing, with no toes pointing to the rear. Three-toed woodpeckers, such as the northern three-toed woodpecker (*Picoides tridactylus*), which have lost the hind toe (hallux, toe I), climb with all three toes facing to the front and sides and swing the inner toe (II) backward when perching on a branch. The hallux tends to be small in those Picidae that have retained it, and its loss has occurred independently in several lines of woodpeckers and piculets.

Foot specialization

Just as birds in general are structurally similar because of requirements imposed by flight, woodpeckers are structurally similar because of requirements imposed by wood-pecking habits. The special modifications responsible for their unique abilities are diverse. The tail feathers (especially the central one or two pairs) are stronger in woodpeckers, resisting the wear caused by their use in propping the bird's body as it hammers with the bill. The toe structure and associated arrangement of tendons and leg muscles form a functional complex of features enabling the woodpecker to climb tree trunks and to maintain its position while pecking the tree.

Another functional complex involves the skull, head musculature, bill, tongue, and associated structures of the woodpecker, collectively its food-gathering apparatus. The skull is unusually thick, and, in the most specialized wood-pecking species, it curves inward at the upper base of the bill, instead of meeting it directly, giving a built-in shock absorber. The head musculature is specialized to power the thrusts of the bill and to help absorb the shock caused by pecking. The long, distensible tongue is variously equipped with tiny barbs and coated with a sticky substance provided by often greatly enlarged and modified salivary glands. The horns of the hyoid apparatus responsible for the ability to extend the tongue are tremendously elongated, and when the tongue is retracted, they often extend around the back of the skull, over the top, and (rarely) even onto the bill and into a nostril (see Figure 2). The tongue can be extended into holes made

Drawing by R. Keane

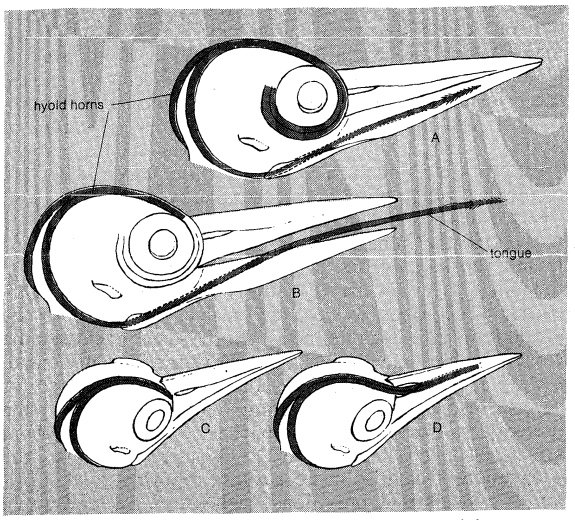


Figure 2: Position of hyoid bones (shaded) with tongue (A) retracted and (B) extended, shown in the hairy woodpecker (*Dendrocopos villosus*). (C and D) Other configurations of hyoids: C ending at the base of the beak; D entering the right nostril and lying within the upper mandible.

by the bill and can be used both as a probe and as a "spear" (by means of the barbs and sticky mucus) to locate and extract insects far beyond the bill's reach. The bill of more specialized woodpeckers is chisel tipped, broad based, and very sturdy. The nostrils are slitlike, preventing sawdust and wood chips from entering. Other parts of the body may be modified for support or to prevent damage from the shock of pecking.

An advantageous side effect of their specialization is that woodpeckers are exceptionally tough-skinned and sturdy and thus are difficult to kill. Unfortunately, this has not saved the magnificent, highly specialized ivory-billed woodpecker (*Campephilus principalis*) and its close relative the imperial ivory-billed woodpecker (*Campephilus imperialis*), by far the largest of all woodpeckers, from extinction through destruction and modification of habitat. A remnant population of the former is known to exist in Cuba, but the future for both species is bleak.

Among woodpeckers, sexual difference in bill structure and other characteristics plays a role in allowing the members of a pair to utilize somewhat different types of food, presumably facilitating more efficient foraging.

Of the other piciform families, the Indicatoridae are specialized in features, such as their thick skin, relating to their beeswax eating and brood parasitism. Other peculiarities, such as the graduated long tail of the lyre-tailed honey guide (*Melichneutes robustus*) and the presence of but nine primaries (there is usually some trace of a tenth primary feather) in the wings of most honey guides, are more difficult to interpret functionally.

The bill of the toucans, another specialized group, is something of an anomaly. Very large, with serrate edges and usually splashed with colour, it is peculiarly constructed within. Many crisscrossing, fine, bony fibres give it great strength for its size and light weight. The serrated bill may be effective in tearing open fruits, although the birds often feed on small fruits daintily picked, held at the tip of the bill, and thrown back into the mouth. Alternatively, the serrations may aid in securing insects or in plucking young birds out of nests. The colour patterns of the bill probably function as species-recognition signals or in some other way during displays. Yet the size of the bill seems greater than necessary for the serving of these functions. Studies of the feeding habits of the larger species may enable the solution of this point.

The toucan bill

EVOLUTION AND PALEONTOLOGY

Fossil species of the order Piciformes are almost unknown. Many avian fossils identified long ago are in need

of verification, including the early fossil "woodpeckers," some of which are now believed to belong in other orders. Fossil woodpeckers are now known only from the Pliocene Epoch (about 6,000,000 years ago). Except for a single specimen from the middle Miocene of Bavaria assigned to the Capitonidae, the other Recent families have no fossil records. A fossil family, Zygodactylidae, was erected in 1971 to receive three species of fossil piciform birds from the Miocene of western Europe.

The dearth of fossils leaves the evolution of existing groups open to speculation. If it is true that existing families of the order are only specialized remnants of a formerly more diverse assemblage, then the generalized groups (families or even suborders) that might have included common ancestors of some of the modern families are now extinct. Therefore, the evolution of the order must be interpreted from inferences based on the morphology of its existing families.

Little can be said about the puffbirds and jacamars. Both families are relatively uniform, and no existing species shows intermediacy or close resemblance between them. They must have evolved very early in the history of the order from a generalized, fly-catching ancestor. If the Galbulae and Pici are indeed closely related, their divergence from a common ancestor must have occurred long ago.

Ancestors of the woodpeckers and barbets

The families within the suborder Pici are definitely interrelated. The honey guides and barbets are morphologically very similar, as are the toucans and barbets. The barbets are rather diverse, foliage and branch-foraging piciform birds. An ancient species of barbet, not greatly unlike modern barbets, easily could have evolved into the ancestral toucan and another such species into the ancestral honey guide. Woodpeckers and barbets share an ancestor probably dating from an earlier period than either the barbet–toucan or the barbet–honey guide ancestor. The peculiar wrynecks, and even the piculets, may represent relicts of early evolutionary radiations along the line leading to woodpeckers. Modern woodpeckers have radiated into many groups, some specializing greatly for an arboreal wood-pecking existence, others remaining generalized, and still others becoming secondarily adapted to life on the ground.

CLASSIFICATION

Distinguishing taxonomic features. Various characteristics set the Piciformes apart from other orders and distinguish subgroups within this order. Among these features are the condition of the toes and their flexor tendons; the condition of the nostrils; the absence of basipterygoid processes at the base of the skull; the arrangement of the pelvic muscles; the size of the deltoid muscle; the condition of the syringeal muscles.

Annotated classification. The classification below is based generally upon those of a British ornithologist, P.R. Lowe, and an American, C.G. Sibley. The groupings are accepted generally among taxonomists, with disagreement centring about the taxonomic rank of the subcategories (*e.g.*, superfamily versus family) and whether or not to merge part of the order with the perching birds (Passeriformes). The family indicated by the dagger(†) is extinct.

ORDER PICIFORMES

Zygodactylous birds (toes arranged 2 in front and 2 behind); flexor tendons with a vinculum (a connecting strip), digits I, II and IV supplied by flexor hallucis, and flexor digitorum supplies digit III; nostrils impervious and holorhinal (not deeply cleft); eutaxic (5th secondary present), 9 to 11 primaries, 10 to 13 secondaries; 8 to 12 rectrices (tail feathers); aftershaft present; no down in adults; hole-nesting birds, laying all-white eggs. About 370 species.

Suborder Galbulae

Desmognathous (vomer bone of palate small or lacking; maxillopalatines connecting); nude oil gland; 2 carotid arteries; no gallbladder; 10 primaries, 10 to 12 secondaries, 10 to 12 rectrices. Catch food, usually insects, on wing; nest in ground or termite nest.

Family Galbulidae (jacamars)

Tropical and subtropical Central and South America. Bill long. Vomer lacking. Legs with long smooth scales behind and scutellate (small scales) in front. Moderately sized, long-tailed birds with iridescent plumage dominated by green. Feed by fly-catching for large insects. Nest in burrows excavated in banks.

†Family Zygodactylidae

Middle and upper Miocene of France and Bavaria; 3 species.

Family Bucconidae (puffbirds)

Tropical and subtropical Central and South America. Distinguished by a shorter, wider bill than jacamars, with a decurved tip. Vomer present. Leg scales scutellate in front and behind. Moderate or small-sized, dull-coloured birds with an unusually large head, plumage lax and noniridescent. Feed by fly-catching for insects and other arthropods and vertebrates as well. Nest in burrows in the ground, or in termite nests on the ground or in trees.

Suborder Pici

Rarely desmognathous, usually aegithognathous (vomer broad; maxillopalatines not connecting) or saurognathous (vomer a delicate rod; maxillopalatines reduced, not connecting); oil gland usually feathered; 9 to 11 primaries, 10 to 13 secondaries, 10 to 12 rectrices. Nest in tree cavity, rarely in ground. Food diverse, secured by various means, but rarely by fly-catching.

Family Capitonidae (barbets)

Tropical and subtropical New World, Africa, and Asia. Small to medium bill, not chisel-like; short, nonextensible tongue; nostrils opening onto bill surface without a fossa (depression); conspicuous bristles at base of bill. Small to medium, usually brightly coloured birds of stocky appearance. Food fruit or insects, often gleaned from tree bark or foliage. Nest hole in trees or termite nests, usually excavated by the birds.

Family Indicatoridae (honey guides)

Tropical and subtropical Africa and Asia. Small bill, not chisel-like; short, nonextensible tongue; nostrils opening onto bill in centres of large fossae, and without conspicuous bristles at base of bill. Small or medium, generally dull-coloured birds with smallish heads. Food beeswax and insects. Brood parasites, laying eggs in nests of other birds, especially piciform birds.

Family Ramphastidae (toucans)

Tropical to temperate Central and South America. Distinguished by a long and often deep and brightly coloured bill, with serrated edges. Tongue nonextensible but long and fringed. Medium to large, long-tailed and often brightly coloured birds. Food mainly fruit, also insects, even eggs and small birds. Nest in natural tree cavities or woodpecker holes, never excavated.

Family Picidae (woodpeckers, piculets, wrynecks)

Generally worldwide, except Australian region. Distinguished by a straight, usually chisel-like bill, long extensible tongue with elongated hyoid (tongue) apparatus, salivary glands secreting a sticky material that coats the tongue. Very small to large birds, often patterned in black, white, and bright colours. Forage on tree trunks and branches, rarely on the ground. Food chiefly insects, some fruit, rarely sap and seeds. Nest excavated (except wrynecks) in tree, post, cactus, or (rarely) in the ground.

Critical appraisal. Although recognizing the relationship of this order to perching birds (Passeriformes), most taxonomists prefer to keep the two groups separate as distinct orders, especially since clear evidence for relationship with a particular group of perching birds is lacking. No perching bird has the zygodactyl toe arrangement. The exact relationship between the suborders Galbulae and Pici has not been established. Lowe advocated relationship of the Galbulae with the Coraciiformes (kingfishers, bee-eaters, etc.), and the merger of the Pici in the Passeriformes. The suborder Pici is often considered to contain three superfamilies, for the toucans, the woodpeckers, and the barbets and honey guides, respectively. The above arrangement within the suborder Pici begins with the barbets, generally recognized as the least specialized group within the suborder and the most likely ancestral family. The honey guides are almost universally considered very closely related to barbets and are placed next. There is agreement that the specialized toucans, placed next, evolved from barbet-like ancestors. The exact relationship of the specialized woodpecker family with barbets is unknown, but it is likely either that woodpeckers evolved from barbets or that the two groups shared a common ancestor. No authority has suggested that woodpeckers are more closely related to toucans or

to honey guides than to barbets. Various authorities have arranged these families diversely, but nearly all place the woodpeckers last. No single arrangement is universally accepted.

BIBLIOGRAPHY. A.F. SKUTCH, "Life-History of the Black-chinned Jacamar," *Auk*, 54:135-146 (1937), is the only relatively complete study of a species of jacamar; "Life History Notes on Puff-Birds," *Wilson Bull.*, 60:81-97 (1948), although not complete on any species, contains most of the information known about the habits of puffbirds. S.D. RIPLEY, "The Barbets," *Auk*, 62:542-563 (1945), is chiefly a taxonomic, technical treatment, but summarizes the habits and characteristics of various groups; A.F. SKUTCH, "The Life-History of the Prong-Billed Barbet," *Auk*, 61:61-88 (1944), presents a fairly complete account of the habits of a barbet, one of few detailed accounts existing. H. FRIEDMANN, *The Honey-Guides* (1955), is a technical but readable account of the habits and relationships of honey guides. J. VAN TYNE, *The Life History of the Toucan* Ramphastos brevicarinatus (1929), provides a good, though technical, account of the habits of a large toucan; H.O. WAGNER, "Notes on the Life History of the Emerald Toucanet," *Wilson Bull.*, 56:65-76 (1944), is a general account of the habits of a small toucan; and A.F. SKUTCH, "Life History of the Blue-Throated Toucanet," *Wilson Bull.*, 56:133-151 (1944), a very readable study of another small toucan; "Life History of the Olivaceous Piculet and Related Forms," *Ibis*, 90:433-449 (1948), presents an interesting account of the habits of these little relatives of the woodpeckers. D. BLUME, "Uber die Lebensweise einiger Spechtarten (*Dendrocopos major, Picus viridis, Dryocopus martius*)," *J. Orn., Lpz.*, vol. 102 (supplement), in German, with English summary (1961), is the classic modern behavioral treatment of woodpeckers, covering three common European species (fairly technical). H. SIELMANN, *Das Jahr mit den Spechten* (1958; Eng. trans., *My Year with the Woodpeckers*, 1959) is a readable, accurate account of the habits of European woodpeckers. L. DE K. LAWRENCE, *A Comparative Life-History Study of Four Species of Woodpeckers* (1966), presents an accurate, not overly technical treatment of the habits of some North American species. A.C. BENT, *Life Histories of North American Woodpeckers* (1939), summarizes, at times anthropomorphically, the knowledge, through about 1938, of the habits of all North American woodpeckers. J.T. TANNER, *The Ivory-Billed Woodpecker* (1942), is an account of the habits and history of one of the largest woodpeckers, now nearly extinct on the North American continent.

(L.L.S.)

Pidgin

The term pidgin is applied to a number of varieties of speech that have grown out of English or other languages and that have been used in various parts of the world since the 17th century. Often termed "bastard jargons," "mongrel lingos," or the like, these tongues in fact are languages like any others and can be accurately delimited and described.

Definitions of lingua franca, pidgin, and creole. When a language is used as a means of communication between persons having no other language in common (*e.g.*, French in 18th-century diplomacy) it is a lingua franca. A lingua franca native to none of those using it and with a sharply reduced grammar and vocabulary is called a pidgin. (This definition of pidgin excludes both the broken English of a beginning learner and the skillful but non-native use of English in such countries as India.) When a whole speech community gives up its former language or languages and takes a pidgin as its mother tongue, the pidgin becomes a creole (is creolized).

Situations in which pidgins developed

Origins. A number of pidgins and creoles have arisen on the basis of various European languages. The first known pidgin, Lingua Franca, or "Westerners' language," of the medieval Levant and the Barbary Coast, was based chiefly on Italian. The American Indians first encountered by Englishmen in the 17th century were a tribe known as Pidians near the mouth of the Orinoco; the reduced language that emerged was termed Pidgin (= Pidian) English. Later in the same century, other varieties of Pidgin English grew up in China as a result of English commercial contacts and in Africa in connection with slave-trading activities. (Some authorities derive the word "pidgin" from a variation of English "business".) Establishment of plantation economies in the Caribbean area, with large groups of Negro slaves from different language backgrounds in West Africa, led to a number of pidgins based on English, French, Spanish, and Portuguese. Many have survived as creoles; *e.g.*, Gullah off the Sea Islands of South Carolina, the Negro English of the Antilles, and Sranantongo (formerly called Taki-Taki) in Surinam (formerly Dutch Guiana), all based on English; the French-based creoles of Louisiana, Haiti, and the Lesser Antilles; and the Papiamento of Curaçao, an outgrowth of Pidgin Spanish and Portuguese. Early contacts between settlers and natives led to the formation of pidgins in Australia and New Zealand, whereas the Pidgin English of the South Seas (called Beach-la-Mar) grew out of whaling, trading, and recruiting native labour. Pidgin English is extinct in New Zealand and the Caroline Islands and moribund in Australia but still flourishes in Melanesia (New Hebrides, Solomon Islands, New Guinea) and has become creolized in Hawaii. One variety, that of the Rabaul area (New Britain), has become the widespread and indispensable lingua franca of the Territory of Papua and New Guinea because of the official sanction given its use under German rule (1884–1914) and later Australian administrations.

Development and persistence of pidgins

Survival. Pidgin languages spring from the initial, non-intimate contacts between speakers of different languages, when quick comprehension is more highly valued than grammatical correctness or fine shades of meaning. As contacts grow closer, normally one group learns the other's language more fully, and pidgins survive the stages of initial contact only in special circumstances. Pidgins persist where a dominant group regards another as child-like or capable only of a simplified version of the "superior" language, as in the relations between Europeans and American Indians, West Africans, or South Sea natives. On plantations and in other situations where European masters were in permanent contact with native servants or labourers, pidgins served as status languages, as in New Guinea. Caste distinction, however, is not a necessary function of a pidgin; Russonorsk, for instance, was a reduced language used by Russians and Norwegians in the Arctic at the beginning of the 20th century. Chinese Pidgin English survived for three centuries and not only in master-servant relations; it also was in use between English merchants and Chinese dignitaries, primarily because each side desired to keep the other at arm's length. Slaves on Caribbean plantations, New Guinea natives in newly founded multilingual villages, and others who have come to live together with no language in common save a pidgin have used the pidgin as the customary language of the group. In such instances, the resultant creole has usually re-expanded its structure and vocabulary by borrowing from the language of a culturally dominant group; *e.g.*, Haitian Creole from French, Sranantongo and Papiamento from Dutch, and Melanesian Pidgin from English.

Spelling and sounds

Orthographies. In its original function as a lingua franca among unlettered folk, a pidgin language is a medium of purely oral communication, as also are creoles in their initial stages. Only afterward, and usually in connection with missionary or other educational programs, are spelling systems devised for pidgins or creoles. Speakers of European languages have often applied the orthographical conventions of their own languages, as when the Melanesian Pidgin sentence for "Why did you hit this policeman?" is written *Belong what name you fight 'im dis fellow police boy*? Such a spelling embodies all the inconsistencies of English orthography and is therefore difficult to learn; it distorts the structure of pidgin; and it confirms the naïve European or American in his belief that pidgin is only a ridiculous reduction of English. Those who have devised orthographies, for ease of learning, accuracy in representing linguistic structure, and emphasis on the independent status of the language, have used phonemically based systems. The most effective orthographies of this type use the letters available on typewriters or in printshops, but consistently and predictably. Thus, the Melanesian Pidgin sentence just quoted reads, in the officially recognized orthography: *Bilong wonem yu faitim dispela plisboi*? In the following discussion, forms are cited only in this type of transcription.

Phonology. The simplification which characterizes pidgin extends to all aspects of linguistic structure (sounds, forms, constructions) as well as vocabulary. In some varieties stress is automatically on the first syllable; *e.g.*, *bíkos* "because," *míshin* "machine." A minimum of five distinctive vowels is necessary, those represented in Latin or Italian pronunciations by *a* ("ah"), *e* ("eh"), *i* ("ee"), *o* ("oh"), *u* ("oo"), as in *antap* "up," *em* "he," *winim* "defeat," *kot* "coat," *tu* "also." The vowel sound *a* ("ah") combines with *i* ("ee") and *u* ("oo") in the basic diphthongs *ai* (English "i") and *au* ("ow"), as in *dai* "cease" and *nau* "now." Some, but not all, speakers make further distinctions—*e.g.*, between the *e* of *em* "he" and the *ei* ("ay") of *neim* "name." In almost all varieties of pidgin English, the two consonant sounds represented by English *th* have merged with *t* and *d* respectively: *saut* "south," *dispela* "this." Many speakers of Melanesian Pidgin merge *f* and *p* (in current official orthography, both are represented by *p*) and also merge *ch* and *sh* with *s*: *tumas* "very" (from English "too much"), *masin* "machine." Users of pidgin often carry over habits of sound production from their native languages; *e.g.*, many Melanesian languages have *mb*, *nd* as variants of *b*, *d* between vowels, and thus Melanesians often pronounce *tabak* "tobacco" as *tambak*, and *sidaun* "sit" as *sindaun*.

Morphology. Grammatical categories—such as number, gender, case, person, tense, mood, voice—are almost absent from pidgin and creole languages, as from many other languages of the world. Pidgin is not, however, "devoid of grammar," as is often asserted. Melanesian Pidgin, for example, has three inflected parts of speech: pronouns, adjectives, and verbs. The suffix *-pela* added to pronouns makes plurals, in *mi* ("I, me") versus *mipela* ("we, us") and *yu* ("you" singular) versus *yupela* ("you" plural). Another suffix, *-pela*, serves as a marker for adjectives of one syllable, demonstratives, indefinites, and numerals: *naispela* "pretty," *dispela* "this," *sampela* "some," *wanpela* "one." Verbs having a direct object (expressed or implied) take the suffix *-im*, and verbs without this suffix are intransitive or passive: *e.g.*, *rausim* "eject, remove" versus *raus* "be out, come out." Other parts of speech—nouns, prepositions, adverbs, conjunctions—are invariable but are distinguished by the types of combinations in which they occur. In other varieties of Pidgin English, the specific criteria for distinguishing classes of linguistic forms are different, but the basic structure is similar. Chinese Pidgin nouns and pronouns, for example, can take the locative suffix *-said*, in *doksaid* "at the dock (docks)," *maisaid* "at my house," and the verb suffix *-em* forms passive participles, as in *dis tri blong spoilem* "this tree is rotten."

Phrases and clauses

Syntax. The basic types of combination—phrases and clauses—found in pidgin are the same as those of English; here again, however, many details of syntax are different. Melanesian Pidgin nouns are followed not only by possessive phrases (*haus bilong mi* "house of me, my house") but also by modifying nouns (*haus pepa* "house [for] paper, office"), verbs (*haus kuk* "house [for] cooking, kitchen"), adverbs (*man nogud* "evil man," *nogud* being an adverb meaning "undesirably"), and clauses (*man mi lukim em* "the man I saw"). In Chinese Pidgin, pronouns simply precede nouns to indicate possession (*hi fes* "his face"), and relative position is shown by a noun preceding an adjective (*Ning-Po mo fa* "further than Ning Po"). With third person subjects, Melanesian Pidgin predicates are normally preceded by the predicate-marker *i-*: *ol i-singaut* "they call"; *balus i-no kamap yet* "the plane hasn't arrived yet." Melanesian Pidgin has the clause type called "equational" (also found in such languages as Russian and Hungarian) in which no verb is present: *dispela kaikai i-gudpela* "this food [is] good." Chinese Pidgin, on the other hand, has the copulative, or linking, verb *blong* "be" with nouns and adjectives in the predicate (*e.g.*, *yu fut blong plenti sor* "your foot is very sore") but uses no verb with an adverb in the predicate indicating location: *tumuchi dast tebal tapsaid* "a lot of dust [is] on the table."

Lexicon. Since vocabulary is restricted (about 700 words in Chinese Pidgin, 2,000 in Melanesian), each word necessarily has a greater range of meaning than its English counterpart. The central meaning of Melanesian Pidgin *sori* is not "sorry" but "emotionally moved," as shown by its extension to "sympathetic, grateful, glad"; similarly, *dai* means "cease" ("die" is *dai tru* "stop for good"); and *stap* is "be located; remain; continue." For many concepts, pidgin uses phrases rather than single words: with *skru* "screw; joint" are formed *skru bilong arm* "elbow," *skru bilong leg* "knee," etc. Some pidgin words belong to different parts of speech from their English counterparts; *e.g.*, the Melanesian Pidgin preposition *belong* "of" and the Chinese Pidgin copulative verb *blong* "be" from English "belong." Non-English meanings of pidgin words often reflect native social structure, as when *papa* means "uncle," since a boy's maternal uncle rather than his father (*papa tru*) is primarily responsible for his upbringing in New Guinea. Speakers of English are often naïvely amused or shocked by certain shifts of meaning, as when *ars* "buttocks" is extended to mean "bottom (of anything), foundation, reason, cause, source"; for example, *ars bilong diwai* "the base of the tree," or *God i-ars bilong olgeda samting* "God is the source of all things." In the context of native society and attitudes, however, these concepts are not taboo and no stigma attaches to the words or their use.

Vocabulary

Non-English vocabulary. The proportion of vocabulary elements in Pidgin English derived from non-English sources is small. Of the approximately 2,000 words in Melanesian Pidgin, not over 10 percent are of non-English origin. Of these, perhaps half are Melanesian (such as *kiau* "eggbomb," *diwai* "tree," *malolo* "rest," *balus* "pigeon, airplane"), and one quarter were borrowed from German, such as *mark* "shilling," *tais* "pond," *langsam* "slow," and *beten* "prayer." The remainder are from various languages—a few from Malay (such as *karabau* "water buffalo") and three from Romance sources: *save* "know," *pikinini* "child," and *pato* "duck." The percentage of non-English elements in Chinese Pidgin is even smaller.

Structural changes from English

Restructuring. In the structural reduction from English to pidgin, the main grammatical characteristics have been kept (the part-of-speech system, the dichotomy between subject and predicate, the use of phrases functioning as single parts of speech), though often with different identifying features. The various kinds of Pidgin English are definitely English and Indo-European, not (as is often said) "native languages spoken with English words." A more sophisticated version of this latter theory is that indigenous vocabulary is simply replaced with new words by "relexification," with indigenous grammatical habits continuing. However, when new functional elements (*e.g.*, pronouns, inflectional suffixes, syntactic patterns) are also taken over, the process is one of complete language-substitution, involving replacement of grammatical structure ("regrammaticalization") as well. Nevertheless there have been extensive carryovers from non-English structural patterns because speakers of native languages have translated their own constructions into pidgin, especially in the early stages of its formation. For instance, in Chinese a numeral modifying a noun must be accompanied by a special word indicating a measure of quantity, called a "numeral-classifier," as in *sān ge rén*, literally "three piece man." This type of combination was reflected in older Chinese Pidgin *tufela man* "two men," *forpisi naif* "four knives," etc., with *-fela* for animate objects, *-pisi* for inanimate objects. However, in modern Chinese Pidgin, *-fela* has not survived, and *-pisi* has lost its independent status and become simply a numeral-suffix, as in *tupisi man* "two men." Similarly, the presence of separate pronouns for "we (excluding the hearer)" and "we (including the hearer)" in Melanesian languages has led to the establishment in Melanesian Pidgin of a parallel contrast between *mipela* "we (but not you)" and *yumi* "we (including you)."

Modern function. With the coming of modern civilization and technology to New Guinea and similar areas, pidgin has proved indispensable in education and political life. Earlier opposition to pidgin—partly on puristic, partly on anticolonialistic grounds—has proved unfounded.

In New Guinea and the Solomons, and in many parts of West Africa, pidgin is no longer a status language or imposed on the people by white colonialists, but is the people's own lingua franca, indispensable for communication and easier to learn than English, which is both more complicated and more foreign to them. If skillfully used, pidgin can serve as both a medium of instruction and a bridge to English. In any case, it is clearly destined to remain as an increasingly useful lingua franca, with already manifest prospects of extensive creolization and resultant permanence as the native language of ever larger groups.

BIBLIOGRAPHY. An extensive discussion and bibliography of the problems of pidgins and creoles is given in R.A. HALL, JR., *Pidgin and Creole Languages* (1966).

(R.A.H.)

Piero della Francesca

One of the most important artists of the Italian Renaissance, Piero della Francesca made his chief contribution through an exploration of problems of perspective, both as an abstract mathematical science and as a specific method of painting. His scientific approach was tempered, however, by his sensitivity to colour, his genuine feeling for nature, and his superb sense of formal composition. Although Piero's restrained and disciplined art did not greatly influence his contemporaries, it has received widespread recognition in the 20th century.

Contributions to perspective

Formative period. The documented facts of Piero della Francesca's life, which are few, permit a reasonably accurate reconstruction of his career and interests but not an exact chronology of his surviving paintings. He was probably born in the Tuscan town of Borgo Sansepolcro (modern Sansepolcro) around 1420. His father, Benedetto de' Franceschi, was evidently a tanner and shoemaker, prosperous enough for his son to become well educated and literate in Latin. Nothing is known about Piero's early training as a painter, though it is assumed that he was instructed by local masters who had been influenced by Sienese art.

In 1439 Piero worked as an associate of Domenico Veneziano, who was then painting frescoes for the hospital of Sta. Maria Nuova in Florence, where the early Renaissance style was beginning to flourish. There he probably studied the statuary of Donatello and Luca Della Robbia, the buildings of Filippo Brunelleschi, and the paintings of Masaccio and Fra Angelico, and he might have read a theoretical treatise on painting by the Humanist and architect Leon Battista Alberti. Undoubtedly, he would have been directed to these luminaries by Domenico Veneziano, whose own works demonstrate a Renaissance emphasis on colour and light as elements of pictorial construction. It was Piero's contact with the early Renaissance art of Florence that provided the foundation of his own style.

Back in Sansepolcro by 1442, Piero was elected to the town council. Three years later the Confraternità della Misericordia commissioned an altarpiece from him. This polyptych shows Piero's indebtedness to the Florentines Donatello and Masaccio, his fondness for geometric form, and the slowness and deliberation with which he habitually worked—for the altarpiece was not completed until 1462.

The Misericordia altarpiece

Periodic retreat to the provincial isolation of Sansepolcro seems to have been necessary for Piero's work. For the rest of his life he alternated between the calm of Sansepolcro and contact with the Humanistic life of the Renaissance in artistic and intellectual centres such as Ferrara and Rimini.

Around 1448 Piero probably worked in the service of Marchese Leonello d'Este in Ferrara, where he may have been influenced by northern Italian art. In 1451, at another northern Italian city, Rimini, he executed a splendidly heraldic fresco (*i.e.*, resembling a heraldic emblem in design) of "Sigismondo Malatesta Before St. Sigismund" in the Tempio Malatestiano, a memorial church that had been built according to the architectural designs of Alberti. Also to this early formative period before 1451 belongs "The Baptism of Christ."

Mature period. Piero della Francesca's mature style is revealed in frescoes painted in the choir of the church of S. Francesco at Arezzo. The decorations had been begun in 1447 by the elderly Bicci di Lorenzo, who died in 1452; Piero presumably was retained to complete the work shortly thereafter. The narrative cycle, depicting "The Legend of the True Cross," was completed by 1466. Its simplicity and clarity of structure, controlled use of perspective, and aura of serenity are all typical of Piero's art at its best. Contemporary with the Arezzo cycle are a fresco of the "Magdalen" in Arezzo Cathedral, the "Resurrection" in the Palazzo Comunale at Sansepolcro, and a "Madonna del Parto" in the chapel of the cemetery at Monterchi. In 1454 a burgher of Sansepolcro, Agnolo di Giovanni di Simone d'Angelo, commissioned an altarpiece for S. Agostino that Piero, characteristically, did not complete until 1469. The surviving panels of the altarpiece reveal Piero's interest of the creation of monumental human figures through the sculptural use of line and light.

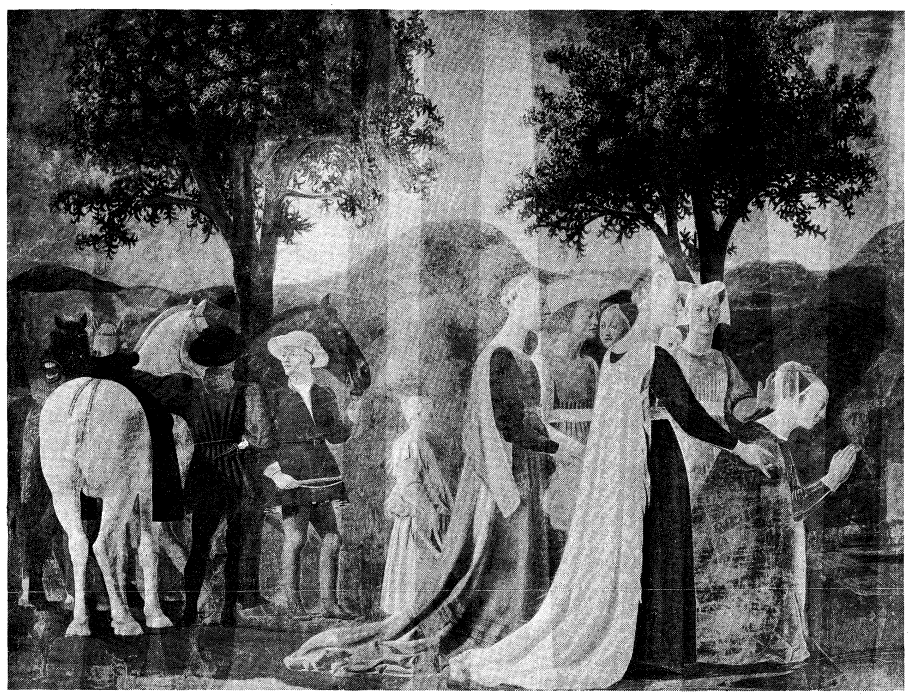
In 1459 Piero was in Rome to paint frescoes (now destroyed) for Pope Pius II in the Vatican. "St. Luke" (Sta. Maria Maggiore), executed at the same time, was probably done by assistants in the studio he had established in Rome. More fruitful was Piero's long association with Count (later Duke) Federico da Montefeltro, whose highly cultured court was considered "the light of Italy." In the late 1450s, Piero painted the "Flagellation of Christ," originally in the sacristy of the Cathedral of Urbino; its lucid perspectival construction contrasts with treatment of the subject wherein Christ is relegated to the background while three unidentified figures dominate the foreground. The content of the picture has become the focus of modern academic controversy. A famous diptych portrait of Count Federico and his consort Battista Sforza (Uffizi, Florence), probably commemorates their marriage in 1465. The paintings show Piero's respect for visual fact in the unidealized features of the Count and in the enchanting landscape backgrounds, which also indicate that he had discovered Netherlandish painting. The reverse depicts the couple in a triumphal procession accompanied by the Virtues. The Count reappears as a kneeling donor in an altarpiece from S. Bernardino, Urbino (now in the Brera, Milan). He, the Madonna and her child, and accompanying saints are placed before the apse (semicircular choir) of a magnificent Albertian church. The painting may have been a memorial to Countess Battista, who died in childbirth, and it has been dated between 1472 and 1474. The altarpiece is one of the most accomplished Renaissance presentations of forms in space and exerted a decided influence on the development of monumental devotional paintings in northern Italian and Venetian art.

Association with Count Federico

Last years. The last two decades of Piero's life were spent in Sansepolcro, where paintings, now lost, were commissioned by local churches in 1474 and 1478. In 1480 Piero became prior of the Confraternità di San Bartolomeo. Among the few extant paintings from this period are the harmonious "Nativity," in London, the "Madonna" from the church at Sta. Maria delle Grazie near Senigallia, now in Urbino, and an awkwardly constructed altarpiece in Perugia, "Madonna with Child and Saints." The "Annunciation" from that altarpiece, however, indicates that Piero's interest in perspectival problems remained keen.

In his old age Piero seems to have abandoned painting in favour of more abstruse pursuits. Between 1474 and 1482 he wrote a treatise on painting, *De prospectiva pingendi* ("On Perspective in Painting"), dedicated to his patron, the Duke of Urbino. In its range of topics and method of organization, the book follows Alberti and the ancient Greek geometer Euclid. The principal manuscript, in Parma (Biblioteca Palatina), was handwritten by the artist himself and illuminated by him with diagrams on geometric, proportional, and perspectival problems. A second treatise, the *De quinque corporibus regularibus* ("On The Five Regular Bodies"), written some time after 1482, follows Plato and Pythagoras in dealing with the notion of perfect proportions. The manuscript,

Scientific writings



"The Queen of Sheba Worshipping the Wood of the True Cross," fresco by Piero della Francesca, 1452–66. In the church of S. Francesco, Arezzo, Italy.
Alinari

again illustrated by Piero, is in the Vatican Library. *Del abaco* ("On The Abacus," Laurentian Library, Florence) is a pamphlet on applied mathematics.

Piero's fascination with geometry and mathematics is a corollary of his own art; his manner of theoretical expression owes much to his mentor Alberti and is analogous to that of his younger contemporary, Leonardo da Vinci; the rigour and logic of the arguments, however, are unique to Piero.

A reliable local tradition of the 16th century claimed that Piero was blind in his last years. If true, this must have occurred only after 1490 because several autographs from that year survive. Moreover, his will of 1486 refers to the painter as aged but sound of mind and body. Piero della Francesca died in Sansepolcro on October 12, 1492.

Piero did not establish a lasting tradition in central Italy. Luca Signorelli and Perugino, who are presumed to be his most important pupils, followed the examples of other masters. Although Piero's reticent art had little influence on the experiments of his great Florentine contemporaries, he enjoyed great fame for his scientific contributions. In 1497 he was described as "the monarch of our times of painting and architecture," and the biographer Giorgio Vasari gave him high praise two generations later. In the 20th century, Piero's career has been reconstructed and his position re-evaluated, giving proper credit to both the science and the poetry of his art.

MAJOR WORKS

The following list includes nearly all of Piero's surviving works—panel paintings except where stated otherwise.

"The Baptism of Christ" (*c.* 1440–45; National Gallery, London); "Polyptych of the Misericordia" ("Madonna della Misericordia"; 1445–62; Palazzo Comunale, Sansepolcro); "St. Jerome Penitent" (1450; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Sigismondo Malatesta Before St. Sigismund" (fresco, 1451; Tempio Malatestiano, Rimini); "The Legend of the True Cross" (fresco cycle, 1452–66; S. Francesco, Arezzo); panels from the altarpiece of S. Agostino in Sansepolcro (commissioned 1454, finished 1469): "St. Augustine" (Museu Nacional de Arte Antiga, Lisbon); "St. Michael" (National Gallery, London); "St. Simon the Apostle(?)" (Frick Collection, New York); "St. Nicholas of Tolentino" (Museo Poldi Pezzoli, Milan); and "Flagellation of Christ" (late 1450s; Galleria Nazionale delle Marche, Urbino); "Hercules" (*c.* 1460–66; Isabella Stewart Gardner Museum, Boston); "Resurrection" (fresco, *c.* 1463; Palazzo Comunale, Sansepolcro); "Portraits of Federico da Montefeltro, Duke of Urbino and of His Wife, Battista Sforza" (1465; Uffizi, Florence); "The Madonna with Child, Angels, Saints and Federico da Montefeltro, Duke of Urbino" (1470s; Brera, Milan); "Senigallia Madonna" (1470s; Galleria Nazionale delle Marche, Urbino); "Madonna with Child and Saints" (*c.* 1472–74; Galleria Nazionale dell' Umbria, Perugia); "The Nativity" (*c.* 1480; National Gallery, London).

BIBLIOGRAPHY. The earliest biography of the artist is VASARI'S "Life of Piero" (1550 and 1568), included in his *Lives of the Painters, Sculptors, and Architects,* trans. by A.B. HIND, vol. 2 (1963). ROBERTO LONGHI, *Piero della Francesca* (1927; 2nd ed., 1946; Eng. trans., 1930), is the definitive critical and scholarly monograph. A useful catalog of paintings is PIERO BIANCONI, *Tutta la pittura di Piero della Francesca,* 2nd ed. (1959; Eng. trans., *All the Paintings of Piero della Francesca,* 1962). KENNETH CLARK, *Piero della Francesca,* 2nd ed. rev. (1969), remains the standard account of the artist in English. It is a rare union of eloquence and scholarship.

(P.F.W.)

Pietism

Beginning as an influential religious reform movement in German Lutheranism in the 17th century, Pietism, with its emphasis on personal faith, spread to other countries and during the course of its history expanded its emphases to include a wide range of social and educational concerns.

NATURE AND SIGNIFICANCE

Evangelical Pietism was a religious movement historically set between two other movements, Protestant scholastic orthodoxy and rationalism (or the Enlightenment), standing in a partial cause-and-effect relationship to both. The 250 years that followed the Reformation (1575–1825) may be roughly divided into three periods in which a particular way of thinking and acting was dominant: 1575–1675, Protestant Scholasticism; 1675–1725, Evangelical Pietism; and 1725–1825, Rationalism.

Evangelical Pietism in some ways was a reaction to what James Hastings Nichols calls the "Secularization of the West." In other ways, it was a voice raised against secularism in the church. In protest to this development, Pietism sought to reassert the power of Christianity to change and shape men's lives and to revitalize the nature and role of the church. Though unsuccessful in its attempts to reverse the movement towards secularization,

it nevertheless released and directed powerful forces that in one form or another enjoyed revivals even into the 20th century.

General characteristics

It has been suggested that there are four general characteristics of Pietism. In the first place, Pietists were convinced that the essence of Christianity was an individually experienced inner change. In the second place, historic Pietism urged the idea of total commitment to the new life in Christ, a break with the old life and a disciplined practice of piety that would provide the opportunity for growth in personal holiness. Thirdly, Pietism gave renewed attention to the role of the Scriptures in Christianity. Not only was the Bible the source of Christian doctrines—this emphasis of Protestant Scholasticism was not to be denied—but it was the source of spiritual life. The Pietists at their best saw the Bible as the Holy Spirit's means for producing repentance, conversion, and the new life in Christ. The Bible mirrored one's sins, demanded the new birth, and guided the reborn in a new quality of life. At their worst, the Pietists used the Bible as a manual for a code morality and a magical book of oracles, a book to be opened at random for divine directives in decision making, a sort of biblical lottery. Fourthly, Pietism was inherently a protest movement; it was "protestant" in its opposition to the religious "establishment." This characteristic was evident in the stance of both Anglo-Saxon Puritanism and continental Pietism. Both forms of piety were in conflict with the representatives of the prevailing ecclesiastical, theological and religious status quo.

HISTORY

Thirty Years' War

Roots of Pietism. The Thirty Years' War (1618–48), though not a cause of Pietism, must be seen as one of the circumstances out of which the desire for renewal emerged. Although modern historical research has modified the exaggerated contemporary accounts of the war's effects, it is unquestioned that distress was wide-spread and profound. In some places the economy was reduced to barter, schools were closed, churches were burned, the sick and needy were forgotten. Not unexpectedly spiritual and moral deterioration accompanied the physical destruction. Drunkenness, sexual license, thievery, and greed were the despair of faithful pastors and earnest laymen.

During the war some notable signs of renewal began to appear. There was, for example, an awakened interest in the earlier devotional literature, some of which reflected the pious mysticism associated with such names as Johannes Tauler (*c.* 1300–61), Thomas à Kempis (*c.* 1380–1471), and other German, Dutch, and even Spanish authors. The mystical tradition had lived on into the Reformation century and found representatives in Kaspar Schwenckfeld (1489–1561), Valentin Weigel (1533–88), and Jakob Böhme (1575–1624). Although both Lutherans and Calvinists opposed these mystics, many of their religious and theological ideas were subsequently absorbed by orthodox theologians.

Influences from English Puritanism reached the Continent through the translation of works by Richard Baxter (1615–91), Lewis Bayly (1565–1631), and John Bunyan (1628–88). Most frequently read were Baxter's *A Call to the Unconverted*, Bayly's *The Practice of Piety*, and Bunyan's *Pilgrim's Progress*.

Dutch Pietism, influenced by the Englishman William Ames (1576–1633) whose *Medulla sacrae theologiae* (1623) and *De Conscientia* (1630) were basic textbooks for "federal theology" and Puritan casuistry in England and New England, was represented by Willem Teellinck, Johannes Cocceius, Gisbert Voetius, and Jodoeus van Lodensteyn. Impulses from these men became a part of the reform movement that had already appeared in German Lutheran circles and was to be known as "Reform Orthodoxy." Older historians of Pietism, notably Albrecht Ritschl, paid little or no attention to this reform phenomenon within Lutheranism. Ritschl saw Pietism as an alien mysticism uncongenial to the spirit of both Luther and the 17th-century theologians. More recent scholars (E. Benz, M. Schmidt, H. Leube, F.W. Kantzenbach) have exposed the Ritschlian prejudice and deepened the understanding of the role played by such representatives of "Reform Orthodoxy" as Johann Arndt (1555–1621) and Johann Dannhauer (1603–1666). The "pectoral heart theology" of these orthodox Lutherans found its highest expression and widest audience in the writings of Arndt, who, rather than Spener, can be called the "father of Pietism." His chief work, *True Christianity* (1610), was soon being read in countless homes. Although Arndt developed devotionally the *unio mystica* (mystical union), a 17th-century Lutheran doctrinal addition to the *ordo salutis* (order of salvation), the central Arndtian theme was not that of mystical union. Rather, he stressed repentance, regeneration, and the new life, and this was the very essence of Pietism.

Alongside the orthodox piety of the 17th century, one of the most significant contributions to spiritual renewal was to be found in the rich treasures of Lutheran hymnody. Examples from this classical period of church song are the works of Philipp Nicolai (1556–1608: "Wake, Awake" and "How Brightly Beams the Morning Star!"), Paul Gerhardt (1607–1676: "O Sacred Head Now Wounded," "O How Shall I Receive Thee," "Put Thou Thy Trust in God"); and Martin Rinkart (1586–1649, "Now Thank We All Our God").

Pietism in the 17th century. The various streams of concern for renewal converged in the life and work of Philipp Jakob Spener (1635–1705). In 1666, after having earned his theological doctorate at Strasbourg, he was called to be Senior of the clergy in Frankfurt am Main, where he was soon distressed by the conspicuous worldliness of the city. His sermons urged repentance and renewal, and each Sunday afternoon he held catechism classes for both children and adults. This led to efforts to revitalize the rite of confirmation, which, since the days of Martin Bucer, had been practiced in the Church of Hesse (Frankfurt).

Philipp Jakob Spener and the *collegia pietatis*

The origin of the so-called *collegia pietatis* (assembly of piety) has been traced to a sermon of 1669, in which Spener exhorted the laity to come together on Sunday afternoon not to drink, play cards, or gamble, as was the custom among Frankfurt's smart set, but to review the morning's sermon and to engage in devotional reading and conversation "about the divine mysteries." The next year, at the request of a few parishioners, such meetings were held each Sunday and Wednesday at Spener's home. Although some of the Frankfurt ministers, over whom Spener was superintendent, took a dim view of the *collegia pietatis*, the practice flourished and in time became a distinguishing feature of the movement. Those who attended the conventicles were soon called pietists.

In a relatively short time, Spener became a household name in Germany. Through his writings and extensive correspondence, especially with men in high places, Spener came to be called "the spiritual counsellor of all Germany." Most significant was the publication in 1675 of his *Pia Desideria* (*Pious Desires*). The book's first part reviewed the low estate of the church. He charged civil authorities, who since before the Peace of Augsburg (1555) were the *de jure* heads of the church, with irresponsible caesaropapism (doctrine of state control over church). He likewise flayed the clergy, many of whom were scandalous and self-seeking, often confusing assent to "true doctrine" with faith. The laymen, too, he claimed, were not blameless. Drunkenness must not be excused as a German peccadillo; prostitution, adultery, fornication, homosexuality, thievery, and assault must be rooted out lest people lose God's promised salvation. The second part of the work reminded readers of the possibility of better conditions in the church: ". . . we can have no doubt that God promised his church here on earth a better state than this." When the full number of heathen (Gentiles) had been brought in, God would even convert the Jews. But the fulfillment of these hopes was not to be achieved by sitting with folded hands. Part three, therefore, set forth a six-point reform program:

The six-point reform program

1. The Word of God—the whole Bible, not merely the pericopes (biblical texts used in a set sequence in worship services)—must be made known widely through public and private reading, group study (conventicles under the guidance of pastors), and family devotions.

2. There should be a reactivation of Luther's idea of the priesthood of believers, which included not only the "rights of the laity" but also responsibility toward one's fellowmen.
3. People should be taught that Christianity consists not only in knowing God's will but in doing it, especially by implementing love of the neighbour.
4. Religious controversies with unbelievers and heretics unfortunately may be necessary. If they cannot be avoided, they should be entered prayerfully and with love for those in error.
5. Theological education must be reformed. Professors must see that future pastors are not only theologically learned but spiritually committed.
6. Finally, preaching should have edification and the cultivation of inner piety as its goal.

Initially the *Pia Desideria* was received with enthusiasm and given wide acclaim. Some clergymen, however, felt threatened by the implications of the reform program's emphasis on the laity. Professors resented Spener's criticism of scholastic theology and advocacy of curricular reform. This was a brash intrusion on their own academic preserve. Spener's response was to emphasize more and more the *collegia pietatis*. Contrary to Spener's wishes, the conventicles in time became divisive and abrasively donatistic (referring to the Early Church heresy that held that priests must be morally righteous or the sacraments would not be valid), tending to develop into "little churches within the church" (*ecclesiolae in ecclesia*). In an attempt to stem separatism and other questionable attitudes, Spener wrote tracts that expounded the doctrines of the spiritual priesthood (1677) and ecclesiology (1684). In the latter, he argued that despite the faults of the church its teachings were not false, and separation from services and sacraments was wrong.

Spener's influence had spread widely by 1686. In many circles, not least among the nobility, he was praised and imitated. In other quarters, his emphases produced vigorous and, in many instances, unjust criticism. Weary of opposition and controversies, Spener accepted a call to be the court chaplain in Dresden, where he was soon disillusioned by the unresponsiveness and vulgarity of the court and the hostility of the pastors in this stronghold of orthodoxy. Two items of special significance from the Dresden period should be noted: (1) Here he wrote his *Impediments to Theological Study* (1690), which was hardly calculated to win friends at the famous Saxon University of Leipzig; (2) here, too, he made the acquaintance of a young instructor, August Hermann Francke (1663–1727), who was to become in a sense Spener's successor and the second great leader of Pietism.

By 1691 Spener welcomed a call from the elector of Brandenburg, who soon brought in other Pietists, opened his domain to persecuted French Huguenots, and made Berlin a strong spiritual centre, thus taking religious leadership away from rival Saxony. All of this was enhanced by the founding of a new university at Halle (1694), the theological faculty of which became, with Spener's and Francke's influence, the academic center of Pietism.

Spener's years in Berlin were not without bitterness. The conflict between Orthodoxists and Pietists had mounted to a high pitch. The theological faculty at Wittenberg, for example, charged Spener with 284 deviations and prayed that God would save "our Lutheran Zion" from the ravages of pietistic heresies.

During his last years, Spener collected and edited several volumes of his papers (*Theologische Bedencken*), continued his friendship with and support of Francke at Halle, and, significantly, served as a sponsor at the baptism of Nikolaus von Zinzendorf, who was to lead evangelical Pietism in a new direction. Spener died on February 5, 1705.

Meanwhile, Francke became the central figure of Pietism. During his student years at Leipzig, he had been engaged in group Bible study, being one of the organizers of a *collegium philobiblicum* (assembly of bible lovers), dedicated largely to the scholarly rather than devotional approach to the Scriptures. A religious experience in 1687 led Francke to make conversion—characterized by a severe penitential struggle and commitment to holy living—the norm for distinguishing the true Christians from unbelievers. Francke's Pietism, going beyond the spirit of Spener, came to stress a legalistic and ascetic way of life. Under Francke's leadership (he became professor in 1698), Halle became famous not only for its university but for the many "Halle institutions" that sprang up: an orphan asylum with affiliated schools, a publishing house and Bible institute, a Collegium Orientale Theologicum (Oriental College of Theology) for linguistic training of missionaries, and an infirmary that the medical faculty welcomed as compensation for the university's lack of a clinic. All of this gave to Halle and Franckean Pietism an energetic and activist character.

Pietism in the 18th century. *Central Europe and England.* One of Francke's institutions in Halle was the *paedagogium* (1698), which was intended for the education of boys whose well-to-do parents lived at a distance. Nikolaus Ludwig Count von Zinzendorf (1700–1760) attended the Halle boarding school from 1710 to 1716. Having been drawn earlier to Spener, his godfather, Zinzendorf was now greatly stimulated by Francke. As a 14-year-old lad he organized the "Order of the Grain of Mustard Seed," whose youthful members pledged themselves to reach out in ever-expanding love to "the whole human race."

By 1721 Zinzendorf had settled down on his estate (Berthelsdorf) near the Bohemian border, where he brought believers together in a nonseparatist *ecclesiola in ecclesia,* which denied the Halle Pietists' demand for penitential remorse as a mark of "heart religion." Zinzendorf formulated the slogan that came to play such a great role in the history of revivals: "Come as you are. It is only necessary to believe in the atonement of Christ."

A small band of Moravian exiles took refuge on his estate in 1722. Looking upon this event as an opportunity to realize his cherished project of "the Mustard Seed," he gave up his position in the Saxon civil service and welcomed other Moravian refugees. They, like Zinzendorf, had been primarily influenced by Pietism and had only a hazy idea that their ancestors were Hussites. Zinzendorf soon organized the colony, now called Herrnhut, into the community of the Bohemian Brethren. They were not to separate from the Lutheran Church of Saxony. They would attend services in the village church at Berthelsdorf and call upon the local pastor for ministerial acts; but they were to look upon themselves as "the salt" of the earth, an *ecclesiola* from which "heart religion" would be disseminated throughout Christendom. Under Zinzendorf's "superintendency" the Herrnhut Brethren became more and more a distinct church, the reborn Moravian Church, or Unitas Fratrum. Although Zinzendorf received a license as a minister in 1734 and three years later was consecrated bishop, he left Herrnhut under pressure in 1736, travelling in western Germany, England, and America. The chief centres of his missionary work in Pennsylvania were Germantown and Bethlehem. He returned to Herrnhut in 1749 and presided over the Church of the Brethren until his death (1760).

Herrnhut

The influence of the Moravians on the Evangelical Awakening in England was significant. By 1775 there were 15 Moravian congregations in England, and it was in one of these that John Wesley, founder of Methodism, had his famous "Aldersgate Street Experience" (1738) as he was listening to a Moravian preacher reading Luther's *Preface to the Romans*:

> . . . while he was describing the change which God works in the heart through faith in Christ, I felt my heart strangely warmed. I felt that I did trust in Christ . . .; and an assurance was given me that he had taken away *my* sins . . . ."

He allied himself with the Moravian society in Fetter Lane, London, and the same year journeyed to Hernnhut to learn at first hand about the people to whom he owed so much. Although Wesley later parted from the Moravians, his initial experience of saving grace in the company of the Brethren shaped the wide-reaching evangelical movement that associated the names of the two Wesleys (John and Charles) and George Whitefield.

Germany. A slightly different type of Pietism appeared in Württemberg, where Spener had established relations

with Swabian churchmen. Avoiding the extremes of Franckean Pietism, it accepted conventicles but opposed all temptation to separatism and sought an evangelical, as opposed to legalistic, sanctification of life in the congregations. Interested in academic theology and a scholarly study of the Scriptures, the leader of Württemberg Pietism, Johann Albrecht Bengel (1687–1752), was a pioneer in textual criticism and biblical, in contradistinction to systematic, theology. His *Gnomon Novi Testamenti* ("Interpretation of the New Testament," 1752) was widely distributed in the Lutheran and English world and gave a fresh approach to the Bible by its emphasis on *Heilsgeschichte* (the history of salvation).

Radical Pietism

Radical Pietism always lurked beneath the surface of Evangelical Pietism. The appeal to mystical and emotional experiences, the exaltation of psychic phenomena as evidences of the activity of the Holy Spirit, and the depiction of the church as "unholy Babylon" were common characteristics of Radical Pietism. Difficult to trace historically because of a tendency to flare up spontaneously, it can nevertheless be divided into two main forms. The first was a fanatic sectarianism in which ecstatic and visionary elements were dominant. A favourite doctrine was chiliasm (referring to the thousand-year reign of Christ at the end of history), in which the *apocatastasis* (the eventual salvation of all men) played a large role. Somewhat different but still under the first rubric were the "inspired congregations," whose inspiration was expressed in convulsive physical phenomena accompanied with *glossolalia*, "speaking in tongues." The second main form was "separatistic" or "nonchurch" and emphasized the "inner light." Because the "inner light" and human reason were often identified, the advocates of "Spiritual Pietism" tended to move toward Rationalism. Chief among these men were Gottfried Arnold (1666–1714), Johann Konrad Dippel (1673–1734), and Gerhard Tersteegen (1697–1769).

18th-century Pietism in Scandinavia, Russia, and America. *Denmark/Norway.* As in Germany, the age of orthodoxy in the Dano-Norwegian kingdom had its deeply spiritual side, which came to expression in men like Bishop Jens Dinesen Jersin (died 1632) and Holger Rosenkrantz (died 1642), both of whom taught the necessity of pious living. Also, as in Germany, the "reform orthodoxy" was evidenced in hymns, especially those of Thomas Kingo (1634–1703). Pietism, as such, arrived in Copenhagen at the turn of the century and was welcomed, strangely enough, by the unpietistic King Frederick IV. It was during his reign (1699–1730) that the royal chaplain, the German R.J. Lütkens, was able to give status to pietistic pastors and to win the King for the cause of missions in India. The King initiated a search for missionaries and, finding none in his domain (Bishop Bornemann causticalally described Danish students as being solely interested in luxury, drink, indolence, and harlotry), he turned to Germany, where Lütken's contacts brought about the connection with two young Halle-trained Pietists, Bartholomäus Ziegenbalg (1683–1719) and Heinrich Plütschau (1678–1747). Ordained at Copenhagen in 1705, these men became the founders of the famous Tamil mission at Tranquebar, India, and stimulated foreign mission interest among the Halle Pietists. To this period belongs the Christian work among the semipagan Lapps in northern Norway carried on by the Norwegian Pietist, Thomas von Westen. Another Norwegian, Hans Egede, became the pioneer missionary in Greenland. King Christian VI, known as the "Pietist on the throne," gave support to numerous pietistic causes: an orphan home and schools modelled after Halle, a missionary institute, and even conventicles (the 1741 decree permitted them only under pastoral leadership). The name of Erik Pontoppidan, court preacher at Copenhagen and later bishop of Bergen in Norway, was to have enduring significance largely because of his excellent exposition of Luther's catechism, entitled *Truth unto Godliness.* Virtually a national reader for many generations, especially in Norway, this "layman's dogmatics" combined Law and Gospel, orthodoxy and Pietism, in such a manner that its power persisted into 20th-century American Lutheranism.

Pietistic missionary activity

Comparable to Pontoppidan's range of influence was that of Hans Adolf Brorson, bishop of Ribe, a lyric poet whose hymns of the incarnation ("Thy little ones, dear Lord are we, and come thy lowly bed to see") and the ultimate glory ("Behold a host") stand out in the hymnody of Danish-Norwegian Christendom.

Sweden, Finland, and Russia. Original royal opposition to Pietism in Sweden was softened only after Francke personally visited King Charles XII on his Russian campaign. Meanwhile, Swedish students at Halle returned to their homeland imbued with Francke's ideas and practices. Following the defeat of Charles XII at Poltava in Russia (1709), thousands of Swedish prisoners of war were quartered in Siberia. Many sought comfort in religion under the leadership of a Swedish Pietist, J. Cederhielm. Correspondence with Halle and the writings of Francke and Arndt produced a strong pietistic movement in the prison camps from which only 5,000 of the original 30,000 captives were able to return to Sweden by 1724. The zealous returnees carried their pietistic convictions back to Swedish parishes. In a short time, both church and government looked upon Pietism as a threat to national unity. The result was the Conventicle Act of 1726, which retarded Pietism and held Swedish church life to conventional forms for the next century. Finns as well as Swedes had followed Charles XII to defeat. Those who returned from Russia were the apostles of a religious awakening. For a time the literature of Pietism was influential, but due to the Conventicle Act of 1726 (Finland was partially a Swedish domain), its role was somewhat limited.

Meanwhile, Pietism came to the Russian-occupied Baltic states where it experienced greater freedom than under the Swedes. In the foreign quarter of Moscow, inhabited mainly by German Lutherans, the work of Francke reached out to influence Peter the Great and some of his governmental ministers. Impulses from the 18th-century activity may have prepared the ground for 19th-century "Stundism," (devotion or Bible reading "hour") which was brought by Württembergian settlers to southern Russia.

America. In 1703 three pastors from New Sweden on the Delaware River ordained Justus Falckner, a Halle-educated Pietist, for service among the Dutch Lutherans in New York. Most of the Dutch Lutherans were of Pietist orientation, as were the many Germans from the Rhineland and Southern German valleys. These "Palatines," who settled in New York and Pennsylvania, and the famous refugee Salzburgers, who settled in Georgia, came via London where the Pietist court chaplain M. Ziegenhagen assisted them on their way to America. Accompanying the Salzburgers were two Francke-selected pastors, J.M. Boltzius and I.C. Gronau, who naturally shaped the spiritual life of the Georgia settlement. Zinzendorf's visit to America (1741–42) led to a clash between his type of Pietism and that of Halle, represented by Henry Melchior Mühlenberg (1711–87). The victory belonged to Mühlenberg, who became the organizing genius and spiritual leader, later called "The Patriarch of American Lutheranism." Meanwhile, the Great Awakening, especially in its "Middle Colony" manifestation among the Dutch Reformed, appeared among the Scottish-Irish Presbyterians. Here the revivals inspired by William Tennent (*c.* 1673–1745) prepared the soil for George Whitefield (1714–70), whose name, together with that of Jonathan Edwards (1703–58), was associated with the final phase of the Great Awakening.

Pietism in the 19th century. *Germany.* Before and for some time after 1815 an awakening occurred in Germany as a reaction against the Enlightenment. In philosophy, literature, and music the reaction found expression in German Idealism and Romanticism. In the congregations the reaction took the form of Pietism. Pietistic orthodoxism and biblicism continued to live on among "the quiet in the land." Some solitary thinkers with pronounced religious interests sought to preserve and awaken genuine Christianity and to point out the banality of the Enlightenment. Among these were Johann Georg Hamann (1730–88), a theologian who understood Luther's *theologia crucis* (theology of the cross) better than any 18th-

century person and a thinker given to brilliant paradoxical thought. Matthias Claudius (1740–1815) was another representative of the antirationalist mood of the dawn of the 19th century. Johann Friedrich Oberlin (1740–1826) mixed his biblicistic piety with a concern for social missions. J.A. Urlsperger (1728–1806) sought to promote piety by organizing the Christentumsgesellschaft ("A Society for Christianity"), the German counterpart of the British "Society for Promoting Christian Knowledge." Out of it grew the Basel Mission Society. G.C. Storr (1746–1804) and J.F. Flatt (1759–1821) represented the "Old Tübingen school" of biblical Supernaturalism.

Revival of Pietism

It was in such a climate that the revival of Pietism occurred. Many of the people involved in the awakening were not interested, at least in the beginning, in reviving former confessional differences. They were satisfied with being known as "Christians" or "evangelicals." But in time some of these new Pietists, influenced by Romanticism's admiration for the past, began to assert the need of linking their pietistic interests with the traditional confessional heritage of the church. True religion (Pietism), they argued, is really Lutheranism properly understood. Thus beginning with a renewal of heart religion (Pietism), they came to a neoconfessionalism. There were three discernible "schools" in this revival of Lutheranism. "The Repristination Theology" (*i.e.*, restoration of earlier norms) made the 17th-century orthodoxy normative for the interpretation of Luther and the confessions, and it fought the rising historical-critical approach to the Bible by affirming the verbal inspiration and inerrancy of the original manuscripts (autographs) of the scriptures. Ernst Wilhelm Hengstenberg (1802–69) was the champion of this school of "Old Lutherans." A second group, the Neo-Lutherans, felt that the Repristinationists, though not basically wrong, needed correction and improvement especially in their view of the church, the ministry, and the sacraments. These Neo-Lutherans ("high churchmen"), influenced by Romanticism, were the German counterpart of the Oxford movement in England. Chief exponents were August Vilmar (1800–68) and Wilhelm Löhe (1808–72), the latter having strong influence in American Lutheranism. The third group, the so-called Erlangen school, rejected Rationalism, Repristination, and Romantic catholicizing of the church. They asserted that theology must see the relationship of faith to history, thus providing a new setting for understanding both the Bible and the Lutheran confessions. Chief representatives were Gottfried Thomasius (1802–75) and J.C.K. von Hofmann (1810–77).

Denmark. The Spener-Francke tradition of Pietism survived the age of Rationalism in Denmark by being nurtured here and there by pietistic pastors and congregations, especially in rural Jutland. The rebirth of Danish spiritual life and the conquest of rationalism in the first half of the 19th century, however, came not from Pietism but from the religious and cultural impact of N.F.S. Grundtvig (1783–1872) and Søren Kierkegaard (1813–55). Both of these men were profoundly religious and at times may have sounded like Pietists, but neither had any essential sympathy for Pietism. Grundtvig was in fact definitely opposed to Pietism, while Kierkegaard, though stressing "the individual" and his existential involvement in the truth, found little time for Pietism as such. The actual renaissance of Pietism in Denmark was associated with the Inner Mission Society (established in the 1850s) and its leader Vilhelm Beck (1829–1901), who, deeply influenced by Kierkegaard's *Øieblikket* ("The Present Moment"), brought some of his emphases into the church that Kierkegaard so bitterly criticized.

Norway. Nineteenth-century Pietism in Norway may be seen in three episodes: (1) the life and work of Hans Nielsen Hauge (1771–1824), (2) the pietistic confessionalism of Gisle Johnson (1822–94), and (3) the conflict over liberal theology (*c.* 1875–1908). Hauge was a layman theologically untrained but at home in the Bible, Luther's catechism, and the works of Arndt, Pontoppidan, and Kingo. Converted in 1796, his sense of mission eventually produced a national revival. Hauge and Haugeanism, though sharply critical of the established church, became an example of conventicle Christianity within the framework of the state church. Arrested no less than ten times for violation of the long-neglected Conventicle Act of 1741, Hauge's final imprisonment lasted from 1804 to 1811. Although he thought of himself solely as a religious awakener, Hauge and his movement contributed to the socio-political revival in Norway through the influence of laymen who had been trained in an activistic type of Pietism. A characteristic feature of Haugeanism was its concept of a person's daily work as a divine calling. Imitating Hauge's example, many Haugeans became successful business men, shippers, and farmers. The second figure in Norwegian Pietism gave his name to a revival that occurred in the 1850s, the Johnsonian Awakening. Influenced by the German "Erlangen school," Johnson was joined on the theological faculty in Christiania (Oslo) by a staunch Hengstenbergian Repristinationist, C.P. Caspari, a converted German Jew. The Johnsonian Awakening, unlike the lay-oriented Haugean movement, was consciously directed toward pastors and church leaders. Like the Beckian revival in Denmark, it produced powerful lay organizations that promoted inner and foreign missions. The third phase of Norwegian Pietism was manifested in the conflict over theological liberalism during the last quarter of the 19th century. Increasingly, the university-oriented Norwegian intellectuals—clergy and lay—were drawn toward liberal positivism, historical relativism, and progressive optimism, the whole structure of which was based on natural science and Biblical criticism. The orthodox Pietists of the Johnsonian school led the attack on the liberal professors now dominating the theological faculty. By the turn of the century the idea of establishing a faculty independent of state control and supported by the faithful in the congregations was born. This was realized in 1908 when the Menighetsfakultetet (The Congregational Faculty) was created.

Sweden. Like Norway, Sweden was visited with a variety of pietistic movements in the 19th century. The first was militant revivalism in northern Sweden, where Moravian Herrnhuters interested in Lapland missions merged their enthusiasms with those of pietistic Lutherans and together were called the "Old Pietists." Lay conventicles, encouraged by some clergymen, emphasized devotional reading of the liturgy, the Bible, and Luther's and Arndt's sermons. The movement, called the Läsare ("The Readers"), soon came under attack, resulting in the emigration of a group under Erik Jansson to Bishop Hill, Illinois. A second revival in the first half of the century was associated with the name of Henrik Schartau (1757–1825), who was pastor and dean at Lund, Sweden. What distinguished Schartauism as a revival movement was its strong churchly character. It was catechetical, liturgical, orthodox, and anticonventicle. Yet its profound piety produced an awakening in southwest Sweden, the results of which remained into the 20th century. The third revival occurred toward the middle of the century under the leadership of Carl Olof Rosenius (1816–68), a lay preacher strongly influenced by George Scott, an English Methodist evangelist. Rosenian Pietism, or the "New Evangelism," as it was called, made much of "objective justification," appealing to sinners to "Come as you are." Again, as in Denmark and Norway, a healthy inner mission society was one of the fruits of revival, the National Missionary Society. Following the death of Rosenius, leadership came into the hands of Paul Peter Waldenström (1838–1917), whose subjective views of the atonement led to the formation of the Swedish Mission Covenant Church (1878).

Schartauism

Generally speaking, the two Swedish universities, Lund and Uppsala, represented high and low churchism respectively. The latter viewpoint influenced Parliament to allow the Church of Sweden its own Convocation (1865) with lay representation.

America. The great 19th-century German and Scandinavian immigration began in 1839–40. The first Germans to arrive were "Old Lutherans" from Prussia whose original pietistic impulses had given way to a high-church confessionalism of the Hengstenbergian and "New Lutheranism" line. Colonies of about 1,000 "Old Lutherans"

under J.A.A. Grabau settled in the vicinity of Buffalo and others in and around Milwaukee. They were the forerunners of the Buffalo Synod (1845). Saxon immigrants, under Martin Stephan and C.F.W. Walther likewise arrived in 1839 and settled near St. Louis to become by 1847 the Missouri Synod. Stephan had practiced conventicle Pietism in Germany and had influenced Walther and others in this direction. Walther and other Missouri Synod leaders later moved to a staunch confessionalism that left little room for conventional Pietism. The Norwegians, who also arrived in 1839, were almost entirely of the Haugean persuasion, one of their first leaders, Elling Eielsen (1804–83), being an extremely legalistic lay follower of Hauge. Most of the subsequent waves of immigrants were sympathetic to Pietism, the laity inclining towards Haugeanism, the clergy towards Johnsonianism. The Danish immigrants, fewer in number, eventually split over the question of Pietism. The anti-Pietists, or Grundvigians, were known as "the Happy Danes," while the pietistic, inner-mission disciples of Beck were denominated "the Sad Danes." The Swedish-Americans reflected "*Läsare*" and Rosenian Pietism initially, but after the Evangelical Lutheran Augustana Synod was formed in 1860 it soon began to evidence a churchly type of Pietism that perhaps could be traced to Schartauism.

Pietism in the 20th century. In central Europe Pietism in the 20th century was most prominent in Württemberg, where the churchly Pietism of Bengel persisted; in Norway, where the Haugean and Johnsonian lay-mission organizations continued to challenge and even to denigrate the national church; and in Finland, where the extreme Lastaedian Pietism (emphasizing lay confession and absolution) and the *glossalalia*-inclined followers of Ruotsalainen were tempered by the Swedish Pietism of Rosenius.

In America, Lutheran Pietism survived primarily among the descendants of 19th-century Scandinavian immigrants, especially in the Evangelical Lutheran Church (Norwegian), Augustana Synod (Swedish), United Evangelical Lutheran Church (Danish), and the Suomi Synod (Finnish).

THEOLOGY, PRACTICES, AND INSTITUTIONS

Although evangelical Pietism looked upon itself primarily as a practical reform rather than as a theological movement, certain theological positions differing from traditional orthodoxy were implicit in its teachings. In the first place, Pietism emphasized regeneration rather than justification by grace through faith. Pietists obviously did not deny the doctrine of justification but felt that overemphasis on its objective (forensic) character led to a weakening of the New Testament teaching concerning new birth and new life. Unconsciously, perhaps, this involved the Pietists in an alteration of basic theological categories. The Reformers, if not the later orthodox dogmaticians whom the Pietists were reacting against, looked upon justification in dynamic, existential, and relational terms. Justification was concerned with man's status, his existence, his relation to God. The Pietists, though assuming this, now shifted the focus from "existence" to "essence." Regeneration meant a change in being; man was a new creation. Theologically, this meant a move from the existential and relational category of justification to the ontological category of regeneration, the "new being" in Christ. This, in turn, modified the concept of the church. The church was no longer the congregation of those called (*ecclesia*) by Word and sacraments, but the association of the regenerate, the new beings. Thus the church, in the true sense, consisted of the *ecclesiola in ecclesia*, the kernels within the husks of corruption and decay. Within the *ecclesiola* of the reborn, each individual under the eye of his fellow Pietists was to occupy himself with growth in holiness, denial of worldiness, and conformity to the indwelling Christ. This "edification of the inner man" (Spener's phrase) led to a preoccupation with individualism, whereas the New Testament concept of edification was invariably related to the upbuilding of the congregation.

Emphases

The chief instrument for edification was the reading of the Bible. Pietists believed that in it one saw mirrored the destructive power of sin and death and the possibility of victory over the forces of evil by daily repentance and faith, prayer, and a life of love and obedience to the will of God. Despite the appearance of joyless and legalistic asceticism in some pietistic quarters, the concern for personal and social ethics was cultivated by this devotional use of the Bible. The Pietistic social concern, seen in the vast network of institutions of mercy, schools, Bible societies, and mission societies, was an unprecedented breakthrough. Some of the inner-mission institutions (for example, Bethel bei Bielefeld in Westphalia, Germany), even in the 20th century, are unique in both Protestant and Catholic institutional ministries to men's needs. Apart from some questionable attitudes that attended an unhistorical biblicism, Pietism made the Bible a more popular book than it had ever been before. Closely related to this was the development of a new theological discipline, biblical theology.

INFLUENCE OF PIETISM TODAY

In the religious world of the 20th century, one of the outstanding features of the post-World War I era was the emergence of biblical theology. In part, this new interest can be traced back to the pietistic preference for the study of Scripture rather than dogmatics and ecclesiastical tradition. It led to an intense scholarly attention to the biblical texts in the original Greek and Hebrew, an activity closely akin to the biblical interests of the Reformation age. This, in turn (to the dismay of the Pietists), led to the historical ("higher") criticism, which seemed destructive of biblical authority. But by the 20th century this same biblical concern had found a new approach, the theological understanding of the Scripture, which introduced the church again to what Karl Barth (1886–1968), the Swiss theologian who was the founder of what is called Neo-Orthodoxy, called "the strange new world of the Bible."

In the area of interconfessional cooperation, the Pietists were in large measure the pioneers of modern ecumenism. Men like Spener, and more especially von Zinzendorf, were far less interested in doctrinal differences that separated Christians than in the confession of Christ that united them. Thus a great simplification in dogmatics took place. Orthodox theology had no salvatory power: Christ alone was Saviour. It might be possible for one to err theologically and yet possess genuine trust (faith) in Christ, and where such faith existed, unitive rather than divisive forces were present. This was echoed with great clarity by Anders Nygren at the Lund Faith and Order Conference (1952): the age of "comparative ecclesiology" has come to an end; the starting point of ecumenism is Christological, that is, the commonality of faith in Jesus Christ.

Ecumenism

In the third place, Pietism reached into the 20th century by its destruction of the notion that one can be a Christian by being religious. "Religionless Christianity" had definite pietistic roots. Adherence to "religion" or the practice of religiosity was as abhorrent to Pietists as to Kierkegaard or Dietrich Bonhoeffer (1906–45), the German Lutheran theologian martyred by the Nazis. The imperative of personal decision and the involvement of the whole man was a tenet that did not originate in the last third of the 20th century.

Although the forms in which Pietism expressed itself in the 18th and 19th centuries will hardly be repristinated in the contemporary world, certain emphases of evangelical Pietism can be expected to reappear perennially because they grow out of what is believed to be the biblical witness to God's redemptive action in history. Among these are (1) the emphasis on religion as personal and experiential; (2) the call for total commitment; (3) the continuing reverence for the Bible as God's address to men; and (4) the noble discontent with, and willingness to protest against, forms without content, bodies without life, men and institutions without soul.

BIBLIOGRAPHY

General church histories (with sections on Pietism): H. HOLMQUIST and J. NØREGAARD, *Kirkehistorie*, vol. 2 (1949), one of the best general church histories with an historical and theological treatment of Pietism; K.S. LATOURETTE, *The Nineteenth Century in Europe: Protestant and Eastern*

Churches, vol. 2 of *Christianity in a Revolutionary Age* (1959), a detailed treatment of 19th-century Pietism in Germany, Scandinavian countries, and Baltic states; C.L. MANSCHRECK (ed.), *A History of Christianity*, vol. 2 (1964), a good introduction to Pietism with illustrative readings from Spener and Francke.

Selected primary sources: J. ARNDT, *True Christianity*, Eng. trans. by A.W. BOEHM (1712), and ed. by C.F. SCHAEFFER (1917), one of the classics of Pietism with a valuable historical introduction by the editor; P.J. SPENER, *Pia Desideria*, trans., ed., and with an introduction by T.G. TAPPERT (1964), the first English translation of Spener's best known work; excellent historical introduction and text annotations, and "On the Spiritual Priesthood of Believers," appendix in H.E. JACOBS, *A Summary of the Christian Faith* (1905), Spener's second best known work; represents his mature views on the relationship of the priesthood of believers to the special ministry of Word and sacrament.

Selected secondary sources: P. GRUENBERG, *Philipp Jakob Spener*, 3 vol. (1893–1906), the most comprehensive work on the life, thought, and influence of Spener; E. HIRSCH, *Geschichte der neuern evangelischen Theologie*, vol. 2 (1951), the best interpretation of the theology of Pietism; J.T. MCNEILL, *Modern Christian Movements* (1954), a brief but excellent analysis of Puritanism, Pietism, and Evangelicalism with a discussion of the similarities and differences between Puritanism and Pietism; A. RITSCHL, *Geschichte des Pietismus*, 3 vol. (1880–86), still the most comprehensive treatment of Pietism, although one-sided and generally unsympathetic; F.E. STOEFFLER, *The Rise of Evangelical Pietism* (1965), the first full statement in English of the early history of Pietism.

(E.C.N.)

Piezoelectric Devices

Piezoelectric devices, widely used for such purposes as converting sound into electrical impulses in sound reproduction, are based on the fact that certain crystals, when deformed by pressure, generate an electrical charge. The phenomenon is due to the electrically polar nature of the crystals; that is, to the fact that they contain positively and negatively charged ions, which separate when the crystals are subjected to stress. About two-thirds of the major classes of crystals can be made to yield an electric charge by the application of stress (Greek *piezo*, "press"). Asymmetrical crystals in particular have this property; the more symmetrical the crystal, the greater amount of stress is required to produce a charge.

Conversely, if an electric charge is impressed on such a crystal, it yields mechanical motion. One of the simplest crystals, zinc sulfide, illustrates how separation comes about. The unit cell shown in Figure 1 consists of two types of atoms: zinc atoms (dark circles), with a positive

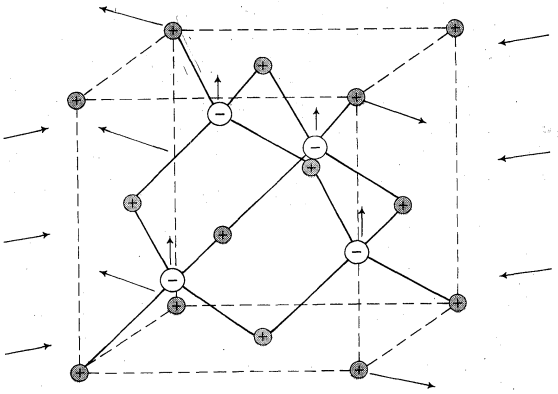


Figure 1: Unit cell of cubic zinc-sulfide crystal with positive zinc and negative sulfur atoms. Long arrows indicate applied shearing strain, short arrows the associated displacement.

charge, and sulfur atoms (light circles), with a negative charge. Compressional stresses do not produce the separation of the charge centres, but a shearing stress, indicated by the long arrows in the illustration, produces a motion of the sulfur atoms, indicated by the short arrows. This motion results in separation of the charge centres (polarization).

History. The piezoelectric effect was discovered experimentally by two brothers, Pierre and Paul-Jacques Curie, in 1880, while still young researchers at the University of Paris. By placing a weight on a crystal, they could observe a voltage on the surface of the crystal. They found this effect in a considerable number of crystals, including quartz, Rochelle salt, and tourmaline. The converse piezoelectric effect, elongation of the crystal with the application of voltage, was predicted theoretically by another investigator in 1881 and verified by the Curies in the same year.

Though Pierre Curie and his wife, Marie, later used the effect in quartz to study radioactivity, piezoelectricity in general remained a scientific curiosity until World War I, when the effect was used to produce acoustic waves in seawater to detect the presence of submarines (see SONAR).

Early applications

Further experimentation, during World War I, led to recognition of the reaction of the mechanical resonance of a vibrating crystal on its electrical behaviour and, hence, to frequency control of oscillators (see below *Frequency control*). Crystals soon found wide application as frequency-control elements in broadcasting and radiocommunication systems. During World War II, the U.S. alone used about 50,000,000 quartz-crystal elements.

Another major application for piezoelectric crystals was in the development of very selective electrical filter circuits; these are employed, for example, to keep undesired signals out of a radio receiver and have been widely used in coaxial-cable, radio, and submarine-cable communications systems (see below *Wave filters*). More than 2,000,000 crystal units a year are now employed for this purpose alone. Studies in the late 1960s made it possible to arrange the contacts on a crystal in such a manner that the desired degree of filtering could be obtained with a single crystal plate. These are called monolithic filters.

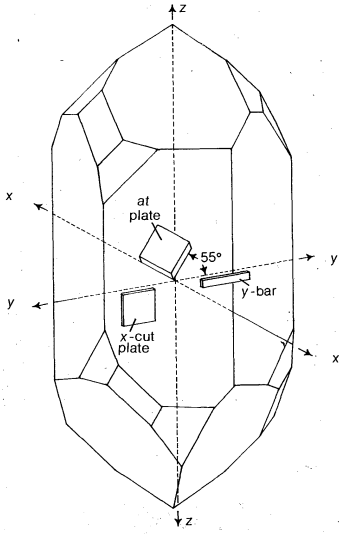


Figure 2: *Quartz crystal with three most important piezoelectric elements.* The *x*-cut is used for expansion or contraction in thickness direction upon the application of an electric field in the *x* direction; the *y*-bar expands and contracts in length direction when the field is applied in the *x* direction. The plate *at* shows shear in side faces when the field is applied in the thickness direction; it excels in low-temperature variation on resonance frequency.

The crystal Rochelle salt was an early discovery. Its properties were originally called Seignette electricity (in honour of its French discoverer) by Soviet researchers, who led in its study in the early 1930s. Since 1945, the term ferroelectricity has been more generally accepted. About 1940 the ferroelectric nature of the ceramic material barium titanate was recognized, and it was soon learned that treatment in a high electric field at temperatures near the Curie point (the temperature below which the material becomes ferroelectric) caused the material

Piezoelectric ceramics

to become strongly piezoelectric when cooled. This process is now known as polarization. Other ceramics, such as lead titanate–zirconate and sodium potassium niobate, with the same properties and a higher Curie point, have largely displaced other materials in uses in which relatively large amounts of power are transformed from electrical to mechanical energy or vice versa.

Piezoelectric crystals. Though Rochelle salt was one of the first and most widely used natural piezoelectric crystals, its sensitivity to temperature eventually caused its replacement, except in phonograph pickups and a few other applications, discussed later.

Because it can be cut in different ways for different applications and because it can produce stable oscillations, which are virtually independent of temperature changes, quartz is the most widely used piezoelectric crystal. The temperature stability of quartz crystals is invaluable in two-way radios. A two-way-radio-equipped policeman, for example, who moves from a warm building to a winter street, exposes his radio to extremes of temperature. Yet the quartz crystals in its transmitter and receiver continue operating exactly on frequency.

Piezoelectric transducers. Transducers are devices that convert one physical effect into another; for example, sound waves into alternating electrical current and vice versa. Piezoelectric devices are used in two types of transducers, nonresonant and resonant, depending on whether the device operates at its normal frequency of vibration (resonant) or is made to operate at a different frequency (nonresonant). Each has special applications.

Nonresonant devices. Nonresonant piezoelectric devices have been used to determine the piezoelectric properties of a crystal; that is, to measure the amount of charge generated by an applied stress or to measure the mechanical movement for a given applied electric field. Such devices are not useful for static measurements, since the charge generated leaks off quite rapidly, but they have been successfully applied in measuring rapidly changing states of motion, as with gauges for measuring the forces generated in an explosion or blast, in which case their natural resonant frequency is far above the frequencies being measured.

Phonograph pickups and microphones

Other nonresonant dynamic applications include phonograph pickups and acoustic microphones. The largest use in terms of manufactured units is the phonograph pickup, in which the piezoelectric device transforms the motion of the needle into an electrical voltage that can drive an amplifier. A typical phonograph-crystal mounting is shown in Figure 3. A lateral motion of the stylus causes a torque (twisting stress) on the two-element (bimorph) unit. The voltage generated is typically one volt for a 5-to-10-gram stylus pressure. For hill-and-dale recording (in which the motion of the needle is vertical), the piezoelectric unit is made up of a bimorph unit obtained by cutting a crystal of Rochelle salt into a pair of expander bars, in which the piezoelectric effect occurs lengthwise in the bars. Expander bars can also be obtained from piezoelectric ceramics; these materials have the advantages of being free from temperature and atmospheric-humidity problems and are easily adapted to lightweight construction. Some lead titanate, lead zirconate compositions are used for this application.

By courtesy of Clevite Corporation

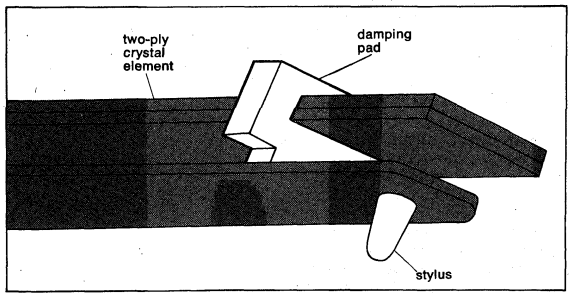


Figure 3: Rochelle salt phonograph pickup.

Another use for piezoelectric pickups is in microphones. In such devices, a diaphragm, actuated by sound waves, is mechanically coupled to the crystal in such a way as to twist or bend it, thus producing an electrical output. Another construction consists of two concentric ceramic pieces with a metallic hemisphere in the middle acting as an air-wave-pickup device. The outside edge of the ceramic is mounted between paper shims, slips of paper that limit the motion but not the bending. Such devices were not extensively employed in the early 1970s, but, with the advent of integrated circuits, which provide amplification very inexpensively, they promised to be of greater importance.

Resonant devices: measurements. Dynamic resonance in piezoelectric crystals is probably their most widely applied characteristic. Freely vibrating crystals are used for frequency and time standards and for very selective electrical wave filters. Resonant crystals and ceramics are also used as transducers to generate mechanical waves in solids and liquids.

To obtain the desired action, piezoelectric elements are cut to closer tolerances with respect to dimensions and to the natural crystal axes. For precision quartz-crystal oscillators, dimensions must be accurate to within one part in 1,000,000,000.

An important characteristic of the crystal is a factor that represents the amount of sharing of electrical and elastic energy through the piezoelectric effect, called the electromechanical coupling factor. This can be determined directly from the resonance measurements. It is an important factor in wave-filter and transducer applications, since it determines the maximum bandwidth, or frequency range, for which the device transmits efficiently. This bandwidth is usually expressed as the ratio of the frequency range to the mean frequency of the device. For quartz, the coupling is about 10 percent, but for many piezoelectric ceramics it may be as high as 70 percent. Ceramic, therefore, is used when wide frequency bands are required, while quartz is particularly applicable when narrower bands but very selective devices or very stable frequencies are required.

Resonant devices: applications. Some of the resonant frequency devices that have received wide application are sonar transducers; delay line transducers; high-amplitude devices for use in testing properties of materials, in drilling, and in welding; and freely vibrating crystals used as oscillators or filters.

Underwater sound

One of the largest applications of resonant transducers is in the production of underwater sound equipment (see SOUND). They are particularly adapted to this application because the mechanical impedances of the devices are well matched to the mechanical impedance of seawater; that is, maximum energy transfer takes place from the device to the water. Most such transducers are of the ceramic type, with lead titanate–zirconate being the most used. Bars of ceramic in the form of a mosaic are mounted on steel backing plates. Power outputs as high as one megawatt have been achieved with some installations.

Similar devices are finding increasing use as ultrasonic cleaning devices, operating in the frequency range of 10 to 100 kilohertz. For the high-frequency range, ceramics vibrating in their thickness mode (the plate becomes alternately thicker and thinner as it vibrates) are usually employed. Curved shapes can also be employed for focussing or concentrating the energy in small areas to produce intensities high enough to remove scale (corrosion) and even to cause plastic flow in metals (see also ULTRASONICS AND INFRASONICS).

Another useful application of piezoelectric transducers is in delay lines; *i.e.*, devices for slowing down or introducing a delay in a signal path. An acoustic wave travels through most materials much more slowly than do electromagnetic or radio waves. A piezoelectric transducer converts the electrical signal to an acoustic wave; this wave travels through some such material as quartz and is delayed in the process; the signal is then reconverted to an electrical signal by another transducer. Delay lines of this type are used in certain kinds of radar systems in which only moving objects are displayed and in computers.

Another use for a resonant transducer is to generate high strains and produce large motions that can be used for processing applications. Figure 4 shows such a transducer, usually made of a piezoelectric ceramic, coupled to a mechanical transformer to produce large motions and strains. Such a device has been used in studying the mechanisms producing metal fatigue.

From *Journal of Applied Physics* (November 1968)

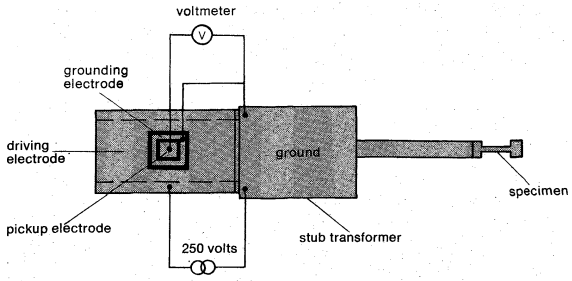


Figure 4: Ceramic resonant transducer for converting a 250-volt ac input to mechanical vibrations, and a mechanical (stub) transformer for coupling these vibrations to the load (specimen; see text).

By mounting a suitable tool on the end of the transformer and by using abrasives, a hole of any shape can be cut in brittle metals and ceramics. Wear studies carried out by the use of tools on the end of the transformer, have shown that most of the wear is caused by tangential motions and is least in polymers and rubbery materials. Large tangential motions coupled with a normal force have also been used for welding thin layers of metals and thin layers of plastics.

Application of freely vibrating crystals. In addition to driving vibrating devices, crystals vibrating freely have been used widely in frequency control of radio transmitters, wave filters in telephone-communications applications, and experimental work on light modulators for optical communications. The first two applications usually employ quartz crystals, although some newer crystals, such as lithium tantalate and lithium niobate, are receiving consideration.

Frequency control. Frequency-control devices such as are required for radio and TV broadcast stations and most radio communications transmitters mainly require low-temperature coefficients and high values of quality factor, or Q (that is, they maintain their oscillations with very little energy input), which are admirably met by the quartz crystals. Specially cut plates, if prepared very accurately, are independent of temperature variations over a wide range. The Q of the crystal is also high, and

Crystal-controlled clocks

such crystals are widely used to control frequency and time standards. Crystal-controlled clocks give extremely accurate time measurement by applying an electric field along the thickness direction, generating transverse oscillations. For 8 megahertz (8,000,000 cycles per second) resonant frequency, a thickness of 0.008 inches (0.02 centimetres) is required. The driving circuit is forced to oscillate at an exact frequency, controlled by the quartz crystal. Without temperature control, a stability within 0.001 percent (one second a day) can be maintained. By holding the temperature constant to within 0.001° C with temperature-controlled ovens and using carefully prepared quartz crystals, stabilities of one part in 1,000,000,000 have been held for many months. Such clocks are more constant than the rotation of the earth itself. Other uses for crystal-controlled oscillators are in broadcasting transmitters, military communication equipment, and on a distance-measuring device called loran (see NAVIGATION). Frequencies from one kilohertz to 100 megahertz can be used by employing various crystal cuts, shapes, and thicknesses.

Wave filters. Freely vibrating crystals are used to make selective wave filters that pass a given frequency band (for example, a voice frequency band of 4,000 hertz for telephone circuits) and eliminate all other frequencies. This permits superimposing a number of telephone conversations on a carrier signal for long-distance transmission. The width of the frequency pass band for a typical quartz crystal is 0.4 percent of the resonant frequency. There are techniques by which bandwidths up to 10 percent are obtainable. Because of the high stability of quartz, it is possible to space the voice bands with very little protective frequency range between them.

Light modulators. Piezoelectric crystals are also employed as light modulators by utilizing the electro-optic effect; that is, a change in the optical properties of the crystal when voltage is applied to it. While this effect depends on the change of the dielectric constant with applied field rather than on the motion of the crystal due to the piezoelectric effect, the constants for the electro-optic effect satisfy the same mathematical expressions as the piezoelectric effect, and, hence, light modulation occurs in all piezoelectric crystals.

This system has achieved great importance since the invention of the laser, which can produce a high-intensity light beam. By using an electro-optic modulator, a large number of communication channels can be transmitted over a laser light beam. The crystal most often used for electro-optic effects is potassium dihydrogen phosphate, or the heavy-water modification (water formed from the heavy isotope of hydrogen deuterium), which has a larger electro-optic effect. In this crystal, the light has to be transmitted in the same direction as the applied field to obtain the greatest effect. Certain cubic crystals, such as copper chloride, zinc sulfide, lithium niobate, and lithium tantalate have the advantage that the light direction and the applied field direction are at right angles to each other. Because this is a rapidly developing field, new crystals are constantly being investigated.

Specialized applications. Piezoelectric materials, particularly the ceramics, have a number of specialized applications, stemming from the high voltage that can be produced by a medium strain. For example, a force of 1,000 pounds (455 kilograms) applied to a block of quartz one centimetre (0.4 inch) on each side can produce an open-circuit voltage of 25,000 volts. A lower voltage but more total energy can be obtained from a ceramic material. This simple technique for generating voltage leads to many uses for ceramic transducers: *e.g.*, in hand grenades, to produce an electrical impulse to trigger an explosive on contact with a solid object; to produce the voltage for the spark plugs on small gasoline engines; as high-voltage generators; as prods for herding cattle and sheep; and for many other high-voltage applications. The reverse effect—that is, the production of a large motion in a ceramic material with an easily obtainable alternating voltage—has been utilized in pumps for oil burners and as generators of ultrasonic sound for burglar alarms, door openers, and ultrasonic cleaning and drilling. High-frequency ceramic units have been used for intermediate-frequency filtering on radio receivers—usually at 455,000 hertz (cycles per second).

Many more applications are receiving attention, and the search for new and improved piezoelectric and electro-optic crystals is part of the expanding field of solid-state physics. The search has already produced a number of new materials that may be expected to take the place of a number of the crystals mentioned here. All of these crystals require specialized methods for growing large-sized specimens.

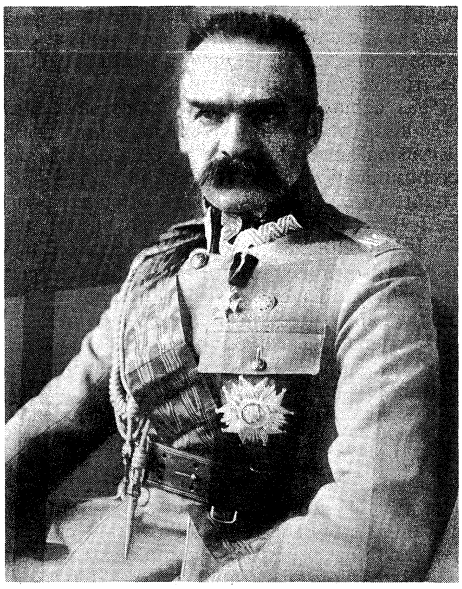
BIBLIOGRAPHY. W.G. CADY, *Piezoelectricity*, rev. ed., 2 vol. (1964), a comprehensive text on piezoelectric theory and applications; W.P. MASON, *Electromechanical Transducers and Wave Filters*, 2nd ed. (1948), on the application of piezoelectric crystals to transducers and wave filters, *Piezoelectric Crystals and Their Application to Ultrasonics* (1950), theory and application of piezoelectric crystals to ultrasonics, *Physical Acoustics and the Properties of Solids* (1958), on the use of ultrasonic methods, principally piezoelectric transducers, in the determination of mechanical properties, and *Crystal Physics of Interaction Processes* (1966), a phenomenological description of crystal properties, including piezoelectricity; R.A. HEISING (ed.), *Quartz Crystals for Electrical Circuits* (1946), a description of all the manufacturing processes used in producing quartz crystals; *Institute of Radio Engineers Standards on Piezoelectric Crystals* (1949, 1957, 1958, 1961).

(W.P.Mn.)

Piłsudski, Józef Klemens

Revolutionary and man of action, Józef Piłsudski was one of the greatest and probably the first of the Polish public figures of the period that saw the rebirth of the Polish state in the 20th century. Poland had been partitioned in the 18th century among Russia, Prussia, and Austria; its revival after World War I stemmed partly from the efforts of two leaders—Roman Dmowski, the cautious statesman, and Józef Piłsudski.

Culver Pictures



Piłsudski.

Early political activities

Piłsudski was born at Żułów on December 5, 1867, the second son of an impoverished Polish nobleman. His mother, *née* Maria Billewicz, inspired him with hatred for the Russian imperial regime, which was treating the Poles with great harshness after their insurrection of 1863. On leaving the secondary school in Wilno (modern Vilnius), Piłsudski studied medicine at Kharkov in 1885 but was suspended as politically suspect in 1886. Returning to Wilno, he consorted with young Socialists and tried to read Karl Marx's *Kapital*, but its abstract argument proved uncongenial. Arrested in March 1887 on a false charge of plotting the assassination of the tsar Alexander III, he was banished to eastern Siberia for five years.

Piłsudski returned in 1892, determined to prepare an insurrection and to work for the re-establishment of Poland's independence. He joined the newly founded Polish Socialist Party (PPS), of which he soon became a leader. He started a clandestine newspaper, *Robotnik* ("The Worker"), in Wilno. In July 1899 he married, in a Protestant church, the beautiful Maria Juszkiewicz, the divorced wife of a Polish civil engineer, and moved to Łódź, where he continued to edit and print his paper.

In February 1900 he was incarcerated by the Russians in the Warsaw citadel. He feigned insanity so successfully that he was transferred to a military hospital in St. Petersburg, from which he escaped in May 1901. He took refuge in Cracow in Austrian Poland, but in April 1902 he was back in Russian Poland looking after the party organization.

When the Russo-Japanese War broke out in February 1904, Piłsudski went to Tokyo to solicit Japanese assistance for an insurrection in Poland. He had been preceded by Dmowski, his rival in the patriotic movement, who had told the Japanese that Piłsudski's plan was impracticable. The two Polish leaders agreed to disagree. Piłsudski returned clandestinely to Russian Poland to help direct the revolutionary movement that was spreading throughout the empire. After the Russian revolution was put down late in 1905, a split occurred within the PPS: the left wing, which proposed to delete from the party's program the stipulation that its main aim was an independent Poland, broke with Piłsudski's group, which insisted on that stipulation.

Aware of the Russian Empire's structural weakness and foreseeing a European war, Piłsudski concluded that it was imperative to organize the nucleus of a future Polish army. In 1908 he formed a secret Union of Military Action—financed with a sum of money stolen from a Russian mail train by an armed band led by Piłsudski himself. In 1910, with the help of the Austrian military authorities, he was able to convert his secret union into a legal Union of Riflemen, actually a school for Polish officers. At a meeting of Polish sympathizers in Paris in 1914, he declared that war was imminent and that

Attempts to organize a Polish army

> . . . the problem of the independence of Poland will be definitely solved only if Russia is beaten by Austria-Hungary and Germany, and Germany vanquished by France, Great Britain and the United States; it is our duty to bring that about.

World War I justified Piłsudski's prediction. Until 1916 the three brigades of the Polish Legion, technically under Austro-Hungarian command, distinguished themselves against the Russians. On November 5, 1916, Germany and Austria-Hungary, short of manpower, proclaimed the independence of Poland, hoping that Polish divisions could be deployed on the eastern front so that German divisions could be moved to the west. Piłsudski, appointed head of the military department of the newly created Polish council of state, accepted the idea of a Polish army on condition that it be part of a sovereign Polish state. His position was unexpectedly reinforced by the Russian Revolution of March 1917. The German government, however, refused to bind itself as to Poland's future, demanding instead that the existing Polish units should swear "fidelity in arms with the German and Austrian forces." Piłsudski, refusing to comply, was arrested in July 1917 and imprisoned in Magdeburg.

Released after the German collapse in the west, Piłsudski arrived in Warsaw on November 10, 1918, as a national hero. Four days later he was unanimously accepted as head of state and commander in chief of the Polish army. From that moment he ceased to be the man of a party, though his main support came from the left and from the centre; the right saw its leader in Dmowski, who had been heading the Polish National Committee in Paris and was now appointed by Piłsudski to be Poland's first delegate at the peace conference.

Piłsudski devoted himself to protecting Poland against the Russian Red Army, which was trying to fight its way into Germany in order to consolidate the revolution there. He led the Polish forces far to the east, occupying large areas that had belonged to Poland before the 18th-century partitions. He envisioned a federal state comprising Poles, Lithuanians, and Ukrainians, whereas Dmowski argued that these areas should simply be incorporated within a unitary Poland. In 1920 a counteroffensive by the Red Army forced the Poles to retreat westward almost to the suburbs of Warsaw, but Piłsudski, made marshal of Poland on March 19, conceived and directed a manoeuvre that in August brought victory to Poland.

After the adoption of a democratic constitution and a new general election, Piłsudski transmitted his powers on December 14, 1922, to his friend Gabriel Narutowicz, the newly elected president of the republic, who two days later was assassinated. Stanisław Wojciechowski, another of Piłsudski's old colleagues, was next elected president, the marshal agreeing to serve as chief of the general staff. When a right-wing government assumed power, Piłsudski resigned his post on May 29, 1923, and went into retirement at Sulejówek, near Warsaw, with his second wife, *née* Aleksandra Szczerbińska, and his two daughters.

Piłsudski became disillusioned with the working of the parliamentary system, and on May 12, 1926, during a time of economic depression, he marched on Warsaw at the head of a few regiments. Wojciechowski and his government resigned two days later. The parliament elected Piłsudski president of the republic on May 31, but he refused the honour, and another of his old friends, Ignacy Mościcki, was elected instead. In the new government Piłsudski assumed the ministry of defense, which he held

till his death. During the ensuing years he was the major influence behind the scenes, especially in the field of foreign policy.

With few exceptions, Piłsudski's former Socialist friends abandoned him and joined a centre–left coalition, which in the summer of 1930 started a mass campaign to overthrow his "dictatorship." Piłsudski's reaction was ruthless; to "cleanse" political life, he had 18 party leaders arrested and imprisoned in the fortress of Brześc. Though all of them were subsequently released, and their political parties were not dissolved, the country was ruled by Piłsudski's men. The most prominent among them was Col. Józef Beck, Piłsudski's former *chef de cabinet,* who became deputy foreign minister in December 1930 and foreign minister in November 1932.

When Hitler came to power in Germany on January 30, 1933, Piłsudski sent to Paris a secret emissary, Count Jerzy Potocki, his former aide-de-camp, to find out whether France, Poland's ally from 1921, would cooperate in joint military action against Germany. The latter had been openly rearming in flagrant violation of the Treaty of Versailles. The French leadership answered in the negative, and consequently Piłsudski accepted Hitler's suggestion of a ten-year German–Polish nonaggression agreement (January 24, 1934). To show that Poland's intentions were above suspicion, Beck was sent to Moscow in February, and the existing Soviet–Polish nonaggression treaty was prolonged to December 31, 1945. Later, Hitler repeatedly suggested a German–Polish alliance against the U.S.S.R., but Piłsudski took no notice of the proposal; he also declined to meet with Hitler. Piłsudski sought to gain time, believing that Poland should be ready to fight when the necessity arose. Such were the last instructions he gave to Beck. Shortly afterward, he died in Warsaw on May 12, 1935, of cancer of the liver. He was buried in a crypt of the Wawel Cathedral in Kraków, among Polish kings.

A romantic revolutionary, a great soldier without formal military training, a man of rare audacity and will power, as well as of great insight into European politics, Piłsudski was nevertheless poorly equipped to rule a modern state. He left Poland undeveloped economically and with an army that was ready to fight heroically but was doomed because of its composition and inadequate armament.

BIBLIOGRAPHY. The best biography of Piłsudski in English is W.F. REDDAWAY, *Marshal Pilsudski* (1939). Two authoritative works in French are J. DE CARENCY (K.M. SMOGORZEWSKI), *Joseph Pilsudski, soldat de la Pologne restaurée* (1929), and *Joseph Pilsudski et les activistes polonais pendant la guerre,* 2 vol. (1930–31). There is also a biography of him by his second wife, A. PILSUDSKA, *Pilsudski: A Biography by His Wife Alexandra* (1941). Of political interest is R. MACHRAY, *The Poland of Pilsudski* (1936). For some of Pilsudski's writings, see D.R. GILLIE, (trans. and ed.), *Joseph Pilsudski: The Memories of a Polish Revolutionary and Soldier* (1931).

(K.M.S.)

Pindar

The tradition of Greek choral lyric culminated during the 5th century BC with the odes of Pindar. These are not easy to evaluate and appreciate, but it is still more difficult to comprehend and assess the poet who composed them. Even to his contemporaries, Pindar must have seemed an aloof and somewhat enigmatic figure. His origins were aristocratic, so that he would more naturally have been a patron of poetry than himself a poet; he came from a part of central Greece that had made a relatively small contribution to literature and the arts. Because ancient biographies, dating from the 4th and 3rd centuries BC, are highly unreliable and of little help, any approach to the poet must be made through his poetry, and this involves a consideration of the historical setting, the religious usages, and the literary conventions that helped to shape his art. A modern reader needs a sympathetic insight into the nature and traditions of Greek aristocratic society before he can begin to understand how Pindar's subject matter—victory in an athletic contest or in a chariot race—could inspire poetry characterized by high seriousness and deep feeling. Pindar cannot, indeed, speak across the centuries with the directness of Homeric epic poetry or Sophoclean tragedy, but he does create, with disciplined mastery of a sophisticated and complex art form, a choral lyric of unsurpassed splendour and sustained nobility.

Early training. Pindar was born in 518 BC, or possibly 522 BC, at Cynoscephalae, a hamlet in Boeotia (central Greece), in Theban territory. He was of noble birth, possibly belonging to a Spartan family, the Aegeids, though the evidence for this is inconclusive. His parents, Daiphantus and Cleodice, survive only as names; his uncle Scopelinus, a skilled flute player, doubtless helped with Pindar's early musical training. The family possessed a townhouse in Thebes (to be spared by express command of Alexander the Great in the general destruction of that city by the Macedonians in 335 BC). Such a background would give Pindar a ready entrée into aristocratic circles in other Greek cities, where his manifest gifts as a poet might be valued more highly than in his native Boeotia, which had little enough to offer in the way of precept and encouragement. (Two contemporary Boeotian poetesses, Myrtis and Corinna, were of a very different tradition, more primitive and essentially feminine, though Corinna is reported to have criticized the lushness of Pindar's early style, advising him to "sow with the hand, not with the full sack.") It was natural, then, that Pindar should be sent to neighbouring Athens to complete his training and education. Athens and Thebes were seldom on neighbourly terms, but relationships between noble families transcended such difficulties. Pindar's horizons would have widened in the exciting atmosphere of a city beginning to realize its destiny, poised as it was on the verge of political and cultural greatness. He must have studied the choral lyric poets of the past, Alcman and Stesichorus in particular, and the work of his elder contemporaries, Simonides of Ceos and Lasus of Hermione. He would have steeped himself in the poetry of Homer and Hesiod, and he must have received detailed and systematic training in the techniques of choral composition in the city where dithyramb (a choral lyric) was cultivated and where tragedy was beginning to evolve from the dramatic ritual dance performed at religious festivals associated with the god Dionysus. Tradition has it that one of his instructors in music was Agathocles, the teacher also of Damon, who was to become a leading figure in the intellectual and musical life of 5th-century Athens.

Training in techniques of choral composition

Professional career. Seventeen volumes of Pindar's poetry, comprising almost every genre of choral lyric, were known in antiquity. Only four books of *epinicia,* odes celebrating victory in athletic contests, have survived complete, doubtless because they were chosen by a teacher as a school book in the 2nd century AD. They are supplemented by numerous fragments, and in recent years finds of papyri have helped toward a deeper understanding of Pindar's achievement, especially in paeans and dithyrambs (both choral religious lyrics). All the evidence, however, suggests that the *epinicia* were Pindar's masterpieces. These are divided as Olympian, Pythian, Isthmian, or Nemean—the games in which the victories he celebrated were held; the *epinicia* number 44 odes in all. The earliest surviving *epinicion* (*Pythian Ode* 10) dates from 498, and Pindar already had an assured mastery of his medium when he wrote it. It would have been quite possible for him to evolve into a cosmopolitan artist like Simonides, welcome all over the Greek world and moving easily from city to city. No doubt Pindar visited the Panhellenic festivals, at Delphi (where the Pythian games were held) and Olympia in particular, to absorb the atmosphere of the games and the victories he celebrated. He would also have seen in person the homes of the aristocrats and the courts of the tyrants whose triumphs he sang. But in general he preferred to remain loyal to his native land and to reside in Thebes: it is Pindar's characteristic that his standards and values, like his poetry, changed little if at all over the years. Such patriotism meant sacrifices. Thebes, like Delphi, collaborated with the enemy in the Persian War—though admittedly Thebes had little alternative. But whereas Delphi's

Earliest surviving *epinicion*

prestige was quickly restored after the retreat of the Persians, Thebes's defection was not lightly forgiven or forgotten. Athens was to dominate the history of the 5th century, and for the first two-thirds of it Athens had very much the better of its long, drawn-out quarrel with Thebes: from 457 to 447 BC Boeotia was virtually an Athenian dependency. And almost everywhere, the aristocratic way of life, integral alike to Pindar's personality and to his art, was threatened. Politically and economically the monopolies of power by the noble families were broken. The aristocratic code, with its ultimately selfish individualism (summed up in a famous line of Homer, "ever to excel and to surpass other men") was undermined by the radical rationalism of a new age. Choral lyric itself had little future as a separate art form: tragedy absorbed into itself what was most vital in the tradition, and Pindar had no worthy successors. It is a tribute to the quality of Pindar's poetry that although he must have regarded these contemporary cultural and political developments with disdain, or at best with indifference (apart perhaps from his reinterpretation of some of the traditional stories concerning the gods), he was universally respected and accepted as a major creative artist.

Pindar's early poems have almost all been lost; it is probable, however, that what gave him a growing reputation beyond the borders of Boeotia were hymns in honour of the gods. Pindar was born at the time of the Pythian festival, and from his youth he had a close connection with the Pythian priesthood, which served the oracular shrine of Apollo at Delphi. Pindar and his descendants, indeed, enjoyed special privileges at Delphi, where his memory was cherished in later times and where an iron chair, in which it was said he had sat to sing, was exhibited. The first commissions for *epinicia* came mostly from aristocratic connections: the Aleuads in Thessaly (*Pythian Ode* 10; 498 BC); the Alcmaeonids in Athens (*Pythian Ode* 7; 486 BC); Agesidamus of Epizephyrian Locri (*Olympian Ode* 11; 484 BC); and above all the Aeacids of the island of Aegina (the series begins with *Paean* 6, dating from 490, and continues with *Nemean Ode* 7). Progress in winning recognition seems to have been steady, if slow. A significant breakthrough came when Pindar established a link with the court of Theron of Acragas through the tyrant's brother Xenocrates, whose chariot won the Pythian contest (*Pythian Odes* 6 and 12; both 490 BC). But the Persian invasion of Greece came before the promise of this new connection could be fulfilled. Pindar faced a crisis of divided loyalties, torn between a sense of solidarity with the aristocracy of Boeotia, who followed a pro-Persian policy, and a growing appreciation of Spartan and Athenian heroic resistance. Pindar was first and foremost a Theban and stood by his friends, many of whom paid for their policy with their lives. But it was Simonides, not Pindar, who wrote the poems of rejoicing at Greece's victories and of mourning for its glorious dead.

First commissions for *epinicia*

It took Pindar some years to re-establish himself: fortunately, his friends in Aegina were staunch (*Isthmian Ode* 8; 478 BC). It is virtually certain that he visited Sicily in 476–474 BC and was made welcome at the courts of Theron of Acragas and Hieron I of Syracuse. They were to elicit much of his greatest poetry, and it was through these connections that Pindar's reputation spread all over the Greek world so that commissions flowed in from the mainland, the islands, and also from the remoter outposts of Hellenism. Promising new contacts were made with the royal houses of Macedon and Cyrene (Alexander of Macedon, Fragment 120; Arcesilas of Cyrene, *Pythian Odes* 4 and 5; 462/461 BC). Theron and Hieron respected and admired Pindar, but his aristocratic temper made him dangerously outspoken. Diplomatic tact and finesse were not among his qualities, and his adroit rivals, Simonides and Bacchylides, were more pliant and adaptable (Bacchylides, not Pindar, celebrated Hieron's Olympic victory in the chariot race in 468 BC). Echoes of Pindar's bitter resentment sound in his poetry. So, too, Pindar's intervention on behalf of Damophilus, a noble exile from Cyrene (*Pythian Ode* 4, 279 ff.), seems to have been taken amiss, and he was not invited to commemorate Arcesilas' triumph at Olympia in 460 BC. Nevertheless, these were the years of supreme achievement, and Pindar found a growing demand for his poetry and a growing appreciation of his skill. His debt to Athens was amply paid in a famous tribute (Fragment 76) that the Athenians never tired of citing, one that earned the poet special honours in that city (and, according to ancient tradition, a fine at Thebes). It was probably in this period that Pindar married.

The subsequent decade of Athenian domination in central Greece coincided with a period when Delphi was controlled by Phocis in northern Greece. These were dark years for Pindar, and his poetic output dwindled. But he continued to celebrate Theban victories (*Isthmian Odes* 1 and 7); and he found inspiration in the achievements of his Aeacid friends of Aegina, though their days of nominal independence were clearly numbered (*Isthmian Odes* 5, 6; *Nemean Odes* 3–8; all celebrate Aeginetan successes). Pindar's last extant poem (*Pythian Ode* 8) appropriately commemorates an Aeacid victory.

The tradition that Pindar lived to the age of 80 may be true, but there is no surviving poem from later than 446 BC. It is known that he died at Argos.

Poetry. The *epinicion* form, which in Simonides' hands seems to have evolved into a relatively simple poem of rejoicing enhanced by touches of realism and humour, is assimilated by Pindar to the religious hymn. The praise and worship of the god whose festival is being celebrated set the tone, and thanksgiving is an integral part of the structure. A second constituent element is the myth, impressionistically treated in a series of short, sharply visualized scenes and meant to link the glorious present to the yet more glorious past, to give a new dimension to the transient moment of victory. A third ingredient is the aphoristic moralizing, often in Pindar of extreme beauty, even sublimity, in which the dangers of excessive pride in achievement are repeatedly stressed. The emotional impulse throughout stems from the aristocratic ideal of self-assertion, competition, and leadership —an ideal expressing itself most finely in battle but also finding fulfillment in athletic contests, in which the palm goes to superior physique and morale deriving from superior birth. His metrical range is exceptionally wide, with no two poems being identical in metre, and he controls difficult and involuted techniques with consummate professional mastery. His dialect is literary and eclectic, with few Boeotian elements; the vocabulary is enriched, poetic, and highly personal. Each poem is fused into a unity by the fire of Pindar's poetic inspiration, by a sweep and soar of imagination that give his poetry power and magnificence, by the shaping and controlling discipline of a fastidious art expressed in an intensely personal style. Delphic religious teaching found in Pindar a ready pupil, and he constantly spiritualized his material, turning away from the cruder traditional stories of the gods, avoiding the mundane details of the contest, striving to catch the fleeting radiance that plays about the moment of supreme endeavour when a man transcends his own limitations of physique and character and so proves worthy of his birth and ancestry. Delphi also profoundly influenced his style, which is frequently cryptic and oracular. He regarded himself as the Muse's prophet: she is the pythoness and he is the priest who puts her inspired message into intelligible shape.

Constituent elements of the *epinicion*

The only other major poet produced by Boeotia was Hesiod, who had flourished in the 8th century BC. The two are poles apart in background and temperament, but they share a deep religiosity, a groping toward something more profound and satisfying than contemporary cults could offer, a fondness for abrupt and violent transitions in thought and mood, and a forthright pungency of speech. A somewhat muted epitaph preserved in the "Greek Anthology" (7, 35) describes Pindar as the servant of the Muses, welcomed by strangers, beloved by his fellow citizens. Perhaps the poet would have regarded it as adequate.

Pindar's odes make great demands on the modern reader, and it is only in recent times that his art has begun to be appreciated for what it is. (The so-called Pindaric

ode has had a long and distinguished history in English literature, but it derives from an almost total misunderstanding and misapprehension of Pindar's own style and technique.) Even so, much essential evidence is missing. The musical settings that he composed to accompany his words are lost forever, though in view of the quality of the poetry it is probable that the words dominated the setting (as must have been the case in most Greek lyric). It is therefore impossible to re-create even in imagination the approximate sound of a Pindaric ode, or indeed to reconstruct visually the appearance and constitution of the choir: how many participated; what range of voices was employed; whether the singers were static, moved in procession, or danced—these are questions that cannot now be answered. Nor is it possible to picture at all clearly the festive occasion that was the background for the poetry. Yet efforts to understand the odes are rewarded by at least a glimpse of the poet behind them. The aristocratic society and standards, which meant everything to Pindar, were dead or dying. But in his art he re-created them, giving them new and permanent existence and value.

MAJOR WORKS

The *Epinikia* (498-446 BC; *The Odes of Pindar*), consisting of: *Olympionikai* (*The Olympian Odes*), *Pythionikai* (*The Pythian Odes*), *Nemeonikai* (*The Nemean Odes*), and *Isthmionikai* (*The Isthmian Odes*).

Although it is difficult to make a choice, special mention may perhaps be made of "Olympionikai i, Hieroni Syrakousio" (476 BC; "For Hieron of Syracuse"); "Olympionikai ii and iii, Theroni Akragantino" (476 BC; "For Theron of Acragas"); "Pythionikai i, Hieroni Aitnaio" (470 BC; "For Hieron of Aetna"); "Pythionikai iv, Arkesila Kyrenaio" (462 BC; "For Arcesilas of Cyrene"); "Pythionikai vi, Xenokratei Akragantino" (490 BC; "For Xenocrates of Acragas"); and "Pythionikai ix, Telesikratei Kyrenaio" (474 BC; "For Telesicrates of Cyrene"). A translation of all Pindar's odes is available in the "Loeb Series" (1915).

BIBLIOGRAPHY. A full account of editions, commentaries, translations, and criticism to 1966 may be found in DOUGLAS E. GERBER, *A Bibliography of Pindar, 1513–1966* (1969). The most up-to-date text is that of BRUNO SNELL, *Pindari Carmina cum fragmentis*, 2 vol. (1953 and 1964). O. SCHROEDER, *Pindari Carmina*, 5th ed. (1923), is the authoritative edition of a distinguished scholar who devoted his life to the study and elucidation of the poet. The best verse translations in English are by R.A. LATTIMORE, *The Odes of Pindar* (1947); and by C.M. BOWRA, *The Odes of Pindar* (1969). A magisterial treatment of Pindar's life and work is provided by ULRICH VON WILAMOWITZ-MOELLENDORFF, *Pindaros* (1922). R.W.B. BURTON, *Pindar's Pythian Odes* (1962), analyzes and explains these *epinicia* poem by poem. C.M. BOWRA, *Pindar* (1964), gives a sensitive appreciation of the poet's personality and art.

(D.E.W.W.)

Piperales

The Piperales, an order of the flowering plants (division Angiospermae, class Dicotyledoneae), comprises two families, the Piperaceae, or pepper family, and Saururaceae, or lizard's-tail family. In some systems of classification the family Chloranthaceae is also included within this order.

General features. Members of this order are mostly herbs or shrubs. The leaves are usually rather broad and alternately arranged on the stems. The densely aggregated flowers are quite small and lack petals and sepals. Reduced leaves, or bracts, are present below the flowers and in a few species are petallike in appearance. The flowers are most often bisexual, with one to ten stamens and usually a single superior ovary (*i.e.*, the ovary is above the other flower parts, the sepals, petals, and anthers, all of which attach at the base of the ovary). The ovary develops into a single-seeded or a few-seeded fruit.

The family Piperaceae is composed of ten genera, mostly shrubs, many vinelike, and largely confined to the tropics of both the Old World and the New World. The genus *Piper*, with centres of distribution in South America and Malaysia, contains around 1,000 species and is one of the larger ones in the plant kingdom. *Peperomia*, a genus of over 500 species, is also widely distributed in the tropics and subtropics of both hemispheres, a few species extending into Florida. A number of geophilous (ground-dwelling) species of *Peperomia* are found at high altitudes in the American tropics. Some taxonomists classify the genus *Peperomia*, along with a few other genera, as a third family, the Peperomiaceae, which differs from the Piperaceae proper by the absence of stipules (appendages at the base of leafstalks) and in characters of the stem anatomy and pollen. The family Piperaceae includes a number of species used by man, the most important of which is *Piper nigrum*, the source of the black and white pepper of commerce.

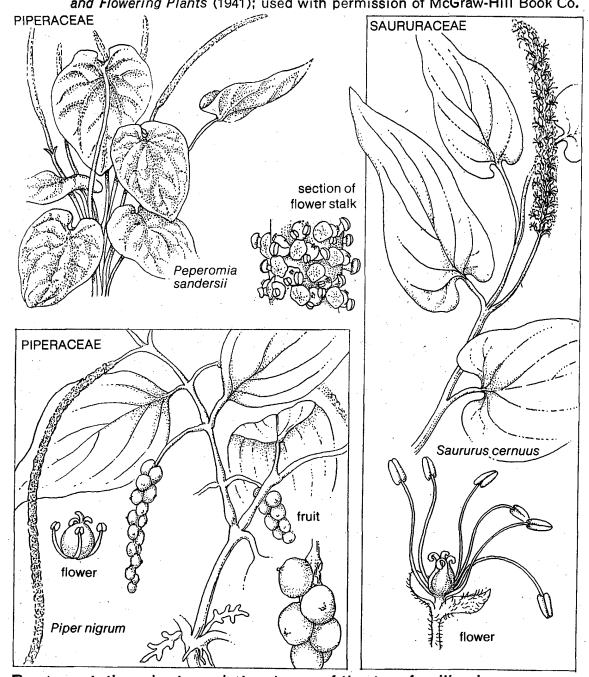
The family Saururaceae, which comprises three to five genera, depending on the authority followed, is made up of perennial herbs found in eastern Asia and North America. The genus *Saururus* (lizard's-tail) has one species in eastern North America and another in Asia. *Anemopsis* (yerba mansa) of the western United States and Mexico is the other member of the family found in the United States. Some species of Saururaceae often have a general resemblance to some of those of the family Polygonaceae (order Polygonales).

The two Piperales families are most readily distinguished from each other by the number of ovules, one in the Piperaceae and two to ten in the Saururaceae. The Piperaceae family also lacks the showy bracts, or upper leaves, found in Saururaceae. The ovary and the stem anatomy of the two families also differ fundamentally.

Little detailed information on the floral biology and natural history of the plants of this order is available. Many species have no obvious specialized means for dispersal of the seeds, but it has been suggested that seeds of many species are eaten by birds and thus disseminated. The small fruits of *Peperomia* are covered with a mucilaginous substance that allows them to adhere to objects. Thus, both birds and other small animals could serve as agents of seed dispersal.

Agents of seed dispersal

(*Peperomia*, *Saururus*) Reprinted with permission of The Macmillan Company from *Taxonomy of Vascular Plants* by G.H.M. Lawrence. Copyright by The Macmillan Company, 1951. (*Piper*) From R. Pool, *Flowers and Flowering Plants* (1941); used with permission of McGraw-Hill Book Co.



Representative plants and structures of the two families in the pepper order.

Economic importance. Pepper, prepared from the fruits, called peppercorns, of *Piper nigrum*, is one of the world's most important spices. It sometimes is referred to as black and white pepper to distinguish it from red, chili, and cayenne peppers, which belong to the genus *Capsicum* of the family Solanaceae (order Scrophulariales). Pepper, one of the first spices known to man, was mentioned in Indian writings more than 3,000 years ago and

was observed by Marco Polo to be in extensive use in China as a condiment in the 13th century; it later became one of the first articles of trade between the Far East and Europe. Although much appreciated by the early Greeks and Romans, pepper did not assume great importance in Europe until the Middle Ages. In addition to its use for seasoning food and in the preserving of meats, it also had considerable application in medicine at that time. Peppercorns were also once used as a medium of exchange to pay rent and taxes.

Pepper production. *Piper nigrum* is adapted to hot, wet tropical climates. It is said still to occur wild in Assam and northern Burma and apparently first became a cultivated plant in that part of the world. The plant, a woody climber, reaches 30 feet or more in height. The aerial portion produces roots that allow the plant to attach itself to the trees or poles on which it is usually cultivated. The shiny, ovate, prominently veined leaves are from five to ten inches (13 to 25 centimetres) long. The inconspicuous, self-pollinated flowers hang in spikes, four to five inches (ten to 13 centimetres) long. Each spike produces from 50 to 60 single-seeded, berrylike fruits, which change from green to red or yellowish red when ripe and dry to a blackish colour.

In commercial planting the pepper is propagated vegetatively from stem cuttings. Harvest begins in the third to fifth year, and the berries mature nine months after flowering. Plants continue bearing well for 12 to 15 years. Harvesting is done by hand when the berries are green but mature. For the manufacture of black pepper, the berries are allowed to dry in the sun after picking until they darken and are ready for shipment. The whole peppercorns are then ground to produce black pepper. For the preparation of white pepper, the fruits, often picked when fully ripe, are placed in water for several days to rot the fruit coat. Then the outer hull of the fruit is removed by trampling underfoot or by other means. After drying, the peppercorns eventually turn white and are ready to be ground into white pepper. Peppers contain the alkaloids piperine and piperidine, the latter in small amounts. A resin known as chavicine, however, is considered responsible for most of their pungency. An essential oil prepared from steam distillation of pepper flavours meat products and various sauces.

Preparation of black and white pepper

Until World War II, Indonesia produced 80 percent of the pepper for the world trade; since then India has become the chief exporter. Other countries exporting pepper include Malaysia, Brazil, Cambodia, and the Malagasy Republic. Annual world exports amount to between 50,000 and 75,000 tons. The United States is the world's chief importer.

Various other Asiatic species of *Piper*, generally called long pepper, also have uses similar to those of *Piper nigrum*. The two most important are *Piper retrofractum*, or Javanese long pepper, which occurs wild in Malaysia, and *Piper longum*, Indian long pepper, native to India. Both species are cultivated to a limited extent but do not enter world trade. In addition to their use as a spice, both are still employed in local medicine for a great variety of complaints. In early times among the Greeks and Romans, long pepper was more highly esteemed than black pepper.

Other products. Several other species of *Piper* are of economic importance. The leaves of the betel pepper, *Piper betle*, are chewed with the nuts of betel palm as a masticatory in much of the Old World tropics from Zanzibar to Oceania. Betel chewing is important at ceremonial occasions and acts as a mild stimulant as well as serving to sweeten the breath. This species also finds medicinal uses among native peoples of the areas where it occurs. Its cultivation is most advanced in India, where it is called *sireh*. Cubeb, *Piper cubeba*, native to Indonesia, is used as a medicine and formerly had some importance as a spice.

Kava, the Polynesian "national beverage"

Kava, or kawa, (*yagona*), made from *Piper methysticum*, has been referred to as the "national beverage" of Polynesians. It is widely cultivated in Oceania for its use as a ceremonial and social drink as well as for medicine. Peeled roots were once chewed, but for making the beverage scrapings from the root are fermented in water. Kava is a narcotic and is used primarily for its relaxing effects; excessive use can lead to intoxication.

The leaves of a few of the species of *Piper* occasionally are used as condiments. In both South America and Malaysia some species have been used as an ingredient of arrow poisons, apparently for their irritating effect, and as a poison to stupefy fish.

Other useful plants in the order include several species grown as ornamentals, especially *Peperomia*, of the Piperaceae family. The peperomias are mostly low herbs, occasionally epiphytic (*i.e.*, growing among the branches of trees), with fleshy leaves. Several species are cultivated as houseplants for their attractive foliage. *Saururus cernuus*, lizard's-tail, so named because of the shape of its inflorescence (flower cluster), is a plant of bogs and marshes of eastern North America and is occasionally cultivated in pool or bog gardens.

Classification.

Annotated classification.

ORDER PIPERALES

Mostly herbs and shrubs; usually with broad, alternate, entire, petiolate (stalked) leaves with stipules attached to the petiole or lacking; flowers bisexual or, rarely, unisexual, perianth lacking, subtended by bracts, borne in dense spikes or, rarely, racemes, hypogynous to epigynous, stamens 1–10; carpels 1–4 or, rarely, 5, free or united, usually with a single ovule; fruit a drupe, follicle, or capsule; seeds with copious endosperm, a mealy perisperm, and minute embryo with 1 or 2 cotyledons. Members mostly tropical.

Family Piperaceae

Erect, trailing, or climbing herbs, shrubs, or (rarely) small trees; vascular bundles of the stem often scattered, as in the class Monocotyledoneae; leaves usually alternate, rarely opposite or whorled, sometimes fleshy, stipules present or absent; flowers minute, without perianth, usually bisexual (plants rarely dioecious), borne in dense, sometimes rather fleshy spikes; stamens 1–10, anthers 2-locular, distinct or confluent, dehiscent by longitudinal slits; ovary superior, stigmas 1–5, carpels 2–5, united, 1-loculed with 1 erect ovule; fruit (frequently called a berry) a dry or fleshy drupe. About 10 genera with perhaps 3,000 species (although conservative estimates would place it at half that number) in the tropics of both hemispheres.

Family Saururaceae

Perennial herbs, usually of moist habitats; vascular bundles of stem united into a single ring; leaves ovate, alternate, stipules present, upper leaves, or bracts, often coloured or showy; flowers bisexual, in dense spikes or racemes; stamens 6–8 or sometimes fewer, anthers 2-locular, dehiscent by vertical or longitudinal slits; ovary superior to half-inferior, styles and stigmas the same number as the carpels, carpels 3–4, free, each with 1-locule, or compound; ovules 2–10 in each carpel on parietal placentas; fruit a fleshy follicle or capsule, apically dehiscent. Three to 5 genera with a total of 4 to 7 species in southeastern Asia, the United States, and Mexico.

Critical appraisal. The family Chloranthaceae is placed in the Piperales order in many systems of classification, whereas others assign it to the order Laurales. Chloranthaceae is readily separated from the Piperaceae and Saururaceae families by its opposite leaves, petioles that are somewhat connate (fused around the stem) at the base, small, free stipules, and more or less inferior ovary.

Modern authorities are generally in agreement that the Piperales is a rather primitive order, representing a slight advance over the Magnoliales order, from which it is thought to have been derived.

BIBLIOGRAPHY. C.K. ATAL and J.N. OJHA, "Studies on the Genus *Piper*, pt. 4, Long Peppers of Indian Commerce," *Econ. Bot.*, 19:157–164 (1965), history, uses, chemistry, and descriptions of various species of long pepper; I.H. BURKILL, *A Dictionary of the Economic Products of the Malay Peninsula* (1935), names, uses, and chemistry of species of *Piper;* ARTHUR CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968), description of order, key to families, discussion of phylogeny, and references; RONALD GATTY, "Kava —Polynesian Beverage Shrub," *Econ. Bot.*, 10:241-249 (1956), general discussion of kava; HOWARD SCOTT GENTRY, "Introducing Black Pepper into America," *Econ. Bot.*, 9:256–268 (1955); J. HUTCHINSON, *The Families of Flowering Plants*, 2nd ed., 2 vol. (1959), description of order and

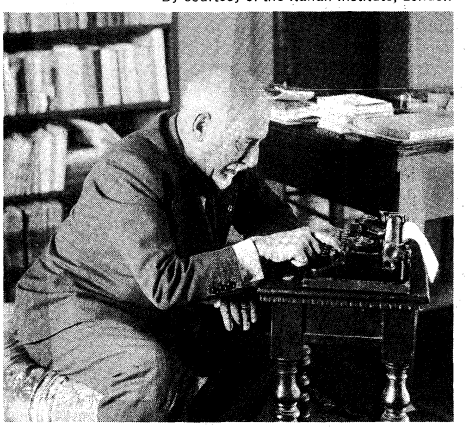
families with a key; GEORGE H.M. LAWRENCE, *Taxonomy of Vascular Plants* (1951), description of order and families, with references; J.W. PURSEGLOVE, *Tropical Crops*, vol. 2, *Dicotyledons* (1968), discussion of economic species of *Piper* with detailed treatment of *Piper nigrum;* FREDERIC ROSENGARTEN, JR., *The Book of Spices* (1969), discussion of black and white pepper with recipes; TRUMAN G. YUNCKER, "The Piperaceae—a Family Profile," *Brittonia*, 10:1–7 (1958), discussion of the taxonomic history and phylogeny.

(C.B.He.)

Pirandello, Luigi

Author of short stories and novels, as well as the plays, for which he is best known, and the winner of the 1934 Nobel Prize for Literature, the Italian writer Luigi Pirandello was one of the revolutionary innovators of 20th-century literature. By his invention of the "theatre within the theatre" (with *Sei personaggi in cerca d'autore*; *Six Characters in Search of an Author*), he provides a metaphor for the contradictions of existence in the equivocal relationship between actors and the characters they represent.

By courtesy of the Italian Institute, London



Pirandello.

Early years and marriage

Pirandello was born near Girgenti (modern Agrigento), Sicily, on June 28, 1867, the son of a sulfur merchant who wanted him to enter commerce. Pirandello, however, was not interested in business; he wanted to study. He first went to Palermo, the island's capital, and, in 1887, to the University of Rome. After a quarrel with the professor of classics, he went in the following year to the University of Bonn, Germany, where in 1891 he gained his doctorate in philology for a thesis on the dialect of Girgenti.

In 1894 his father arranged his marriage to Antonietta Portulano, the daughter of a business associate, a wealthy sulfur merchant. This marriage gave him financial independence, allowing him to live in Rome and to write. He had already published an early volume of verse, *Mal giocondo* (1889), which paid tribute to the poetic fashions set by Giosuè Carducci; and this was followed by other volumes of verse, including *Pasqua di Gea* (dedicated to Jenny Schulz-Lander, the love he had left behind in Bonn) and the translation of Goethe's *Römische Elegien* (*Elegie romane*). But his first significant works were short stories, at first contributed to periodicals without payment.

In 1903 a landslide shut down the sulfur mine in which his wife's and his father's capital was invested. Suddenly poor, Pirandello was forced to earn his living not only by writing but also by teaching Italian at a teacher's college in Rome. As a further result of the financial disaster, his wife developed a persecution mania, which manifested itself in a frenzied jealousy of her husband. His torment ended only with her removal to a sanatorium in 1919 (she died in 1959). It was this bitter experience that finally determined the theme of his most characteristic work, already perceptible in his early short stories—the exploration of the tightly closed world of the forever changeable human personality.

Pirandello's early narrative style stems from the *verismo* ("realism") of two distinguished Italian novelists of the late 19th century—Luigi Capuana and Giovanni Verga. The titles of Pirandello's early collections of short stories —*Amori senza amore* (1894; "Loves without Love"), *Beffe della morte e della vita* (1902–03; "The Jests of Life and Death")—suggest the wry nature of his realism that is seen also in his first novels: *L'esclusa* (1901) and *Il turno* (1902). Success came with his third novel, often acclaimed as his best, *Il fu Mattia Pascal* (1904). Although the theme is not typically "Pirandellian," since the obstacles confronting its hero result from external circumstances, it already shows the acute psychological observation that was later to be directed toward the exploration of his characters' subconscious.

First novels and stories

Pirandello's understanding of psychology was sharpened by reading such works as *Les altérations de la personnalité* (1892), by the French experimental psychologist Alfred Binet; and traces of its influence can be seen in the long essay *L'umorismo* (1908), in which he examines the principles of his art. Common to both books is the theory of the subconscious personality, which postulates that what a person knows, or thinks he knows, is the least part of what he is. Thus, before the work of Sigmund Freud, the founder of psychoanalysis, was known to Pirandello, and before the French novelist Marcel Proust had published any of his works, the themes of psychology had become material for literature. They found their most complete expression in the volumes of short stories, *La trappola* (1915) and *E domani, lunedì . . .* (1917), and in such individual stories as *Una voce, Pena di vivere così*, and *Con altri occhi.*

Later novels

Meanwhile, he had been writing other novels, notably *I vecchi e i giovani* (1913) and *Uno, nessuno e centomila* (1925–26). Both are more typical than *Il fu Mattia Pascal.* The first, a historical novel, reflecting the Sicily of the end of the 19th century and the general bitterness at the loss of the ideals of the Risorgimento (the movement that led to the unification of Italy), suffers from Pirandello's tendency to "discompose" rather than to "compose" (to use his own terms, in *L'umorismo*), so that individual episodes stand out at the expense of the work as a whole. *Uno, nessuno e centomila,* however, is at once the most original and the most typical of his novels. It is a surrealistic description of the consequences of the hero's discovery that his wife (and others) see him with quite different eyes than he does himself. Its exploration of the reality of personality is of a type better known from his plays.

He had first turned to the theatre in 1898 with *L'epilogo,* but the accidents that prevented its production until 1910 (when it was retitled *La morsa*) kept him from other than sporadic attempts at drama until the success of *Così è (se vi pare)* in 1917. This delay may have been fortunate for the development of his dramatic powers. *L'epilogo* does not greatly differ from other drama of its period, but *Così è (se vi pare)* began the series of plays that were to make him world famous in the 1920s. A demonstration, in dramatic terms, of the relativity of truth, and a rejection of the idea of any objective reality not at the mercy of individual vision, it anticipates Pirandello's two great plays, *Sei personaggi in cerca d'autore* (1921) and *Enrico IV* (1922). *Sei personaggi . . .* is the most arresting presentation of the typical Pirandellian contrast between art, which is unchanging, and life, which is an inconstant flux. Characters that have been rejected by their author materialize on stage, throbbing with a more intense vitality than the real actors, who, inevitably, distort their drama as they attempt its presentation. And in *Enrico IV* the theme is madness, which lies just under the skin of ordinary life, and is, perhaps, superior to ordinary life in its construction of a satisfying reality. The play finds dramatic strength in its hero's choice of retirement into unreality in preference to life in the uncertain world.

Fame and influence

The production of the first of these two plays in Paris in 1923 made Pirandello widely known, and his work became one of the central influences on the French theatre; French drama from the existentialistic pessimism of Jean Anouilh and Jean-Paul Sartre to the absurdist comedy of

Eugène Ionesco and Samuel Beckett is tinged with "Pirandellianism." His influence can also be detected in the drama of other countries, even in the religious verse dramas of T.S. Eliot.

In 1920 Pirandello said of his own art:

> I think that life is a very sad piece of buffoonery; because we have in ourselves, without being able to know why, wherefore or whence, the need to deceive ourselves constantly by creating a reality (one for each and never the same for all), which from time to time is discovered to be vain and illusory . . . My art is full of bitter compassion for all those who deceive themselves; but this compassion cannot fail to be followed by the ferocious derision of destiny which condemns man to deception.

This despairing outlook attained its most vigorous expression in Pirandello's plays, which were criticized at first for being too "cerebral" but later recognized for their underlying sensitivity and compassion. The plays reflect the *verismo* of Capuana and Verga in dealing mostly with people in modest circumstances, such as clerks, teachers, and lodginghouse keepers, but from whose vicissitudes he draws conclusions of general human significance.

World tour with his own company

The universal acclaim that followed *Sei personaggi . . .* and *Enrico IV* sent Pirandello touring the world (1925–27) with his own company, the Teatro d'Arte in Rome. It also emboldened him to disfigure some of his later plays (*e.g.*, *Ciascuno a suo modo* [1924]) by calling attention to himself, just as in some of the later short stories it is the surrealistic and fantastic elements that are accentuated.

After the dissolution, because of financial losses, of the Teatro d'Arte in 1928, Pirandello spent his remaining years in frequent and extensive travel. He died in Rome on December 10, 1936. In his will he requested that there should be no public ceremony—only "a hearse of the poor, the horse and the coachman."

MAJOR WORKS

VERSE: *Mal giocondo* (1889); *Pasqua di Gea* (1891); *Pier Gudrò* (1894); *Elegie renane* (1895); *Elegie romane* (1896), translations from Goethe's *Römische Elegien*; *Zampogna* (1901); *Scamandro* (1909), verse play; *Fuori di chiave* (1912).

CRITICISM: *Arte e scienza* (1908); *L'umorismo* (1908).

SHORT STORIES: (a chronological list of the first book editions of collections of *novelle*): *Amori senza amore* (1894); *Beffe della morte e della vita* (1902–03); *Quand'ero matto . . .* (1902); *Bianche e nere* (1904); *Erma Bifronte* (1906); *La vita nuda* (1910); *Terzetti* (1912); *Le due maschere* (1914); *La trappola* (1915); *Erba del nostro orto* (1915); *E domani, lunedì . . .* (1917); *Un cavallo nella luna* (1918); *Berecche e la guerra* (1919); *Il carnevale dei morti* (1919). English translations have appeared in the following selections: *Horse in the Moon* (1932), contains *Un cavallo nella luna* and 11 other stories; *Better Think Twice About It!* [*Pensaci, Giacomino!*] *and Twelve Other Stories* (1933); *The Naked Truth* [*La vita nuda*], *and Eleven Other Stories* (1934, new ed. 1947); *A Character in Distress* (1938, U.S. title, *His Medals*, 1939), "La tragedia d'un personaggio" and 18 other stories; *Short Stories* (1959), translations of 22 stories, including "Pena di vivere così," "Such Is Life"; *Short Stories*, selected, translated, and introduced by Frederick May (1965), includes "Capannetta," "The Little Hut," Pirandello's earliest known prose work, and "La giara," "The Jar," probably his best known story.

NOVELS: *L'esclusa* (written 1893, published 1901; *The Outcast*, 1925); *Il turno* (written 1895, published 1902); *Il fu Mattia Pascal* (1904; *The Late Mattia Pascal*, 1923); *Suo marito*, later title *Giustino Roncella nato Boggiòlo* (1911); *I vecchi e i giovani* (1913; trans. by C.K. Scott-Moncrieff, *The Old and Young*, 1928); *Si gira . . .*, later title, *Quaderni di Serafino Gubbio operatore* (1916; trans. by Scott-Moncrieff, *Shoot! Si gira . . . the Notebooks of Serafino Gubbio, Cinematograph Operator*, 1927); *Uno, nessuno e centomila* (1925–26; *One, None and a Hundred Thousand*, 1933).

PLAYS: Pirandello wrote more than 50 plays. The following is a selection from the basic 31-volume Bemporad-Mondadori edition of the plays collected under the title *Maschere nude*, "Naked Masks" (1920–35): *Tutto per bene* (1920: *All for the Best*, 1960); *Sei personaggi in cerca d'autore* (1921; trans. by F. May, *Six Characters in Search of an Author*, 1954); *Enrico IV* (1922; *Henry IV*, trans. by F. May, 1960); *Vestire gli ignudi* (1923; *To Clothe the Naked*, 1962); *La vita che ti diedi* (1924; *The Life I Gave You*, 1959); *Ciascuno a suo modo* (1924; *Each in His Own Way*, 1924); *Così è (si vi pare)* (1925, first pub. 1918; *Right You Are* [*If You Think So*], trans. by F. May, 1960; by Eric Bentley as *Right You Are*, 1954); *Sagra del Signore della Nave*; *L'altro figlio*; and *La giara* (1925: *Our Lord of the Ship*; *The House with the Column*; and *The Jar*, in *The One-Act Plays of Luigi Pirandello*, 1928); *Il piacere dell'onestà* (1925, first pub. 1918; *The Pleasure of Honesty*, in *To Clothe the Naked, and Two Other Plays*, 1962); *Il giuoco delle parti* (1925, first pub. 1919; *The Rules of the Game*, in *To Clothe the Naked, and Two Other Plays*, 1962); *All'uscita* (first pub. 1916); *Il dovere del medico* (first pub. 1912); *La morsa* (first pub. 1898 with the title *L'epilogo*); *and L'uomo dal fiore in bocca* (1926: *At the Gate; The Doctor's Duty*; *The Vise*; and *The Man with the Flower in His Mouth*, in *The One-Act Plays of Luigi Pirandello*, 1928); *Diana e la Tuda* (1927; *Diana and Tuda*, 1950); *La nuova colonia* (1928; *The New Colony*, in *The Mountain Giants and Other Plays*, 1958); *Lazzaro* (1929; *Lazarus*, 1952); *Questa sera si recita a soggetto* (1930; *Tonight We Improvise*, 1932); *Come tu mi vuoi* (1930; *As You Desire Me*, 1931); *Quando si è qualcuno* (1933; *When Someone Is Somebody*, in *The Mountain Giants and Other Plays*, 1958).

BIBLIOGRAPHY

Editions: The most complete edition of Pirandello's works is the so-called *edizione definitiva* in the series *I classici contemporanei italiani* (1956–60). It consists of the following volumes: *Novelle per un anno*, 2 vol. (1956–57); *Tutti i Romanzi* (1957); *Maschere nude*, 2 vol. (1958); *Saggi, poesie, scritti varii* (1960).

Pirandello had planned to write 365 short stories, one for each day of the year, and himself began to edit the stories in a series entitled *Novelle per un anno;* the first volume appeared in 1922 and the 15th and last in 1937.

Bibliography: MANLIO LO VECCHIO MUSTI, *Bibliografia di Pirandello*, 2nd ed. (1952).

Biography: F.V. NARDELLI, *L'uomo segreto: Vita e croci di Luigi Pirandello*, 3rd rev. ed. (1962), first written in 1932, with the collaboration of Pirandello himself, revealed the tortured background of Pirandello's life with Antonietta Portulano. It is fundamental for an understanding of the author, although from the circumstances of its writing, it appears now somewhat hero-worshipping in tone. GASPARE GIUDICE, *Luigi Pirandello* (1963), presents the life and work of Pirandello in the setting of his time; it is probably the best general account of the author (in Italian).

Criticism: W. STARKIE, *Luigi Pirandello, 1867–1936*, 4th ed. (1967), a very personal, critical account, with extracts and synopses, of Pirandello's work in the context of European literature; THOMAS BISHOP, *Pirandello and the French Theater* (1960), a useful discussion of Pirandello's ideas, and an account of his lasting influence on the French theatre; J.H. WHITFIELD, "Pirandello and T.S. Eliot: An Essay in Counterpoint," in *English Miscellany* (1958), a study, not of influences, but of affinities and contrasts, not always expected; and "Pirandello and the Question-Mark of the Twentieth Century," in *A Short History of Italian Literature*, 2nd ed. (1969), a thumbnail account of Pirandello's work and importance; GLAUCO CAMBON (ed.), *Pirandello: A Collection of Critical Essays* (1967), a helpful set of discussions of various aspects of Pirandello criticism; OSCAR BUDEL, *Pirandello*, 2nd ed. (1969), a little book that will be more useful to those who already have some idea of the European background of Pirandello.

(J.H.W.)

Pisa

A university town in central Italy, Pisa is the capital of the province of Pisa in the region of Tuscany. It lies in the alluvial plain of the Arno River, about six miles from the Ligurian Sea. Its population was about 100,000 in 1970.

History. Pre-Roman Pisa (Pisae), the origins of which have been a matter of controversy since ancient times, was possibly inhabited by the Ligurians, before coming under Roman control as a naval base and a frontier post against the Ligurians to the north.

Situated on the northern bank of the Arno, Pisae became a Roman colony soon after 180 BC and flourished as a station on the coastal road leading northward from Rome. Fortifications, theatres, baths, and other public buildings were constructed, and by AD 313 Pisa had a Christian bishop.

As the Roman Empire disintegrated in the 4th and 5th centuries, the land route northward lost importance, and



Pisa, the baptistery, cathedral, and campanile (the leaning tower).
Fotocielo

Pisa was isolated from the chief areas of conflict and maintained its urban character, securing from its Lombard conquerors a degree of autonomy and freedom to trade that made it the only Lombard port of any importance. Pisa's navy defended the adjacent coasts, while some shipping continued to arrive from Corsica and Sardinia. Despite Saracen (Muslim) incursions in the 9th century, which led to a sharp decline in trade, Pisa was the greatest city in Tuscany in the 10th century. Agriculture and local trade, both by road and water, probably brought more wealth than did overseas commerce. During the 11th century Pisa not only exploited the products and markets of its large Tuscan hinterland but it also took the initiative against Muslim raiders; in 1016 the Pisans and Genoese drove the Saracens from Sardinia, and in 1063 the Pisan fleet sacked Muslim Palermo. Participation in the Crusades secured valuable commercial positions for Pisan traders in Syria.

Pisa grew in wealth and strength to rival Venice, as the establishment of overseas settlements and protectorates —especially in the western Mediterranean—brought trading opportunities and raw materials. These developments were accompanied by territorial expansion in Corsica and Sardinia as well as in Tuscany, by a growth of population and of industrial production in wood and leather, by the emergence of a commune, and by a republican form of government that provided a measure of participation and power for the middle and merchant groups in society. Pisa usually enjoyed the support of the German emperors in its long conflicts with Genoa at sea and with its Tuscan rivals, Lucca and Florence, on land. These struggles culminated in Pisa's defeat by the Genoese fleet at the decisive Battle of Meloria in 1284.

Battle of Meloria

Despite the loss, after Meloria, of its power at sea and of much of its trade overseas, Pisa experienced an astonishing industrial revolution late in the 13th century. The accelerated transfer of woollen manufacturing from the countryside to the town permitted industry to increase in efficiency and to employ cheap labour, thus reducing its costs; this, in turn, encouraged a shift of population to the city and the profitable enclosure of cultivated land for sheep grazing. As agricultural activity dwindled, the population of the town and its suburbs rose from perhaps 15,000 in the mid-12th century to some 50,000 in about 1315. The Levant trade was largely abandoned, but Pisan ships still brought grain to Tuscany, exported local products, and carried timber, wines, and other goods all over the western Mediterranean. Large quantities of merchandise continued to pass through the city until the 15th century, when serious silting, accompanied by changes in land levels and in the course of the river, made the movement of laden galleys up the Arno almost impossible. Porto Pisano and subsequently Leghorn, both on the coast, were then used for the actual loading and unloading of large ships.

Pisan prosperity was reflected in the characteristic *casa-torre*, a tall inhabited tower usually built of brick and stone, and in the city's churches, particularly the grandiose and spectacular group of cathedral, baptistery, and campanile—the leaning tower, which formed a harmonious ensemble constructed of white marble in a distinctively Pisan blend of Romanesque and Gothic. The cathedral and the baptistery were decorated by a succession of distinguished sculptors, including Guglielmo Pisano, who made the cathedral's first pulpit (1162); Bonanno Pisano, sculptor of its bronze doors (1180); Nicola Pisano, who carved the baptistery pulpit (1260); and Nicola's son Giovanni Pisano, sculptor of the existing cathedral pulpit (1310).

Economic change was accompanied by social tensions. During the 13th century the *popolo*, a group that included vigorous petty merchants of rural origins, provided a new political element; yet popular constitutions and anti-magnate legislation never really deprived the nobility of power and influence. Social divisions, in fact, remained vague and shifting, as aristocrats and businessmen rose, fell, and intermingled. Industry flourished, but foreign trade and agriculture were in decay, and the town, overcrowded, unhygienic, and surrounded by the stagnant swamps of the Arno's floodplain, suffered from endemic malaria. The catastrophic visitations of plague, the Black Death, which first came in 1348, brought further disaster. Military and economic failures, accentuated by internal faction, led to political apathy; to increasingly tyrannical rule by captains and local despots; and finally, in 1406, to Pisa's subjugation by Florence, which had long outdistanced its rival in population and industrial power. Pisa's independence was over.

Florentine rule

Florentine rule, though oppressive and galling, brought a revival of trade and other material advantages. Pisans were deported, heavily taxed, and excluded from office and from the full profits of commerce but they benefitted from the organization of a Florentine fleet and the consequent restoration of Pisa as a trading and shipbuilding centre. There was a new citadel and an unwelcome garrison but the Florentines, who had strong public and private interests in Pisa, improved the countryside, encouraged immigration into the town, and often resided there themselves.

When French armies invaded Italy in 1494, Pisa temporarily reasserted its independence; the city sustained a series of wars and sieges until Florence reconquered it in 1509. Thereafter it continued to decay as a provincial Tuscan town; by 1550 the population had fallen below 9,000. The Florentine rulers of Pisa were not always unenlightened; they encouraged Pisa's university, to which the great scientist Galileo Galilei—born at Pisa in 1564—was connected; they founded an office for the reclamation of marshlands; and in 1562 they instituted the Order of the Knights of Saint Stephen to combat the Barbary pirates.

The modern city. Pisa survived as a tranquil and elegant city, the seat of an archbishopric and of a great university. It is an important railway junction, and its light industries have brought prosperity. Widespread destruction in World War II diminished Pisa's charm, but tourists continued to flock there to see the world-famous leaning tower, now in some danger of collapse.

BIBLIOGRAPHY. The only full treatment in English, now outdated, is J. ROSS and N. ERICHSEN, *The Story of Pisa* (1909). In addition to the detailed scientific studies and bibliographical references in the bulletin, *Bollettino Storico Pisano* (1932–), see A. NEPPI-MODONA, *Forma Italiae Regio VII: Etruria*, vol. 1, *Pisae* (1953); and E. TOLAINI, *Forma Pisarum: problemi e ricerche per una storia urbanistica della città di Pisa* (1967), for ancient and medieval history and topography; W. HEYWOOD, *A History of Pisa: Eleventh and Twelfth Cen-*

turies (1921); D. HERLIHY, *Pisa in the Early Renaissance: A Study of Urban Growth* (1958); and E. CRISTIANI, *Nobiltà e popolo nel comune di Pisa dalle origini del podestariato alla signoria dei Donoratico* (1962), on the 13th and 14th centuries; M. MALLETT, "Pisa and Florence in the 15th Century: Aspects of the Period of the First Florentine Domination," in *Florentine Studies*, ed. by N. RUBENSTEIN (1968).

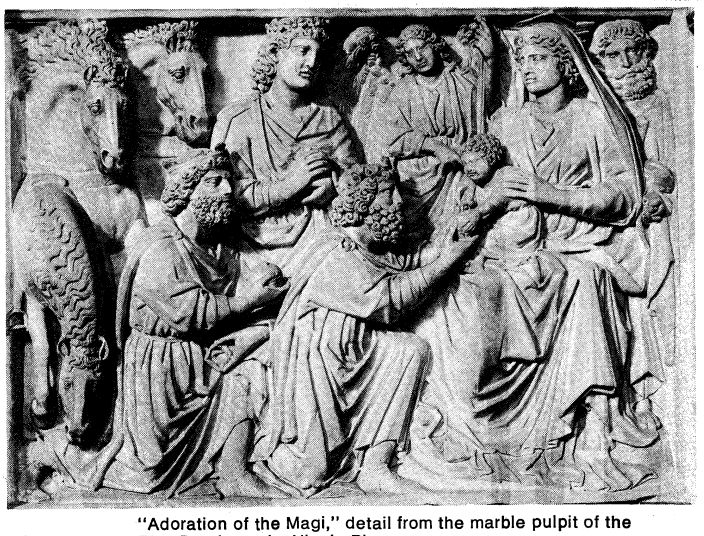
(A.T.L.)

Pisano, Nicola and Giovanni

Nicola Pisano, a sculptor who worked in Tuscany between 1260 and 1278, was reponsible for introducing a classical style inspired by the art of antiquity into the mainstream of Italian medieval art. His son, Giovanni (active 1265–1314), began his career in his father's workshop and so thoroughly assimilated the ideas he found there that his early work is difficult to distinguish from that of his father. After Nicola's death Giovanni carried on the traditions of this studio and lent to them his own distinctive artistic personality. These two sculptors and their workshops created a new sculptural style for the late 13th and 14th century in Italy; the echoes of this tradition remain noticeable throughout the Italian Renaissance. Nicola's origins are unclear. He is first recorded in 1260 in Pisa (or perhaps 1259, if corrections are made for the medieval Pisan calandar), where he had just completed a hexagonal freestanding pulpit in the baptistery of the cathedral. Although his signature on the pulpit clearly reads "Nicola Pisanus" (thus Nicola Pisano), or "Nicola, the Pisan," documents of 1266 relating to a later work twice call him "Master Nicola from Apulia," Apulia being the province located in the southeastern section of the Italian peninsula. While most scholars now accept an Apulian birth for Nicola, there is still much dispute about his artistic training. Since the sculpture of the Pisa pulpit indicates that Nicola was a mature artist by 1260, his birth date has been postulated to have been some time in the decade 1220–30, thus making him an artist in his 30s at the time of the carving of the Pisa pulpit. Later sources indicate that his son, Giovanni Pisano, was born in Pisa somewhere around 1250, a date that does not contradict the supposed date for Nicola's own birth and that also proves he had already settled in Pisa by mid-century.

Nicola's Apulian origins

Alinari



"Adoration of the Magi," detail from the marble pulpit of the Pisa Baptistery by Nicola Pisano.

Nicola's sculptural training, in line with medieval practices, was probably obtained through an apprenticeship in an already established workshop. If Nicola began his training in Apulia he would most likely have been taught by one of the army of craftsmen whom the Hohenstaufen emperor Frederick II had employed to decorate his vast new building projects. Frederick's artists consciously imitated an ancient Roman style of sculpture in order to give visual support to his role as the Holy Roman emperor. If, on the other hand, Nicola had actually left Apulia before receiving his sculptural training, he might have apprenticed himself either to local Tuscan or Lombard workshops; strong echoes of both regional styles appear in his work.

Although no work can definitely be attributed to Nicola before the Pisa pulpit, the strong classical spirit which motivates its forms suggests more than simple first-hand experience and fascination with then visible ancient Roman sculpture. Quite specific formal motifs in the figures of the Pisa pulpit compare closely with sculptural fragments representing Jupiter and imperial advisers from Frederick's triumphal gateway over the Via Appia at Capua (now in the Museo Provinciale Campano in Capua) and argue persuasively for Nicola's training in the artistic workshops of the Emperor. Pisa's close alliance with the empire of Frederick II, even after Frederick's death in 1250, would have provided good reasons both for Nicola's emigration to that city and for the rather sudden emergence of the overtly classicizing style of the pulpit which he carved.

Artistic importance of the Pisa pulpit

The Pisa pulpit marks one of the extraordinary moments in the history of Western art when a new style, distinct from all its predecessors, though indebted to them, clearly asserted itself and opened new avenues for artistic expression. Whatever Nicola's actual formal artistic training may have been, the Pisa pulpit indicates the widest possible range of artistic experiences. Nicola's familiarity with antique style could have resulted from study of extant Roman reliefs at such places as Benevento (on the via Appia from Apulia to Rome) and at Rome itself; actual borrowings from Roman reliefs preserved in the Campo Santo in Pisa have long been recognized in Nicola's Pisan reliefs. Roman minor arts, such as gems and cameos, which formed part of Frederick II's avidly amassed collections, could also have played a role in forming Nicola's style, as could early Christian fresco and mosaic decorations in the city of Rome. Nicola's work also shows references to localized Tuscan and Lombard forms, as well as isolated motifs deriving from French Gothic sculpture and architecture about which he may have learned either by visiting French-influenced centres in Apulia or by an actual trip to France. Nicola assimilated this encyclopaedic array of artistic expressions and transformed them into a brilliantly unified whole that gave new grandeur and new energy to his narratives and a new sense of direction to art in Tuscany from 1260 onward.

Nicola's style changed dramatically during the carving of the Pisa pulpit from the amplitude of form and rhythmic fluency of movement evident in the relief panel of "The Presentation of Christ in the Temple," to a greatly more agitated treatment of the space and forms in which figures become smaller in relationship to the entire surface area of the relief and pile up on one another in a surface rather than a spatial organization. The figures themselves become more animated and are twisted to emphasize their dramatic potential more fully. Rather than being a rejection of Nicola's earliest known style, however, this later style in fact grew directly out of his concern for presenting the human emotional content of his subject matter. Some of this stylistic change, especially the last two reliefs from the Pisa pulpit, may also be attributable to apprentices or members of his workshop, including Giovanni Pisano, Nicola's son.

Commissions in Siena and Perugia

Nicola's workshop, including Giovanni Pisano, Fra Guglielmo, and Arnolfo di Cambio, also shared in his next commissions, the pulpit for the Cathedral of Siena, 1265–68, and the Fontana Maggiore in the main square of Perugia of 1278. In the Siena pulpit, the form of which is much like that of the Pisa pulpit, Nicola continued the investigations of expressive human figures which permeate the Pisan reliefs. Yet although Nicola's authorship of the design is evident throughout the pulpit, the individual dispositions of the various assisting sculptors toward ancient Roman or Gothic forms is equally obvious; the separate strains of the two traditions which Nicola had so successfully united in the Pisa pulpit again reassert their independence at Siena. This uneasy relationship between the Antique and the Gothic becomes

increasingly clear at Perugia where the 25-sided fountain mixes tales from Romulus and Remus with fables from Aesop, classical personifications of nearby places with representations of the liberal arts and the labours of the months, and contemporary historical figures with Old Testament characters and heraldic animals. The shifts in the style of the sculpture at Perugia were also influenced by the fact that single figures were employed rather than complex narratives as in Nicola's earlier work.

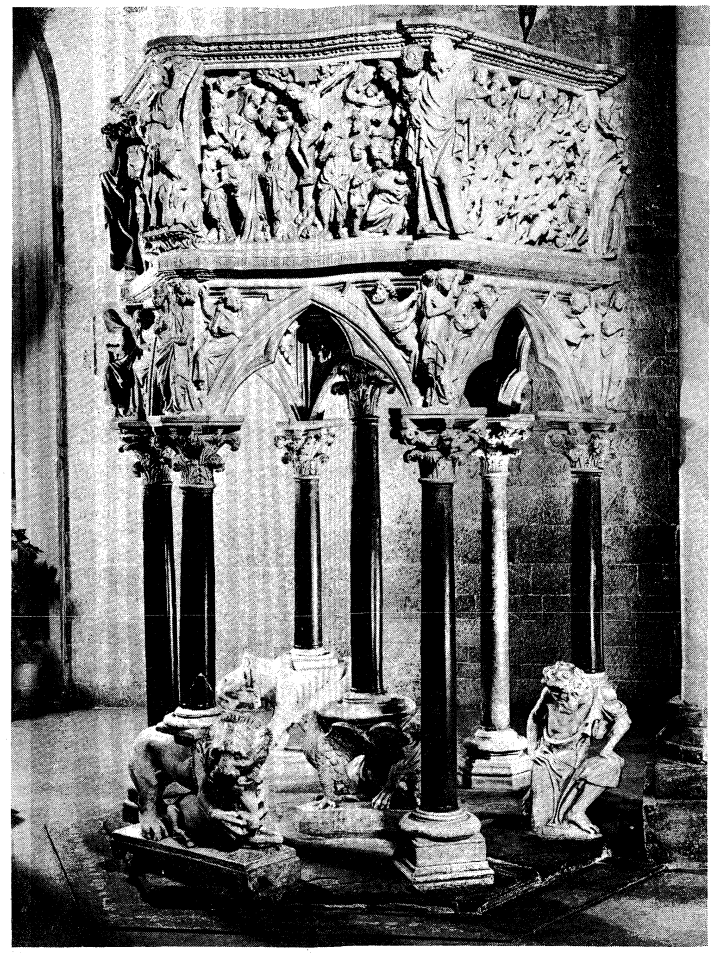
It was in the contract (1265) for the Siena pulpit that Giovanni Pisano is first specifically mentioned as an assistant to his father. Since he was at that time not referred to as "magister," or independent master craftsman, Giovanni must still have been in his teens. Although there is much scholarly dispute over assigning particular sculptural work from the Siena pulpit or the Perugia fountain to Giovanni, there is factual evidence that by September 1285 Giovanni had rejected his Pisan citizenship and had become a resident of Siena. Around this time he began his work on the design and sculptural ornamentation of the facade for the Cathedral of Siena which became, in its lavishness and ordering, the model for virtually all future Gothic facade decoration in central Italy. Unlike French examples, in which figural ornament pulsates over the entire facade, Giovanni's designs for the Siena facade offer a much more architectural approach to the problem. The lower story is simply decorated with colonnettes (small columns) and a restrained foliate pattern, which follows the vertical movement of the colonnettes. Aside from the carved lintels over the doors, figural sculpture begins at the level of the arches over the entrances with slightly larger than life-size figures of prophets and sibyls and continues throughout the rest of the facade. Although each figure inhabits a discrete niche, agitated, forward-bending poses cause them to converse across the vast space of the facade and soften the otherwise clearly stated architectonic lines that order the structure. Recent research has emphasized particularly close relationships in design between the sculpture of the Siena facade and French foliate patterns and figural reliefs, especially from the cathedral at Auxerre. Since there are no known documentary references to Giovanni Pisano between 1268 and 1278, the possibility of a trip through France during these years seems extremely likely.

Giovanni's works in Siena and Pistoia

Next to the Siena Cathedral facade, Giovanni's pulpit in Pistoia, completed in 1301, is his greatest achievement. The five narrative reliefs of this pulpit roughly parallel the subject matter of Nicola's Pisa pulpit 40 years earlier, as does the overall architectural format; but the style pushes the expressive qualities innate in Nicola's Pisa pulpit to a new level of intensity. In the "Annunciation," the "Nativity," and the "Annunciation to the Shepherds," the extreme agitation that characterizes all the reliefs for the Pistoia pulpit pulsates throughout the panel. Figures, animals, drapery, and landscape features are wrenched into physically impossible configurations; light shatters over the broken surfaces and deeply cut relief; and each figure responds convulsively to the individual situations in which he acts as a participant.

In Pistoia, as well as in the Siena pulpit and the Fontana Maggiore at Perugia, there is a strong interest in Gothic forms such as twisted and bent figures, angular contortions of bodies, and a heightened narrative emotional intensity; yet a simple labelling of such forms as Gothic ignores the roots of the relief style and denies the consistent use that both Nicola and Giovanni made of antique sculpture. The later reliefs of the Pisani are explicitly Roman in inspiration and derive from battle or hunt type sarcophagi of the 3rd century. The careful layering, or distinct tiering, of French and German reliefs is not used by the Pisani until after 1302. What is critical to the change in style from the first Pisa pulpit reliefs to the Pistoia reliefs is a preference for an overall agitated and deeply cut surface as opposed to the earlier more massive and monumental organization of forms.

Giovanni never repeated the frenzy of forms which covers the Pistoia pulpit. Instead he returned to the more stately, classical spirit which had been at the heart of



Pulpit, marble, of the church of S. Andrea, Pistoia, Italy, by Giovanni Pisano, 1297–1301.
SCALA, New York

Nicola's earliest work. The reasons for this cannot be documented, but they most likely stem in part from Giovanni's experience with Giotto's monumental and heroic style which was already in the ascendancy by the time that the Pistoia pulpit was completed. Giovanni, in fact, carved a marble Madonna and Child for the Arena Chapel in Padua at approximately the same time that Giotto painted his profoundly moving fresco cycle there (*c.* 1305). In addition, the quasi-imperial political movements established by Pope Boniface VIII at the turn of the 14th century may also have prompted Giovanni to return to more overtly Classical quotations.

Giovanni's return to the classical spirit

From 1302 to 1310 Giovanni again worked in Pisa, this time for a pulpit for the cathedral. In this pulpit, now badly reconstructed after having been disassembled, the relief style is considerably more docile than that of the Pistoia reliefs. Although a decidedly Gothic tiering of the narrative appears for the first time at Pisa, the decorative motifs, column capitals, and freestanding caryatids (a sculptured figure used as a column) show a new severity of style and a renewed interest in direct copying of Roman sculpture. These developments parallel aspects not only of Giotto's painting but of contemporary French sculptors as well.

Giovanni's last recorded work was a tomb sculpture for Margaret of Luxembourg in Genoa in 1311 (fragments now in the Palazzo Bianco). He was last recorded in Siena in 1314, and it is presumed that he died shortly thereafter. If, as is clear from his work in Siena, Giovanni was Italy's only Gothic sculptor, it is also true that he never lost sight of the heritage of Rome that underlies all of the artistic thinking of central Italy. While at times the Antique style may seem to submerge beneath his fascination for contemporary Gothic forms, nonetheless Giovanni, like his father, continuously insisted on a reintegration of antique forms and the Antique style into his sculp-

ture. Thus the Pisani created in their work a focus on the antique past that led ultimately to the revivals of the 15th century now commonly known as the Renaissance.

MAJOR WORKS

GIOVANNI PISANO: Fontana Maggiore (with Nicola Pisano) (1278; Perugia, Italy); external decoration of the Baptistery, Pisa (1278–85; Pisa, Italy); sculptural decoration of the facade, Siena Cathedral (1284; Siena, Italy); pulpit (1297–1301; S. Andrea, Pistoia, Italy); pulpit (1302–11; Cathedral, Pisa); tomb of Margaret of Luxembourg (originally in S. Francesco di Casteletto, Genoa, but now destroyed, 1311; fragments in Palazzo Bianco, Genoa; Palazzo Rosso, Genoa).

NICOLA PISANO: pulpit (*c.* 1255–60; Baptistery, Pisa); pulpit (1265–68; Siena Cathedral).

BIBLIOGRAPHY. The most comprehensive writings on the Pisani in English are by J. POPE-HENNESSY, *Italian Gothic Sculpture* (1955); and J. WHITE, *Art and Architecture in Italy, 1250 to 1400* (1966), both of which give monographic discussions of the individual works of art. Long articles by M. WEINBERGER on Nicola in the *Encyclopedia of World Art*, vol. 10, col. 641–647 (1965), and by J. POPE-HENNESSY on Giovanni, *ibid.*, vol. 6, col. 357–365 (1962), are models not only for their review of scholarly opinions and attributions, but also for the comprehensive bibliographies that they include. M. WUNDRAM also catalogs recent publications in his article "Toskanische Plastik von 1250–1400," *Zeitschrift für Kunstgeschichte*, 21:243–270 (1958).

E. BERTAUX in *L'Art dans l'Italie méridionale* (1904), is still important for his early discussion of Nicola's Apulian connections. G.H. and E.R. CRICHTON carried this discussion into English in *Nicolo Pisano and the Revival of Sculpture in Italy* (1938), but their arguments lack focus and oftentimes substance as well. C. GNUDI, *Nicola, Arnolfo, Lapo* (1948), is important for his discussion of Nicola's studio (in Italian). C. SEYMOUR, JR., "Invention and Revival in Nicola Pisano's 'Heroic Style'," *Studies in Western Art: Acts of the Twentieth International Congress of the History of Art*, vol. 1, pp. 207–226 (1963), reviews recent literature for Nicola and gives cogent political reasons for Nicola's classicizing style; he also carries the heroic mode of Nicola's sculpture into the 15th century. W.R. VALENTINER's article on Nicola's Apulian connections, "Studies on Nicola Pisano," *Art Quarterly*, 15:9–36 (1952), is still useful.

(J.T.Pa.)

Pissarro, Camille

Camille Pissarro was the only Impressionist painter who participated in all eight of the group's exhibitions; his kindness, warmth, wisdom, and encouraging words cast him in a fatherly role to struggling younger artists—Claude Monet, Auguste Renoir, Paul Cézanne, and Paul Gauguin—who were exploring new means of personal expression. Despite financial hardships that continued until he reached his 60s, Pissarro never lost faith in the new art, believing that "one must be sure of success to the very end, for without that there is no hope!"

Pissarro was born July 10, 1830, in St. Thomas, Danish West Indies, son of a prosperous Jewish merchant, Abraham Gabriel Pissarro, and Rachel Manzano-Pomié. At the age of 12 he left home for studies in Paris, where he showed an early interest in art. Returning to the West Indies after five years to work in his father's store, he began making sketches of the exotic island and its people. Because he was unable to obtain his father's permission to study art, he ran away to Caracas in 1853 and remained there for two years with the Danish painter Fritz Melbye. Finally, Pissarro's father relented, and in 1855 he returned to France. His earliest canvases, dating from this period, are figure paintings and landscapes of the tropics and of the French countryside; although broadly painted, they show the careful observation of nature that was to remain a characteristic of his art throughout his life.

Earliest canvases

The uninspired academic teaching at the École des Beaux-Arts, where he was enrolled, led Pissarro to seek out the painter Camille Corot, who permitted Pissarro to call himself Corot's "pupil" at a Salon exhibition in 1864. At this time Pissarro was also attracted to the rural, sentimental paintings of the Barbizon artist Jean-François Millet and to the works of Gustave Courbet, the leading proponent of everyday Realism. During the 1860s Pissarro participated in the famous Parisian Café Guerbois discussions, in which artists and writers exchanged ideas, and worked with the younger painters Renoir and Monet. To escape the Franco-Prussian War, in 1870 Pissarro fled to England; there he and Monet, who had also fled France, visited the museums, where they viewed British landscape paintings. It was in London that Pissarro married Julie Vellay, formerly his mother's maid, who had already borne him two of their seven children.

When Pissarro returned to France in 1871, he found his house in Louveciennes looted and a great number of his paintings destroyed. Soon he was to look for another house in Pontoise. (Because it was costly to live in Paris, Pissarro, like several of his painter friends, lived in villages not far from the city.) His surroundings formed the theme of his art for some 30 years and were always carefully chosen: "I require a spot that has beauty!" At Pontoise he was joined by Cézanne in 1872, and the two of them painted out of doors, even in the middle of winter.

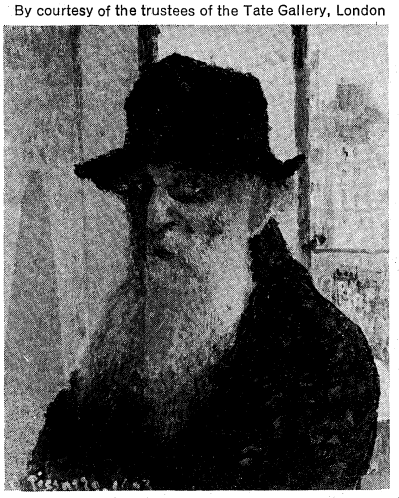
Pissarro's paintings are never dramatic; on the contrary, his leading motifs during the 1870s and 1880s are simply houses, factories, trees, haystacks, fields, labouring peasants, and river scenes. Forms do not dissolve but remain firm, and colours are strong; during the latter part of the 1870s his comma-like brushstrokes frequently recorded the sparkling scintillation of light, as in "Orchard with Flowering Fruit Trees, Springtime, Pontoise" (1877). Although his paintings were sold by the dealer Paul Durand-Ruel, who represented several of the Impressionists, Pissarro continued to experience financial hardships, which he described in letters to his eldest son, Lucien; this remarkable correspondence began in 1883 and lasted for 20 years.

Motifs during the 1870s and 1880s

In some of the letters to his son, Pissarro expressed dissatisfaction with his own work. Preoccupied by problems of style and technique, he eagerly embraced the Neo-Impressionist theories of Georges Seurat, whom he met in 1885 through the painter Paul Signac. Seurat's technique, consisting of meticulously painted small dots of juxtaposed colours, was adopted by Pissarro; for about five years he painted in this "divisionist" manner, a style which made his works unpopular with dealers, collectors, and even his old fellow artists. Overwhelmingly discouraged by their continuing state of poverty, Madame Pissarro considered drowning herself and their two youngest children. Finally, Pissarro abandoned the style, not, however, because of the opposition he met but because "it was impossible to be true to my sensations and consequently to render life and movement, impossible to be faithful to the effects, so random and so admirable, of nature. . . . " At about this time, also, there was an estrangement from Paul Gauguin, who had formerly worked at his side but now was involved with the new Symbolist movement.

A large and successful retrospective exhibition of Pissar-

By courtesy of the trustees of the Tate Gallery, London



Pissarro, "Self-Portrait," oil painting, 1903. In the Tate Gallery, London.

Late works

ro's paintings was held by Durand-Ruel in 1892, giving the artist greater financial stability, although by this time he was troubled by a chronic eye infection that frequently made it impossible for him to work out of doors. Both in 1893 and 1897 he took hotel rooms in Paris from which he painted 24 canvases of the city's streets by day and by night, in sun, rain, and fog. During the 1890s he also did a series of river scenes in Rouen, likewise depicting the various effects of nature. From 1900 until his death in Paris on November 13, 1903, Pissarro continued working, mainly in Paris, Éragny, Dieppe, and Le Havre, with freshness of vision and increasing freedom in his technique. More than 1,600 works, consisting of oils, gouaches (a method of painting with opaque watercolours prepared with gum), temperas, pastels—and even paintings on fans and on porcelain—as well as nearly 200 prints, give testimony to the high quality of Pissarro's half century of work.

MAJOR WORKS

"The Hermitage at Pontoise" (*c.* 1867; Justin K. Thannhauser Collection, New York); "Still Life" (1867; Toledo Museum of Art, Ohio); "Jallais Hill, Pontoise (La Côte du Jallais, Pontoise)" (1867; Metropolitan Museum of Art, New York); "Louveciennes, the Road to Versailles" (1870; Emil G. Bührle Collection, Zürich); "Lower Norwood, London, Snow" (1870; National Gallery, London); "Crystal Palace, London" (1871; Henry J. Fisher Collection, Greenwich, Connecticut); "Quarry Near Pontoise" (1874; Kunstmuseum, Basel, Switzerland); "The Little Bridge" (1875; Städtische Kunsthalle Mannheim, Mannheim, West Germany); "Orchard with Flowering Fruit Trees, Springtime, Pontoise" (1877; Louvre, Paris); "The Red Roofs, View of a Village in Winter" (1877; Louvre); "Peasant Girl with a Stick" (1881; Louvre); "The Cafe au Lait" (1881; Art Institute of Chicago); "Towpath at Pontoise" (1882; Durand-Ruel Collection, Paris); "Île Lacroix, Rouen—Effect of Fog" (1888; Philadelphia Museum of Art); "Paris, the Boulevard Montmartre at Night" (1897; National Gallery, London); "Boulevard des Italiens, Morning, Sunlight" (1897; National Gallery of Art, Washington, D.C.); "Place du Théâtre Français" (1898; Minneapolis Institute of Arts, Minnesota); "Rue de l'Épicerie, Rouen—Morning, Gray Weather" (1898; Durand-Ruel Collection, Paris); "The Church of Saint-Jacques, Dieppe—Morning, Rainy Day" (1901; Nate B. and Frances Spingold Collection, New York); "Louvre from Pont Neuf" (1902; Sterling and Francine Clark Institute of Art, Williamstown, Massachusetts); "Bridge at Bruges" (1903; City Art Gallery, Manchester, England).

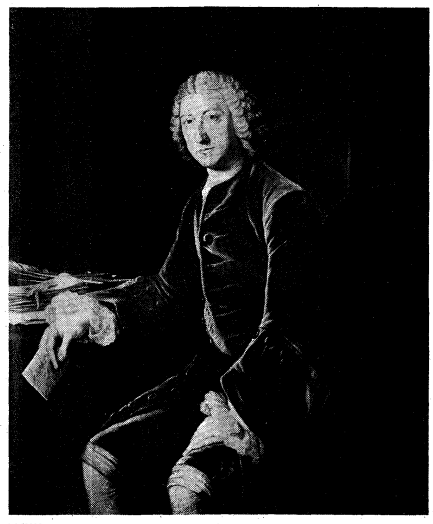
BIBLIOGRAPHY. A most useful annotated bibliography of the artist appears in JOHN REWALD, *The History of Impressionism*, rev. ed., pp. 637–39 (1961). A good biographical study that includes colourplates and a selected bibliography is JOHN REWALD, *Camille Pissarro* (1963). Indispensable for the study of the artist's work is the critical study and catalog by LUDOVIC RODOLPHE PISSARRO and LIONELLO VENTURI, *Camille Pissarro: son art—son oeuvre*, 2 vol. (1939). Its extensive bibliography lists over 500 publications. Pissarro's graphic works were cataloged by LOYS DELTEIL, *Le Peintre-Graveur illustré*, vol. 17, *Camille Pissarro, Alfred Sisley, Auguste Renoir* (1923). For the artist's own writings see CAMILLE PISSARRO, *Letters to His Son Lucien*, ed. with the assistance of LUCIEN PISSARRO by JOHN REWALD (1943).

(Mi.V.)

Pitt, William, the Elder

William Pitt the Elder (later the 1st earl of Chatham), who dominated the British government during the Seven Years' War (1756–63), is often regarded as the greatest British statesman of the 18th century. He held effective directing power of government for no more than five years, but in that period (1756–61) his energy, vision, and practical ability in conducting the war secured the transformation of his country into an imperial power. Though the rest of his life was politically active, it seems by comparison to be either a preparation for that burst of national and personal glory or an aftermath of it, for his formation of another ministry in 1766 was clouded by an illness that compelled his withdrawal two years later.

Background and education. Pitt was born in London on November 15, 1708. His mother, Lady Harriet Villiers, daughter of Viscount Grandison, belonged to the Anglo-Irish nobility; his father, Robert Pitt, member of



William Pitt, the Elder, painting from the studio of W. Hoare, 1754. In the National Portrait Gallery, London.

By courtesy of the National Portrait Gallery, London

Parliament, was the son of Thomas ("Diamond") Pitt, governor of the East India Company's "factory" at Madras, India, where he made a vast fortune and secured one of the world's largest diamonds (sold in 1717 to the regent of France). "Diamond" Pitt had returned from India with a despotic temper made devilish with spleen and gout; he quarrelled violently with his wife and declared war on "that hellish confusion that is my family," but he treated his grandson William with affection. Father Robert was mean and cantankerous, and the Villiers blood was notoriously unstable. William inherited the gout, as well as a haughty temper and a strain of manic depression not understood at the time but politely called "gout in the head."

Such was the background and the smoky, explosive inheritance that was suddenly to blaze into genius. But William's passionate temper and Pitt truculence had to be disciplined, so he was sent to Eton College, where he acquired social polish and learned to be aloof and yet agreeable, to be politely insolent. Delicate health and the early onset of gout deprived him of field sports and hunting, but he learned to ride with a good seat and take his port wine, and he enjoyed the select company of clever and well-connected friends—the two Grenvilles (one to be Earl Temple; the other, George, to be first minister to George III), George Lyttelton, Charles Pratt (to become a follower of Pitt and, as the 1st Earl Camden, a member of his 1766 ministry), and other men who would later become influential in politics, as well as Henry Fielding, author of *Tom Jones*. But Pitt hated the brutal harshness of Eton and determined to have his own sons educated at home. He continued his education at Trinity College, Oxford, but left after a year without taking a degree. He then spent several months at the University of Utrecht in the Netherlands, probably studying law.

Education

His classical education made him think, act, and speak in the grand Roman manner. His favourite poet was Virgil, and he never forgot the patriotic lessons of Roman history; he constantly read Cicero, the golden-tongued orator who could yet lash offenders with his indignation. Later, in Parliament, his organ-like voice could be distinctly heard outside the House. This voice, perfect timing, and splendid gestures were worthy of David Garrick, the greatest actor of the day and a personal friend; Pitt's lean, tall, commanding figure, combined with a Roman beaky nose and hawklike eyes—large and gray but turning black when he was roused—overwhelmed all onlookers. To his countrymen he was to become a divine portent, a voice from the Delphic oracle, with the mind of Julius Caesar. Frederick II the Great of Prussia was to call him a "true Roman."

For the present, possessed of a mere £100 a year, he nevertheless rejected the church, a younger son's last re-

sort as a career. While he was vegetating on a small family property in Cornwall, which he called a "cursed hiding-place" in one of his many letters to his adored sister and confidante, clever Nan (Ann) Pitt, help came from a politically powerful millionaire nobleman, Lord Cobham, who lived in splendour in a palatial mansion and vast park at Stowe, in Buckinghamshire, to which William and his friends paid visits. Cobham sent William abroad on "The Grand Tour" of Europe (only France and Switzerland were visited, however) and later bought him a cornetcy—a commission—in his own regiment of horse (1731).

Early political career. In 1735 Pitt was offered one of the "pocket" boroughs his brother controlled, Old Sarum in Wiltshire, and entered Parliament. He belonged to the small group known as "Cobham's cubs" and the "boy patriots," the connection of family friends and place hunters whom Cobham was mobilizing to oppose the ministry of Sir Robert Walpole (later the 1st earl of Orford). Walpole had governed England since 1720, monopolizing patronage, and had—they thought—become too ready to compromise in foreign affairs for the sake of peace. The "patriots" joined other discontented Whigs such as John Carteret (later Earl Granville) and William Pulteney (later the 1st earl of Bath) to rally opposition forces behind Frederick Louis, prince of Wales, who was vehemently estranged from his father, King George II.

Attack on Walpole

There were no formal political parties in the 18th century, and political power, together with the financial opportunities it brought, was a gift of patronage from a handful of landowning family oligarchies and from the monarch himself; nor was there a formal opposition in Parliament, and opposition to the king's ministry was regarded as factious and even traitorous. Pitt's maiden speech in Parliament was so critical of the ministry that it provoked Walpole to deprive him of his military commission, to "muzzle this terrible young cornet of horse."

In 1737 the Prince of Wales made Pitt one of his court officials with a salary of £400 a year. He was still a relatively poor dependent of a powerful Whig clan but already showed an independence of mind and a readiness to appeal to public opinion outside Parliament that were new in English political life: when Walpole dismissed him from his cornetcy, he ostentatiously drove about London in a one-horse chaise to underline his poverty. His talents as an orator had already become clear. He repeatedly referred to the "voice of England," which had to be sought outside Parliament because Parliament was so packed with placemen and sinecurists. He claimed to speak for the commercial interests and even for the colonies overseas, the latter scarcely represented in the Commons. He was using arguments that carried far beyond the close interests of the Whig families; but he made lasting and valuable friendship with the rich sugar planter-aldermen of the City of London in his opposition to Walpole's cautious handling of the disputes with Spain over West Indian trade.

Walpole at last fell from power in 1742 and was replaced by a ministry that included his old colleagues Thomas Pelham-Holles, the 1st duke of Newcastle, and Philip Yorke, the 1st earl of Hardwicke, with Carteret as secretary of state. Pulteney was silenced by the grant of a peerage. The "boy patriots," of whom Pitt was the acknowledged leader, were still excluded from power. They opposed Carteret even more vigorously than they had Walpole. In the War of the Austrian Succession (1740–48), Pitt, a former warmonger, now vigorously opposed the sending of men and subsidies to the Continent to check the French by protecting Hanover (the king's territory in Germany) and condemned Carteret as a "Hanover troop minister": for this he was never forgiven by his sovereign. Pitt insisted that French power should be opposed at sea and in its colonial possessions, not on the Continent. When Carteret was forced to resign in 1744, Newcastle and his brother Henry Pelham took office and wanted to include Pitt in their ministry, but George II refused to accept him, though he did accept Cobham, Lyttelton, and Grenville. It was at this time that Pitt first appeared in Parliament swathed in bandages, on crutches, and with a huge gout boot on his foot, parading his illness. But, in the Jacobite rising of 1745 (the Forty-five Rebellion), Pitt gained new stature as the one effective statesman.

Opposition of George II to Pitt

In February 1746, the King agreed to appoint Pitt joint vice treasurer of Ireland at £3,000 a year, and two months later he became paymaster general of the forces; he bowed very low as he kissed the King's hands, but George II wept with rage. The post of paymaster was one of the most sought after in government, with ample opportunity for corruption. There was an outcry from Pitt's friends, who suspected he had been bought, but he proceeded to show both his contempt for moneymaking and his integrity as an honest but comparatively poor man by ostentatiously refusing to take for himself any more than the official salary of more than £4,000 a year: he put the interest that paymasters before had appropriated to themselves, together with the accounts of the paymaster's funds, into the Bank of England and won the people's hearts again. He introduced many reforms into the administration, and, though he supported the Pelhams' alliance with Hanover (which was a change of tack), he tried also to strengthen British naval power.

A legacy of £10,000 from the old Duchess of Marlborough at this time, left to Pitt "for the noble defence he has made for the support of the laws of England, and to prevent the ruin of his country," enabled him to indulge in more lavish expenditure and generosity. He spent a good deal on landscape gardening and bought a new property near London. At this time, after a furious quarrel, he became estranged from his sister Nan, who had been his hostess for many years.

When Henry Pelham died in office in 1754, Pitt hoped for advancement, but, after much reshuffling and intrigue, Newcastle and Henry Fox (later 1st Baron Holland) abandoned him for the sake of expediency. He then became ill and retired to a new house at Bath, groaning "I wish for nothing but a decent and innocent retreat, not to afford the world the ridiculous spectacle of being passed by every boat that navigates the river." An invalid and an aging bachelor, he suddenly fell in love with Lady Hester Grenville, became engaged at once, and was married by November 1754. She was 33 and he 46, but she adored him, possibly from the times in her childhood when he had visited Stowe with her brothers. She was attractive, clever, patient, and eminently practical—particularly about money, arranging mortgages, satisfying creditors, and pouring away her own fortune, in his last years of grandiose extravagance, to protect him from disillusionment. It proved to be an ideally happy marriage with a well-ordered, loving home and family ("the infantry" Pitt called them); later he was to be found making hay with them at Hayes Place, his house in Kent, going for picnics, and chasing butterflies. He magically became healthy and happy, ready for his last big parliamentary fight for high office. But first, because of his attacks on Newcastle's ministry, he was dismissed, penniless, from the pay office in 1755. His brother-in-law Earl Temple helped with an annuity of £1,000.

Marriage to Lady Hester Grenville

Leadership during Seven Years' War. The outbreak of the Seven Years' War gave Pitt his supreme opportunity for statesmanship. The war began with heavy losses and considerable confusion of policy. The popular demand for Pitt became irresistible, and he declared, "I am sure I can save this country, and nobody else can." In November 1756 he formed a ministry that excluded Newcastle, with the Duke of Devonshire as its nominal head. In June 1757 Newcastle returned to office on the understanding that he should control all the patronage and leave Pitt to conduct the war.

Pitt determined that it should be in every sense a national war and a war at sea. He revived the militia, re-equipped and reorganized the navy, and sought to unite all parties and public opinion behind a coherent and intelligible war policy. He seized upon America and India as the main objects of British strategy: he sent his main expeditions to America, to ensure the conquest of Canada, and supported the East India Company and its "heaven-born general," Robert Clive, in their struggle against

the French East India Company. He subsidized and reinforced the armies of Frederick the Great of Prussia to engage the French on the Continent, while the British Navy harassed the French on their own coasts, in the West Indies, and in Africa. Choosing good generals and admirals, he inspired them with a new spirit of dash and enterprise. His hand, eye, and voice were everywhere. By 1759, the "year of victories," Horace Walpole, man of letters and son of Sir Robert Walpole, wrote with reluctant admiration, "Our bells are worn out threadbare with ringing for Victories." Pitt, the "Great Commoner," was known and feared throughout the world. This resolute and concerted policy was too much for Bourbon France, and, by the terms of the Treaty of Paris in 1763, Great Britain remained supreme in North America and India, held Minorca as a Mediterranean base, and won territory in Africa and the West Indies.

The Treaty of Paris

Pitt had given Britain a new empire besides preserving and consolidating the old. But, before the war ended, he had been forced to resign. In 1760 George III came to the throne resolved, as was his chief adviser, the Earl of Bute, to end the war. When Pitt failed to persuade his colleagues to declare war on Spain to forestall its entry into hostilities, he resigned in October 1761. He alone was not tired of war. He never considered its carnage or the ruin facing a bankrupt country. He had tended to concentrate the whole conduct of government into his own hands and worked with furious energy. His haughty and aloof manner, which alienated so many of his colleagues, and his high-handed treatment of affairs had earned him respect and admiration but little friendship. When his resignation was accompanied by a peerage for Hester and an annuity to her of £3,000, there was again an outburst of abuse and scurrility. Just as when he had accepted the pay office, this acceptance of a peerage and a pension for his wife seemed to be the result of a political bargain. As rewards for his immense services they were meagre enough, but it was some measure of his unique reputation for high-minded disinterestedness that his accepting them should provoke so much bitter disillusionment. His effigy was burned in the city, and Hester was reviled as Lady Cheat'em. Pitt attacked the terms of the Treaty of Paris as an inadequate recognition of Great Britain's world-wide success. But, though his popular appeal was soon restored, his career as war minister was over.

Last years. Pitt fell back on his gout and his gardening. In 1765, an admirer left him a splendid new estate at Burton Pynsent in Somerset, where he could plant avenues of noble trees. He was frequently at Bath, where they stood up in the Pump Room when he drank the water. He now had the attacks of "gout in the head" that led to bouts of insanity.

When Bute resigned in 1763, he was succeeded by George Grenville, and Pitt's attacks on the administration completed a breach between the two brothers-in-law. Pitt was becoming a champion of liberty, condemning the high-handed action taken by the ministry against a member of Parliament, John Wilkes, whose paper, the *North Briton*, had attacked the King's speech at the opening of Parliament and who eventually had to flee abroad. Inactive in 1764 and 1765, Pitt re-entered the stage in January 1766 to deliver a passionate plea for imperial liberty on behalf of the American colonists who had resisted the Stamp Act and to demand that act's repeal. In July the King asked him to form a ministry drawn from all sections of the houses of Parliament. Pitt's judgment and wisdom were impaired at this time, and, never having paid attention to manoeuvring among political connections, he found it difficult to form a coherent ministry. It was a fiasco rightly called a "tessellated pavement" by his political opponent Edmund Burke. Pitt himself chose the secondary post of lord privy seal, for which he was created earl of Chatham, but this meant abandoning the House of Commons and the possibility of influencing it directly by his oratory there. The "Great Commoner" retired to the Lords and fell ill for another two years, leaving a rudderless government, under the luckless Duke of Grafton and Charles Townshend, to abandon his policies. Engulfed in a black fit of insanity, Chatham withdrew completely and in 1768 resigned office. He acquired a group of followers in the House of Commons and, in an alliance with Lord Rockingham's group of opposition Whigs, offered a threat to Lord North's ministry, but this opposition was, in the end, without results.

Defense of the American colonists

Chatham's last years were clouded by illness, yet he was to reappear in the House of Lords—with ever greater difficulty—as an elder statesman. He continued to plead for generous treatment of the American colonists though he did not wish to grant them independence, partly for fear of their falling into the hands of France; in 1775 he hurriedly introduced a bill designed to suspend repressive measures at Boston and to maintain the legislative authority of Parliament over the Colonies while using the Continental Congress established at Philadelphia as a body for assessing the monetary contributions of each colony. Although the bill was summarily rejected, it indicates how Chatham would have handled the American problem. His last speech, against any diminution of an empire based on freedom, closed a political career that had become devoted to a reconciliation of imperial power with constitutional liberty. Chatham died at Hayes on May 11, 1778, falling back into the arms of his son William who was reading to him the passage in Homer's *Iliad* on Hector's farewell. He was buried in Westminster Abbey with all the funeral pomp he could have desired and with public grief.

BIBLIOGRAPHY. LORD MACAULAY'S review (1834) of Rev. FRANCIS THACKERAY'S *History of the Right Honourable William Pitt, Earl of Chatham* in his *Essays and Lays of Ancient Rome* (1884, reprinted 1954), brilliant but dated; J.H. PLUMB, *Chatham* (1965), a short, interesting, readable, and up-to-date account; LORD ROSEBERY, *Chatham: His Early Life and Connections* (1910), long, but worth digging into; O.A. SHERRARD, *Lord Chatham: A War Minister in the Making*, 3 vol. (1952–58); and W.C.B. TURNSTALL, *William Pitt, Earl of Chatham* (1938), both scholarly and detailed.

(V.M.W.)

Pitt, William, the Younger

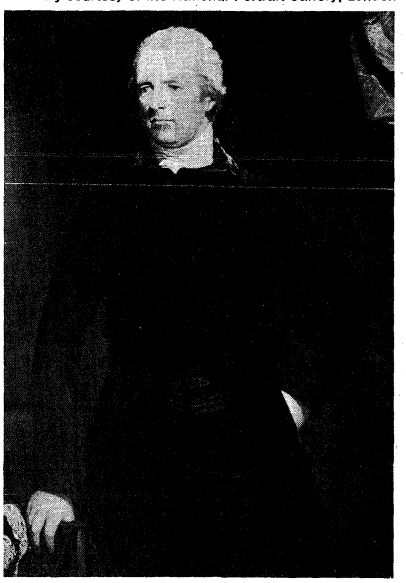
One of the most notable prime ministers of British history, Pitt the Younger, who held office from 1783 to 1801 and from 1804 to 1806, has nevertheless been more highly regarded by posterity than by his contemporaries. Certainly, his political precocity was as obvious in his own time as to later historians; entering Parliament at the age of 21, he became prime minister at 24. But his meteoric rise to power was made possible by the irreconcilable political splits and personal animosities of the period and by George III's determined participation in politics. It was undoubtedly to Pitt's advantage that one of his most brilliant contemporaries, Charles James Fox, had hopelessly antagonized the King by attacking the crown's influence and had also alienated his own natural supporters by his coalition (April–December 1783) with a reactionary politician, Lord North. Pitt's outstanding gifts of oratory were also recognized by his colleagues; even in the late 18th century, the classic age of brilliant British parliamentary debates, he again and again demonstrated his power to dominate the House of Commons. But because of his extreme shyness and his total inability to court popularity, he made little personal impact on his colleagues or the public, and many of his abilities remained unnoticed. It was, therefore, left to the historians of a later age to assert his greatness. Constitutionalists, who could view his work and position in perspective, claimed that his premiership constituted a significant development in the history of the office of prime minister (a title that Pitt himself would never accept), while others have regarded him as a reviver of Toryism. He is also thought of as a great war leader.

Meteoric rise to power

Early life. William Pitt was born at Hayes, Kent, on May 28, 1759, the second son of William Pitt, 1st earl of Chatham, a famous statesman of the mid-18th century, whose energy contributed much to Britain's successful prosecution of the Seven Years' War (1756–63) with France. His mother was Lady Hester Grenville, sister of George Grenville (who headed the government from 1763 to 1765). Both because he was extremely delicate

and because his father disliked public schools, he was educated at home. A precocious boy, he went to Pembroke Hall, Cambridge, at the age of 14. Because of persistent ill health, he made use in 1776 of the antiquated privilege that allowed noblemen's sons to graduate without examination. Left, after his father's death in 1778, with an income of less than £300 a year, he was called to the bar in 1780 and joined the western assize circuit. In September 1780, because of his youth, he failed to secure election to Parliament for Cambridge University but four months later was provided with a seat for Appleby in Westmorland, on condition that he should resign it should his views and those of his patron diverge. Appleby was one of the many pocket boroughs that, before the Reform Act of 1832, were virtually in the control of some great landowner who could ensure that his nominee won the election. Pitt made a very successful maiden speech and, in March 1782, when it was clear that a new ministry would soon be formed, announced with astonishing self-confidence that he had no intention of accepting a subordinate position. He did, in fact, refuse a vice treasurership of Ireland, offered to him by the next prime minister, the Marquess of Rockingham.

By courtesy of the National Portrait Gallery, London



William Pitt the Younger, oil painting by John Hoppner (1758–1810). In the National Portrait Gallery, London.

Under Lord Shelburne, who succeeded as prime minister in July 1782, Pitt became chancellor of the exchequer. With Shelburne's consent, Pitt, in February 1783, sought a compact with Charles James Fox, who had gone into opposition, but Fox absolutely refused to serve under Shelburne. According to his contemporary biographer, his former tutor and friend George Tomline, Pitt then replied that he had not come there to betray Lord Shelburne. This interview marked the beginning of the long political duel between Pitt and Fox and also rapidly brought to an end Shelburne's ministry, since Fox now surprisingly allied himself with Lord North, a politician who had always been willing to work in accordance with the King's wishes, and defeated Shelburne. King George III, unwilling to accept the coalition that would give office to Fox, whom he hated, invited Pitt to form a government; but Pitt declined, knowing he would not have a majority in the House of Commons, and the King had to commission Fox and North. To embarrass the new government, which, under the nominal premiership of the Duke of Portland, consisted of an admixture of reformers and anti-reformers, Pitt brought forward the question of parliamentary reform, with which he had already once, a year earlier, concerned himself. He suggested no extension of the franchise but recommended measures to prevent bribery and to make the representation more realistic. Although his resolutions were defeated, reformers now looked to him, rather than to Fox, as their parliamentary leader.

Loyalty to Lord Shelburne

Pitt's first ministry, 1783–1801. In December 1783, after the defeat in the House of Lords of Fox's East India Bill, George III at once took the opportunity to dismiss the coalition and asked Pitt to form a government. Pitt clearly did not take the premiership as the King's tool, for his first step was to try, on his own terms, to include Fox and his friends in the new ministry. But Fox would not consent to join a government from which his ally Lord North would have been excluded.

When Parliament reassembled in January 1784, the government was at once defeated by 39 votes on a virtual motion of censure, but Pitt refused to resign, and George III was prepared to abdicate rather than again surrender to the Fox–North coalition. Pitt admitted that his situation was without precedent but denied that he was prime minister through backstairs influence. He hung on, and gradually the coalition's majority in Parliament began to crumble; many members, fearing the loss of their seats at a general election, went over to Pitt's side during February and March, doubtless in the hope that he would gain a majority in the existing house sufficient to make a dissolution unnecessary. By March 8 the majority against him was one vote, and on March 25 Parliament was dissolved.

No 18th-century government lost a general election, and Pitt's success in 1784 was never a matter of doubt. The "influence of the Crown" ensured that the new House of Commons was chosen by the Treasury. The statement of Pitt's early-20th-century biographer J. Holland Rose that Pitt "never stooped to associate with jobbers" is no longer accepted. Patronage and corruption gave Pitt a majority, and secret service money was used to pay election bills. Although public opinion aided Pitt in the open constituencies, it is nevertheless misleading to say that he was "the choice of the people"; he was merely the dispenser of royal patronage. Pitt himself was returned for the University of Cambridge; only once again (1790), at subsequent elections, did he have to stand a contest.

Financial policy. When Pitt became prime minister, the national credit was seriously impaired by the heavy cost of the recent American Revolution. The debt was about £250,000,000, a staggering amount for those days. Pitt imposed new taxes to wipe out the deficit, checked smuggling by reducing the high duties that encouraged it, and reduced frauds in the revenue by establishing an improved system of auditing accounts. He also simplified the whole range of customs and excise duties, bringing them into a single consolidated fund, out of which all classes of public creditors were to be paid. In 1786 he introduced a sinking fund on a new principle: an annual surplus of £1,000,000 was to be appropriated to the purchase of stock and allowed to accumulate at compound interest for 28 years, by which time the income from it would amount to £4,000,000 a year. In 1792 another act provided that a sinking fund of 1 percent should be attached to every new loan, which would thereby be redeemed within 45 years. The system worked reasonably well in peacetime because there was an annual surplus of revenue, but, after the outbreak of war in 1793, the government redeemed debt bearing a low interest by fresh borrowing at a higher rate of interest.

Pitt's fiscal policy revealed the influence of the ideas of the political economist Adam Smith, author of *The Wealth of Nations* (1776). In 1785 Pitt tried to bring about the commercial union of the British Isles; Ireland was to purchase the advantage of complete free trade with England and the Colonies by an annual contribution to the cost of defense. But the opposition of British traders caused the plan to be modified in such a way as to compel the Irish to reject it.

Problems with India, Canada, and Australia. Fox's East India Bill had been defeated, but the problems it was designed to solve remained. Britain's increased possessions in India made it necessary for the administration there to be supervised by the government rather than be left in the hands of the commercial East India Company. Pitt, therefore, introduced his own East India bill, which was passed in the summer of 1784. He set up a new

government department, the Board of Control, to supervise the directors of the company. He also ended an inappropriate division of authority in India by making the governor general supreme over the subordinate governments of Bombay and Madras. In 1786 a supplementary act increased the authority of the governor general over his own council. Warren Hastings, governor general of Bengal since 1773, returned home in 1785, having greatly strengthened British power in India, only to undergo the ordeal of an impeachment (trial in Parliament) for his conduct. Pitt honestly believed that there was a case against Hastings and, determined that the British name should be freed from the suspicion of injustice or oppression in the government of Oriental peoples, supported the demand for an inquiry. But those who conducted the impeachment acted with unwarrantable rancour; the trial dragged on for seven years (1788–95) and, although Hastings was finally acquitted, the expenses almost ruined him.

Impeachment of Warren Hastings

Another imperial problem with which Pitt had to deal was that of the future of Canada. By the Constitutional Act of 1791 the then province of Quebec was divided into a predominantly French province of Lower Canada and a predominantly English province of Upper Canada. Pitt, who was in office when men were first transported to Australia, never regarded that country as anything more than a convict settlement.

Foreign policy and war with France. Pitt's foreign policy was only moderately successful. In 1788 he made alliances with Prussia and with Holland, aimed at restricting French influence. But, in effect, the alliance served only one useful purpose: Prussia's diplomatic support enabled Pitt in 1790 to triumph over the Spanish without having to go to war in the Nootka Sound dispute, when the Spanish attacked English ships at what is now Vancouver Island. Thus, the Spanish claim to a monopoly of trade and settlement on the western seaboard of North America was finally destroyed. Pitt's intervention in eastern Europe, however, bore no such marks of triumph. Catherine II of Russia, at war with Turkey, was bent on establishing her supremacy in the Black Sea. In March 1791 Pitt sent her an ultimatum demanding the restoration to the Sultan of all conquests except the Crimea. But his policy of bolstering up the effete Turkish Empire was supported neither by the entire Cabinet nor by public opinion, and the government, badly shaken, had to reverse its policy.

Although the British government clung to neutrality as long as possible, in face of the European wars started by the leaders of the French Revolution, war proved in the end unavoidable. It was not the execution of the French king Louis XVI in January 1793 that made a continuation of peace impossible but was the provocative French decrees of November and December 1792, which authorized their armies to violate neutral territory and which promised military assistance to any European people wishing to depose its rulers. The French, confident of victory after their successes against the Austro-Prussian forces and believing that England was ripe for revolution, declared war on England and Holland on February 1, 1793. Pitt refused to intervene to restore the French monarchy. He fought to protect Britain's vital commercial and colonial interests.

The French Revolution had revived the agitation for parliamentary reform, dormant since a bill introduced by Pitt in 1785 had been defeated. But the cause of reform was soon discredited because its advocates were thought to approve of the violence in France. The unwise demonstrations of the radicals caused the government to have recourse to repressive legislation. In May 1792 a proclamation against seditious publications was issued; and the Habeas Corpus Act, which normally prevented the detention of persons without trial, was suspended in 1794 and remained so until 1801.

Suspension of Habeas Corpus

The crisis over Ireland. The French Revolution had disastrous repercussions in Ireland, too, creating new hatreds to exacerbate the old religious feuds and a rebellion in 1798. As early as 1792 Pitt had held that an ultimate union of the two countries was the only solution of the Irish religious problem; the events of 1798 convinced him that union was most urgently necessary. Large-scale corruption carried the measure through the Irish Parliament, but opposition from Pitt's Cabinet and particularly from the King prevented him from carrying his supplementary proposals—Catholic emancipation and state provision for Catholic and Dissenting clergy. As a result, Pitt resigned on February 3, 1801, and his friend Henry Addington (afterward Viscount Sidmouth), formed a government. The crisis again drove the King insane, and after his recovery in March he accused Pitt of having caused his illness. Pitt replied that he would never again press the Catholic question during the King's reign.

Addington's ministry, 1801–04. Patriotic motives induced Pitt to support the new ministry, but for several months during the session of 1802–03 he never attended Parliament, living in Walmer Castle, where, holding the ancient office of warden of the Cinque Ports, he organized a local volunteer force. In March 1803 Addington invited Pitt to join the government, but Pitt made it clear that he would return only as prime minister. War broke out again in mid-May 1803, and by 1804 Pitt was becoming increasingly critical of the government's financial polity and its measures to meet the growing danger of invasion. Addington's majority fell steadily, and he decided to resign. On April 30 Pitt was informed that the King wished him to plan a new ministry. Pitt replied on May 2 that a nonparty government was desirable but fell in with the King's determination that Fox be excluded and also renewed his pledge not to reintroduce measures for Catholic relief.

Pitt's second ministry, May 1804–January 1806. Pitt's second ministry was weaker than the first, for the Addington group, as well as others, went into opposition. The Third Coalition against Napoleon's France, an alliance with Russia, Sweden, and Austria engineered by Pitt, collapsed after the battles of Ulm and Austerlitz in 1805, and the year closed in disaster, in spite of Nelson's victory at Trafalgar in October, which ended the invasion threat and ensured Britain's naval supremacy for the rest of the war. Pitt's health, never robust, was now failing. He made his last public speech at the Guildhall in London on November 9, 1805. The Lord Mayor proposed his health as "the saviour of Europe." He replied,

> I return you many thanks for the honour you have done me. But Europe is not to be saved by any single man. England has saved herself by her exertions, and will, as I trust, save Europe by her example.

By January 15, 1806, some of his colleagues were determined to force him to resign as the only means of saving his life, and the King was thinking about his successor. He died on January 23, 1806, at Putney and was buried in Westminster Abbey on February 22. A motion for a grant of £40,000 to pay his debts was unanimously carried in the Commons. Earlier (1801), his friends had raised £12,000 in order to relieve him from embarrassment. The famous inscription under his monument in the Guildhall, composed by the statesman George Canning, records that "dispensing for near 20 years the favours of the crown, he lived without ostentation, and he died poor." Careless to a fault about money and engrossed with public affairs, he allowed his large official income to be squandered by servants and tradesmen.

Private life and character. Though eloquent and forceful in Parliament and Cabinet, Pitt made no impact in society and altogether lacked the common touch. He was always notably withdrawn. He never married, and no wholly convincing explanation has ever been put forward of the insuperable obstacle that, at the eleventh hour, prevented him marrying Eleanor Eden, daughter of Lord Auckland. In his career, women played an astonishingly insignificant part. He had few friends; an increasing loneliness, characteristic even when he was a young man, was the self-chosen feature of his private life. Even members of the government complained of his inaccessibility. In 1801 his resignation from office caused extraordinarily little sensation; a contemporary wrote that "nobody speaks of him; no addresses, no subscriptions, no stir of any kind any where." Long before his death, bodily in-

Pitt's personal inaccessibility

firmities, increased by his excessive addiction to port, curtailed his working day. In appearance he was tall and thin, with exceptionally bright eyes and a prominent nose.

Pitt's experience was remarkably limited. He never set foot in Scotland or Ireland; the greater part even of England was unknown to him. He was once in France—for a few weeks. He never came into touch with men of letters or original thinkers; in his official patronage he neglected literature, science, and the arts. Even his friend William Wilberforce, the anti-slavery campaigner, was struck by his ignorance of human nature and his exaggeratedly favourable estimate of men and of political prospects. He was long overconfident of success in every cause he espoused; in the end only the weight of ill-health and Napoleon's great victories of 1805 began to shatter his optimism. Although at first connected with the movement for parliamentary reform, he made no attempts to reintroduce the issue after the failure of his bill in 1785. He made no effort to deal with the social problems caused by the Industrial Revolution; and in all his long years of office, nothing was done to reform the barbarous criminal law, the harsh game laws, prison administration, and local government.

But, to be fair, little was done in these matters until the third decade of the 19th century. Moreover, many of Pitt's limitations stemmed from the real if not apparent misfortune of his becoming premier when little more than a boy. If he had had time to broaden his mind with varied experience and to know his fellowmen, he might have governed the country with more mature views and a greater store of practical wisdom. Nevertheless, by reason of his superb debating powers he dominated the House of Commons, even in that age of notable oratory. His conduct in Parliament had a mixture of prudence, firmness, and transcendent ability never before seen and hardly ever again surpassed.

Historical importance. The constitutional significance of Pitt's career has often been misunderstood. He was not a prime minister of the modern type. At no time was he the leader of a well-organized, coherent party commanding a majority of the House of Commons, which itself owed its existence to the will of the electorate. He was not at all the choice of the nation; he was the nominee of the King, and he retained office only as long as he retained the King's confidence. He had to resign in 1801 because his Irish policy was not acceptable to George III. Even though the inadequacy as a wartime prime minister of his successor, Henry Addington, made Pitt's return to office almost inevitable three years later, Pitt did not return on his own terms but on the King's. He was more dependent on the King's favour than on the support of the House of Commons. His most serious crisis came in the winter of 1788–89, when, during George III's madness, Pitt lost the all-important support of the crown. Had the dissolute Prince of Wales, who favoured the opposition, become regent, Pitt would have been dismissed. Without the support of the crown, neither he nor anyone else could remain long in office. Moreover, there were obvious limitations to his absolute authority in the Cabinet, where colleagues opposed him on all the great questions of the day. And, finally, Pitt had to deal with a sovereign of narrow intellect and with intense and irrational prejudices—though, indeed, these were shared by a great many of George III's subjects.

Although Pitt's supremacy in the Cabinet has often been exaggerated, the necessity for a prime minister who would supervise and coordinate the work of the various departments and possess the chief confidence of the king was never again questioned after his ministries. When the question of his returning to power had been discussed in 1803, he asserted to Viscount Melville that it was absolutely necessary

Pitt's view of the prime ministry

> that there should be an avowed and real minister, possessing the chief weight in the council and the principal place in the confidence of the king. . . . Power must rest in the person generally called the first minister, and that minister ought . . . to be the person at the head of the finances.

Pitt's achievement of this status, while depending upon his forcefulness of character, was only made possible by his very long tenure of office. His total of 19 years in power exceeded by almost seven years the tenure of office, earlier in the 18th century, of Sir Robert Walpole, often regarded as "the first" British prime minister, and that of Lord North, nearer Pitt's own time. In the early 19th century, Lord Liverpool was nominally the head of the administration for 15 years but was never the dominant personality in his cabinets.

It is sometimes claimed that Pitt emerged as the leader of a new Tory Party. Certainly, as a minister who accepted the royal prerogative, he represented the traditions of the Tory, or Court, Party, as distinct from those of the Whigs, who sought to dictate to the crown the choice of its servants; but he was far from being a great party leader commanding the votes of a majority in the House of Commons. He had a personal following of little more than 50. In spite of persistent efforts, great speeches, and the support of powerful and eloquent members, he failed to pass a slave trade abolition bill, a parliamentary reform bill, and Catholic relief bills. Had he led a homogeneous party, this legislation would have been carried; but his supporters always deserted him when he was advocating reforms that either their prejudices or their vested interests opposed.

A cartoon of World War II depicted Pitt and Lloyd George, the prime minister during part of World War I, helping Winston Churchill to climb up and join them on a plinth for famous war leaders. Into this role, too, Pitt's long tenure of office helped posterity to mold him; he was in power during the eight years (1793–1801) of the French Revolutionary Wars and during two years (1804–06) of the Napoleonic Wars. Yet military historians have often severely criticized Pitt for the unprepared condition of the British armed forces in 1793, for his failure to mobilize the country's manpower during the war's progress, and for allowing such military resources as the country did possess to be wasted in ruinous enterprises. Forty thousand troops, sent to capture the French West Indies in the mistaken impression that the French economy would be seriously affected by the loss of the sugar islands, died of yellow fever there. The Duke of Wellington later declared that it was Pitt's great misfortune that he had no one to advise him about the conduct of the war. But alliances formed by Britain were the basis of the great coalitions that waged war against Revolutionary and Napoleonic France. Pitt was the architect of the first three coalitions, those of 1793, 1798, and 1804–05. His steadfast resolution despite all setbacks and his undoubted patriotism justly made him recognized at the time as his country's "guardian angel" and gave rise to various stories associated with him, such as that which alleges that he died murmuring of the plight of his country.

BIBLIOGRAPHY. PHILIP HENRY STANHOPE, 5th Earl Stanhope, *Life of the Right Honourable William Pitt*, 2nd ed., 4 vol. (1862; new ed., 3 vol., 1879), is written in fine literary style and contains some important letters. The standard large-scale biography has been that of J. HOLLAND ROSE: *William Pitt and National Revival* (1911), *William Pitt and the Great War* (1911), and *Pitt and Napoleon* (1912). JOHN EHRMAN, *The Younger Pitt*, vol. 1 (1969), ends at 1789. When completed, Ehrman's biography will replace Rose. Of shorter biographies, the best is LORD ROSEBERRY, *Life of Pitt* (1891, reprinted 1969). PHILIP W. WILSON, *William Pitt, the Younger* (1930), is a semi-popular short life. DONALD G. BARNES, *George III and William Pitt, 1783–1806* (1939), has long been the standard treatment of the political roles of the King and Pitt. New information and interpretations require modification of some views. ALAN F. FREMANTLE, *England in the Nineteenth Century*, 2 vol. (1929–30), is a detailed political history. *The Later Correspondence of George III*, ed. by ARTHUR ASPINALL, vol. 1–5 (1962–70), includes a generous selection of the King's letters (1783–1810) but needs to be supplemented by other correspondence at certain points; *The Correspondence of George, Prince of Wales, 1770–1812*, also ed. by ARTHUR ASPINALL, vol. 1–8, (1963–71), is useful on the parliamentary opposition. Both series have valuable prefaces and footnotes.

(A.As.)

Pius II, Pope

Pope Pius II (Aeneas Silvius Piccolomini) was both an outstanding humanist and pope in his era (15th century). In various ecclesiastical and civil posts he played an important role in European politics before becoming pope. As pope he tried to unite Europe in a grand crusade against the Turks. He recorded the vicissitudes of his life in his many writings in flowing Latin prose.



Pius II, bronze medal by Andrea Guacialoti (1435–95). In the Samuel H. Kress Collection, National Gallery of Art, Washington, D.C.
By courtesy of the National Gallery of Art, Washington, D.C., Samuel H. Kress Collection

From cardinal's secretary to pope

Aeneas Silvius was born October 18, 1405, at Corsignano (later Pienza), near Siena, Italy, into a noble family in very reduced circumstances. He acquired his humanistic education by assiduous work under unfavourable conditions. To earn a living he became secretary to Cardinal Domenico Capranica and went with him to the Council of Basel, a meeting of bishops concerned with church reform (1431–37), which was already at loggerheads with Pope Eugenius IV. With Cardinal Niccolò Albergati he visited many European countries on a diplomatic mission. On returning to Basel in 1436, he became an official of the council, which gave him opportunities to show his great skill as an orator. He became secretary to the antipope Felix V, elected on November 5, 1439, by the remnant of bishops at Basel, who refused to obey Pope Eugenius' order to transfer the council to Ferrara and Florence. As representative of the Basel remnant at the Diet (imperial assembly) of Frankfurt, he attracted the attention of Frederick III of Austria, who invited him to Vienna (1442) and made him imperial poet laureate and his private secretary. Thereupon, he broke his connection with the antipope in 1445 and was absolved of the ban of excommunication that he had been under. A serious illness is said to have led him to amend his dissolute life (he was the father of several illegitimate children). In Frederick's name he proposed to end the rivalry between the papal council at Florence and the rebellious council at Basel by summoning a third council but could persuade neither Eugenius nor the bishops at Basel. Hitherto a layman, Aeneas received sacred orders in 1446. He next managed to calm the storm raised by Eugenius' deposition of two of the German archbishop electors and was chiefly responsible for reconciling the German princes with the Pope and for Frederick's withdrawal of support for the council at Basel.

Made bishop of Trieste by the new pope, Nicholas V, in 1447, he continued his successful mediation between the German states and the Holy See, explaining in a "letter of retractation" his change of role from supporting Basel to being advocate of the papacy. He was transferred in 1449 to the see of Siena, where he was still able to be of service to King Frederick by negotiating his marriage with a Portuguese princess and his coronation as Holy Roman emperor in Rome by Nicholas V (1452).

Nicholas' successor, Calixtus III (1455-58), made Aeneas cardinal-priest of Sta. Sabina as a reward for negotiating peace with Alfonso V, king of Aragon and Naples, and persuading him to cooperate in the crusade against the Turks that Calixtus was energetically promoting. On Calixtus' death Aeneas Silvius was elected pope as Pius II (August 19, 1458). As pope he had one main purpose: to organize a grand crusade to drive back the Turks, who, having captured Constantinople in 1453, were threatening to overrun the rest of Europe. He summoned the Christian princes to a congress in Mantua to study and meet the danger. When he arrived on the appointed day, June 1, 1459, he was alone. Very gradually some came but only to squabble for advantages to themselves.

Abortive crusade against the Turks

The patient diplomacy of the Pope achieved little. A condition for success was to restore peace to the West. In Italy, Pius slowly regained control of the Papal States. His negotiations with France for the repeal of the Pragmatic Sanction of Bourges (of 1438, which made France largely independent of the papacy in ecclesiastical jurisdiction) failed with King Charles VII but succeeded with his son Louis XI (1461). The Greek cardinal Bessarion was sent to Germany (1460) to promote the crusade, but local feuds and wars blocked his efforts. He was equally unsuccessful in Vienna and returned to Rome in the following year but had some success later in Venice. Wars in the Tyrol and discord in Bohemia increased the general unrest. A lull in some hostilities and promises of support from the emperor Frederick and Philip the Good, duke of Burgundy, encouraged the intrepid but sick Pope to proclaim in October 1463 a crusade, which he himself would lead. Pius left Rome on June 18, 1464, for the rendezvous of the armies in Ancona, an Adriatic seaport on the east coast of Italy, where he arrived to find nearly no one. Two Venetian ships arrived on August 11. Pius died during the night of August 14–15. His heart was interred at Ancona, still facing as it were, the infidel East. His body was taken to Rome and there buried in St. Peter's; it was transferred to the church of S. Andrea della Valle when the new St. Peter's was being built.

Pius' achievements

Pope Pius II's great knowledge of conditions in Germany and elsewhere inspired a scheme of wide church reform, but the political conditions of the time and the resistance of the various vested interests rendered its application impossible. He did what he could, particularly by encouraging the movements toward strict observance of the rules of life within religious orders. On the whole, he created worthy cardinals, and on January 17, 1460, he issued a bull condemning appeals from a pope to a general (ecumenical) council of the church. A voluminous writer about the events he had participated in, he wrote also general history and geography, poetry, and at least one scurrilous novel (*The Tale of Two Lovers*). As pope he was a patron of humanists but not to excess either in the use of money or in being complacent about their morals.

BIBLIOGRAPHY

Writings: Pius II's complete works, *Opera omnia*, ed. by M. HOPPERUS, were printed in 1551. Other collections are *Orationes politicae*, ed. by J.D. MANSI (1755–59); and *Opera inedita*, ed. by J. CUGNONI (1883).

English translations: His novel *De duobus amantibus* was much translated (first Eng. trans., *The Goodly History of the . . . Ladye Lucres of Scene in Tuskan*, 1560); new Eng. trans., *The Tale of Two Lovers*, by F. GRIERSON (1929). "The Commentaries of Pius II," ed. by F.A. GRAGG and L.C. GABEL, were published in *Smith College Studies in History*, vol. 22, 25, 30, 35, and 43, abbreviated into the autobiographical account *Memoirs of a Renaissance Pope* (1959). The *De gestis Concilii Basiliensis commentariorum libri II*, ed. by D. HAY and W.K. SMITH (1967), was written in 1440 supporting conciliarism.

Lives: B. PLATYNA, *Vita Pii II in RIS* (Rerum italicarum scriptores), vol. 3, pp. 246–263, by a contemporary curial official; R.J. MITCHELL, *The Laurels and the Tiara: Pope Pius II, 1458–1464* (1962); G. PAPARELLI, *Enea Silvio Picollomini, Pio II* (1950).

(J.Gi.)

Pius V, Saint and Pope

An ascetic and reformer who inaugurated one of the most austere periods in the history of the Roman Catholic Church, Pius V, pope and saint, succeeded in rooting out

the last vestiges of Protestantism in Italy. Pius V was born Antonio Ghislieri on January 17, 1504, at Bosco in northern Italy. He came from a poor family and was a shepherd until the age of 14, when he became a Dominican friar. His first important appointment was as inquisitor, a high office of the Inquisition, then the Roman Catholic Church's judicial system for discovering, examining, and punishing heretics. Ghislieri's methods, prompted by excessive zeal, provoked such opposition from his bishop's officials as well as his chapter that he was recalled in 1550. The chief inquisitor in Rome, Giovanni Pietro Carafa, convinced of his value, sent him on a mission to Lombardy and, in 1551, appointed him commissary general of the Roman Inquisition. When Carafa became pope (as Paul IV), Ghislieri was made bishop of Nepi and Sutri (1556), cardinal (1557), and finally grand inquisitor of the Roman Church (1558). He was continued in this office by Pius IV, whom, however, he antagonized by his censoriousness and obstinacy. After the death of Pius IV, the adherents of strict religious rules, led by Cardinal (later St.) Charles Borromeo, the nephew of Pius IV, had no difficulty making him pope (January 7, 1566). Retaining his ascetic mode of life, Pius immediately began the work of reform. Decrees and ordinances were issued rapidly; the papal court became a model of sobriety; prostitutes were driven from the city or confined to a certain quarter; penalties were fixed for Sunday desecration, profanity, and animal baiting; clerics holding benefices were required to spend definite periods in their administrative districts; members of convents were compelled to live in strict seclusion according to their vows; instruction in the catechism, the short manual outlining the principles of Catholicism, was ordered. A new catechism appeared in 1566, followed by an improved breviary (the daily prayers for clergy and nuns [1568]) and an improved missal (a book containing the prayers and responses for celebrating the mass [1507]). The use of indulgences—*i.e.*, the remission of temporal punishment due for sin—and dispensations from vows was restricted, and the whole system of penance reformed.

Reform activity



Pius V, contemporary medallion. In the coin collection of the Vatican Library.
Leonard von Matt—EB Inc.

Pius was an avowed enemy of nepotism. Though it is true that he made one nephew cardinal, he was allowed to have no influence, and the rest of the family was kept at a distance. By the constitution *Admonet nos* (March 29, 1567), he forbade the reinvestiture of fiefs—those landed estates held under feudal tenure that were intended to revert to the Holy See—and bound the cardinals by oath to observe it. In March 1569 Pius ordered the expulsion of the Jews from the States of the Church, though for commercial reasons they were allowed to remain under humiliating conditions in Rome and Ancona. In February 1571 the Humiliati, a corrupt monastic order of Milan, was suppressed on account of an attempt upon the life of the archbishop, Cardinal Borromeo.

The rules governing the Inquisition were sharpened; old charges, long suspended, were revived; rank offered no protection but rather exposed its possessor to fiercer attack. None was pursued more relentlessly than the intellectuals, among whom many of the Protestant doctrines had found acceptance. Princes and states withdrew their protection of heretics and courted the favour of the Holy See by surrendering distinguished offenders. Philip II of Spain in 1566 surrendered Bartolomé de Carranza, the Spanish theologian and former confessor to Queen Mary of England, and Cosimo de Medici in 1567 gave up Pietro Carnesecchi, the Florentine heretic who had been suspected even during Paul IV's papacy (receiving two years later as a reward the title of grand duke of Tuscany). In March 1571 the special Congregation of the Index, a list of books condemned as dangers to faith and morals, was established distinct from the Inquisition, and hundreds of printers took flight to Switzerland and Germany. The regret of Pius was that he had sometimes been too lenient. He encouraged Philip II of Spain to use the most ruthless tyranny to preserve his Dutch subjects in the Catholic faith and sent troops to France to help Catherine de Médicis repress the Huguenots; he protested against the tolerance shown by the Holy Roman emperor Rudolf II.

Establishment of the Index

In all this work, Pius confirmed Borromeo's belief that a spiritually minded pope was above all else necessary if the decrees of the Council of Trent (1545–63) that were intended to rebut Protestant doctrines and to reform Catholicism were to be put into practical effect. The moral standing of the papacy was greatly raised, and its effectiveness was immensely increased by the obliteration of heresy in Italy, and the morale of the church was much improved by the universal insistence on interpreting church doctrine according to the precepts established at the council. Yet it is open to doubt how much of this improvement was due to Pius' continuous use of the Inquisition. Some of the credit belongs to the new generation of higher and lower clergy in the several provinces whose attitude was so different from that of their predecessors. The council left him to finish the reform of the missal and breviary, but he left the medieval canon virtually unchanged, as it remained until 1970.

Moreover, Pius' policy had practical disadvantages. It took too little account of the wishes of secular rulers at a time when their support was essential for the defense of the church against heresy. He excommunicated Elizabeth I of England and declared her an usurper (February 25, 1570) without possessing the means to enforce his judgment, and he antagonized not only England but Spain, France, and the Holy Roman Empire as well. Much more successful was his organization of a crusade against the Turks, which resulted in a decisive naval victory of Lepanto (October 7, 1571). He made October the month of the rosary because of his victory. Pius died on May 1, 1572, and was canonized by Clement XI in 1712. His feast day is May 5.

BIBLIOGRAPHY. The fullest biography is F.A.P. DE FALLOUX, *Histoire de St. Pie V*, 2 vol. (1844). Slighter biographies (*Vita di Pio V*) are two that appeared soon after his death: GIROLAMO CATENA (1586) and GIOVANNI ANTONIO GABUZIO (1605). There is a study of the Pope's policy towards France by C. HIRSCHAUER, *La Politique de St. Pie V en France 1566–1572* (1922). H. DANIEL-ROPS, *Une Ère de renouvenu: la Réforme catholique* (1955; Eng. trans., *The Catholic Reformation*, 1962), has a good section on Pius V.

(I.F.B./D.Wo.)

Pius IX, Pope

Pius IX, Pope (Giovanni Maria Mastai-Ferretti), also known as "Pio Nono," was pope from 1846 to 1878, the longest pontificate in history. It had far-ranging effects for the development of the Roman Catholic Church during the period from the mid-19th to the mid-20th century. It was in his pontificate that the papacy lost control of the papal states, the dogma of the Immaculate Conception was declared, the Syllabus of Errors of 1864 was issued, the First Vatican Council was convened, and the doctrine of Papal Infallibility was authoritatively defined.

Prepontifical life. Pius IX was born on May 13, 1792, at Senigallia, Italy, the fourth son of Girolamo Mastai-



Pius IX.
Felici

Ferretti, gonfalonier (standard-bearer) of Senigallia, and the countess Caterina Solazzi. He first came into prominence as archbishop of Spoleto from 1827 to 1832, a time of revolutionary disturbance. He was made bishop of the important diocese of Imola in 1832, but it was not until 1840 that he received the hat, as cardinal priest of SS. Piero e Marcellino, the delay being generally ascribed to the disquiet caused by his liberal views. He was not, in 1846, the most prominent liberal candidate likely to succeed Gregory XVI; but it took the conclave only two days to determine his election (on June 16, 1846) and so prevent that of the conservative Luigi Lambruschini. He took the name of Pius in deference to the memory of Pius VII, who had been his friend and who had, like him, been bishop of Imola. The choice was in some sense prophetic, for "Pio Nono," as his predecessor had done, began his career as a supporter of liberal ideas only to learn from bitter experience that liberals often tended to be anticlerical. In 1846, however, all this lay in the future and Europe was agog at the unusual spectacle of a liberal pope.

Early reign. The new pope was confronted by a difficult situation. All Europe, save perhaps Metternich of Austria, considered that the papal states stood in urgent need of reform. A memorandum drawn up in 1831 by the French, Austrian, Russian, and Prussian ambassadors in Rome (with the support of the British representative in Florence) had suggested that councils should be elected to assist in local government, that a central body, composed partly of elected representatives, should control finance, and that the dominant position held by the clergy in the administration and in the judicial system should be terminated. Liberal opinion clung to these measures as absolutely essential throughout the pontificate of Gregory XVI. In addition, the papacy was constantly under attack by Italian nationalists as one of the instruments through which Austria maintained its domination over the peninsula.

Pius recognized that the papal administration, if not so corrupt and despotic as its opponents asserted, was inefficient and intolerant, but he was sufficiently Italian to resent Austrian interference. Above all, he saw the need for the papacy to possess a greater measure of popular support. Yet he was never a revolutionary. He believed that the interests of the church must always come first, and he appreciated that the conservative Catholic powers such as Austria were entitled to consideration for the practical support that they had given the church in the past. That he should eventually disappoint his revolutionary supporters is not surprising, and Pius himself was early aware of this possibility. A criticism more substantial than that of betrayal of the liberal cause is that he chose to depart from the traditional policies of his predecessors without possessing any definite plans for avoiding the pitfalls that must surround the paths of a pope endeavouring to ally himself with liberalism and nationalism. Pius wasted no time. He declared a general political amnesty on July 16, 1846, and followed it by a number of practical reforms, including a reorganization of the criminal procedure. Some financial reforms of an administrative character came next, and the censorship was mitigated. Finally came Pius' most daring innovations, the political reforms announced in 1847. The council of ministers became a cabinet of departmental heads of which only the secretary of state was of necessity to be a cleric, it was to be advised by a council of 24 elected members, and a municipal government was established for Rome that included an elected representative body. At the same time, the papacy brought support to the cause of Italian nationalism when Pius ordered the Austrian government to withdraw from Ferrara the reinforcements sent to its garrison there and opened negotiations with Piedmont for a national customs union.

Papal responses to Austrian domination of Italy

The revolution of 1848. The year of revolutions, 1848, began with revolution in Sicily; soon all Europe was ablaze and Pius was faced with demands, both liberal and nationalist, much beyond what he had been prepared to grant. On March 14 he was compelled to grant a constitution establishing a two-chamber parliament with full legislative and fiscal powers subject only to the pope's personal veto. On March 23 Charles Albert of Sardinia (Piedmont) declared war on Austria. For a time Pius continued to endeavour to steer a middle course, claiming in his address to the cardinals of April 29 that he was a disinterested spectator of the revolutionary activities sweeping Italy and that his program of reform was merely the fulfillment of the program long pressed upon the papacy by the powers. In the atmosphere of the time such sentiments were judged as displaying absolute hostility to the national cause, and the papacy was never again able to appear in Italy as anything other than a bulwark of reaction.

To prevent revolution from breaking out in Rome itself, Pius consented to the appointment of popular ministries, but none of the appointees were able to control the situation. A steadily deteriorating situation culminated in the assassination of one of them, Pellegrino Rossi, on November 15. A radical ministry was appointed under Emanuele Muzzarelli and Giuseppe Galletti, and when the Swiss guards were disbanded the pope was a virtual prisoner. On November 24–25, with the aid of the French and Bavarian ambassadors, he fled to Gaeta in the kingdom of Naples. In his absence, elections were held for a constituent assembly; and this, on February 9, 1849, declared the temporal power at an end and a democratic republic to be established. The papacy thereupon issued a formal appeal to the rulers of France, Austria, Spain, and Naples, but not Piedmont, for assistance. Although it was generally considered that the pope's restoration could take place only with some sort of undertaking to maintain constitutional government in the papal states, and although Louis-Napoleon, the newly elected president of France, was in favour of such a policy, Pius held out against any concessions and asserted his determination to exercise his temporal power without any restrictions whatsoever. The upshot of a period of military and diplomatic manoeuvre on the part of France and Austria was the unconditional restoration of papal rule by the French general N.-C.-V. Oudinot on July 15. The revolution was over, and Pius himself returned to his capital on April 12, 1850.

The Roman question. It has often been asserted that Pius returned to Rome a changed man, that the former liberal had become a narrow reactionary. That his policy had changed there is no doubt, but his fundamental attitude remained the same. The interests of the church had always been his first concern, and he could never be an unconditional ally of nationalism or liberalism. He had been prepared to countenance both while they left the church intact, but experience had taught him that both led to revolution, which he had never been prepared to countenance. Furthermore, political concessions on his part had led to attacks on his spiritual power, and he considered that it could be protected only by his continued exercise of a temporal authority. Once these two aspects of his dominion had become indissolubly linked, it is easy

to see why Pius considered himself obliged to oppose any alteration of his position as a temporal ruler.

In 1846 Pius had considered that a new departure was necessary to meet the legitimate demands for reform within the papal states and perhaps also those for a change in the Italian system of states; he had experimented and failed, and he now considered that he could do no more. But the problems that he had tried to solve remained, and to substitute no policy at all for one that had failed was no answer. The new secretary of state, Cardinal Giacomo Antonelli—a layman, to whom diplomacy was now entrusted, since Pius recognized his own inadequacies in the field—considered throughout that, in endeavouring to retain the temporal sovereignty untouched, he was fighting a losing battle, but that he had no practical alternative to a policy of inaction. So, not for the first time in history, the papacy was a spectator while events changed the map of Europe.

Role of papal secretary of state

Most of the administrative reforms carried out immediately after Pius' accession remained, and the papal territories benefitted from the general increase in European prosperity after 1850. But constitutional government was never restored; the amnesty granted on the pope's return was so riddled with exceptions as to be worthless; and to all expressions of national sentiment the papacy was utterly hostile. It was not that papal government was tyrannical but that it formed an absolute barrier in the way of Italian unification upon which politically minded Italians were set. So relations with Piedmont, where Count Camillo Cavour became chief minister in 1852, grew steadily worse. What began as a dispute over the terms of a new concordat took on a much more serious aspect when Cavour found it politically expedient to encourage anticlerical feelings during the passage of a law in 1855 dissolving a large number of religious houses. Certain of papal opposition, Cavour began in the same year his plans to achieve Italian unification under the leadership of Sardinia-Piedmont alone. At Plombières in 1858 he agreed with Napoleon III to combine to eliminate Austria from the Italian peninsula and so to settle the fate of Italy between them. Pius declared his neutrality in the war which broke out in April 1859, fearing that the victory of either of his protectors would leave him at its mercy. The Austrian defeat and the armistice of Villafranca resulted in the pope's losing control over the Romagna but nowhere else. Sporadic negotiations for the establishment of an Italian confederacy under the presidency of the pope went on during the winter of 1859–60, but the pope's refusal to contemplate the loss of any part of his dominions brought them to an end. In January 1860 Cavour (who had resigned in protest against the armistice) returned to power, possibly as a direct consequence of the pope's intransigence, and agreed with France to cede Savoy and Nice in return for a free hand everywhere in Italy except at Rome itself. The newly raised papal army was defeated at Castelfidardo on September 18, and the papal states were quickly overrun. French troops protected the Patrimony of St. Peter, which alone among his dominions still remained in the pope's hands.

The loss of temporal power

For ten more years the remnant of the temporal possessions were retained by virtue of the presence of a French garrison at Rome. Cavour himself was favourable toward an agreed settlement of the Roman question on the basis of a continuing papal sovereignty over all the states of the church but with its exercise entrusted to the king of Italy. A start was made to begin negotiations, but then the pope suddenly withdrew. In less than a year Cavour had died (June 1861), and it was clear that the pope would be able to hold his remaining territories just so long as Napoleon III was willing and able to maintain the French garrison in Rome. When the Franco-German War broke out that was no longer possible. On Sept. 20, 1870, Italian troops occupied the city; and in October a plebiscite was held in which an overwhelming majority of the votes cast were for the incorporation of Rome in the kingdom of Italy. Pius remained for the rest of his days a prisoner, as he regarded himself, in the Vatican. He refused any intercourse with the Italian government, so that their relations rested upon a law passed by the Italian parliament in November. The sovereignty of the pope was declared to be untouched by the loss of his dominion in compensation for which he was to receive an annual sum of money. He was to be entitled to conduct his own diplomatic relations with other powers and to have exclusive authority within the Vatican itself and a small district around it. In the rest of Italy, church and state were to be separated. So, though the papacy did not formally recognize the fact until the concordat of 1929, the Roman question had been settled.

Ultramontanism. Important as the events just described were for the papacy, the doctrinal developments of Pius' pontificate, which spring directly out of these political disasters, constitute its most significant contribution to papal history. Ultramontanism began with Joseph de Maistre, as a reaction against Gallicanism and against Josephinism, seeking to free the church from the chains of secular control by binding it more closely with the papacy. H.-F.-R. de Lamennais developed it by suggesting that the church would benefit from a general increase in political freedom. Gregory XVI condemned Lamennais's teaching because he saw that freedom might mean freedom to deny religion altogether. Pius IX decided in 1846 to experiment with liberalism but later became convinced that Gregory XVI had rightly suspected it. Nevertheless, if Italy taught Pius one lesson, developments in France, where the church prospered more under the liberal regime of Louis-Philippe than it had under the clerical Charles X, suggested quite the opposite conclusions to the liberal Catholics there, whose spokesman was Charles de Montalembert. On the other hand, the coming of the Second Empire stimulated the party led by Louis Veuillot, whose Ultramontanism was of the older sort, completely divorced from liberalism and seeking freedom for the church in an authoritarian state that would guard it against revolution.

For a period after 1850, Pius' policy took little heed of either brand of Ultramontanism. Antonelli followed the paths of Consalvi, Pius VII's secretary of state, in seeking to procure more favourable concordats with Catholic rulers. Most important was that with Austria (1855), the home of Josephinism; and others were concluded with Spain (1851, 1859), with Portugal (1857), with Württemberg (1857), with Baden (1859), with Nassau (1861), and with a number of Latin-American states (1852–62). Such agreements might be politically valuable but were no defense against intellectual anticlericalism; and Pius became increasingly convinced that the real danger to the church lay in the modern secular ideas that the liberal Catholics were endeavouring to incorporate into its doctrines. The events of 1860 finally convinced him that the notion of a "free church in a free state," as dear to Montalembert as it was to Cavour, was a snare. The encyclical *Jamdudum Cernimus* (March 18, 1861) denounced not only Piedmontese aggression but all modern political doctrines, which, it said, encouraged non-Catholic religions, allowed full freedom to attack the church in print, and sought always to undermine the authority of the church. The risorgimento not only convinced Pius that liberalism in the church must be destroyed but also placed the liberal Catholics in the difficult position of appearing to support those who had caused him so much distress. In every way the ground was being cleared for a new departure. The alternative to Montalembert's doctrine was no longer an unconditional attachment to the principles of the *ancien régime* but a new kind of Ultramontanism, asserting the need for concentrating church authority in the pope's hands. The ground was being prepared for the first Vatican Council and the doctrine of papal infallibility.

Pius' attack on liberalism

But first the strong liberal party in the Catholic Church had to be defeated. In 1863 Montalembert was invited to address a large Catholic congress at Malines, and he took the opportunity to defend the concept of a free church in a free state and to condemn intolerance in principle, citing as an example the revocation of the Edict of Nantes. Pius was content in reply to point out that on these two points he was running counter to authoritative pronouncements of Pius VI and Gregory XVI. This was sufficient to deter Montalembert from accepting a second invitation to Malines in 1864, but his supporter F.-A.-P. Dupanloup

proved an able substitute. Meanwhile, at a congress at Munich in 1863, J.J.I. von Döllinger had pleaded for the right of a scholar to pursue independent inquiry. It was clear to everyone that the church stood in need of authoritative pronouncements about its relations with the state and with modern society, and discussion began about the possibility of calling an ecumenical council for this purpose. But once again the Roman question intervened decisively in the struggle.

On Sept. 15, 1864, the French and Italian governments came to an agreement whereby the French garrison was to be withdrawn from Rome within two years. In fact, it was not finally withdrawn until the outbreak of the Franco-German War, but the conclusion of the September convention was sufficient to make Pius decide to take immediate action against liberalism. On Dec. 8, 1864, he issued the encyclical *Quanta Cura* with, attached to it, the famous *Syllabus* listing 80 of the "principal errors of our times." As the errors listed had already been condemned in allocutions, encyclicals, and other apostolic letters, the *Syllabus* said nothing new and so could not be contested. Its importance lay in the fact that it published to the world what had previously been preached in the main only to the bishops, and that it made general what had been previously specific denunciations concerned with particular events. Thus perhaps the most famous article, the 80th, stigmatizing as an error the view that "the Roman Pontiff can and should reconcile himself to and agree with progress, liberalism, and modern civilization," sought its authority in the pope's refusal, in *Jamdudum Cernimus*, to have any dealings with the new Italian kingdom. On both scores, the *Syllabus* undermined the liberal Catholics' position, for it destroyed their following among intellectuals and placed their program irrevocably out of court.

The *Syllabus*

Though Dupanloup tried to explain away the *Syllabus* by insisting upon its context and by stressing its purely negative aspect (thus, though the concept of toleration as the final and universal answer was condemned, it did not follow that intolerance was universally valid), the *Syllabus* nevertheless dealt a mortal blow at liberal Catholicism, which ceased after 1864 to be the main issue taxing Catholic controversialists. While some of Louis Veuillot's followers hoped that at the forthcoming council a positive statement of the orthodox doctrine of the position of the church in society would replace the negative denunciations of the *Syllabus*, the majority looked upon that battle as won and so turned to the question of defining the pope's infallibility, the keystone of the neo-Ultramontane program of centralizing the authority of the church in Rome so as to escape from the control of the secular state. Conversely, the former liberal Catholics were ranged in opposition to this project. When the council at last met, on Dec. 8, 1869, everything was ready for a decisive engagement.

The Vatican Council. In the doctrine of papal infallibility itself there was nothing new. It had been employed to define, on Dec. 8, 1854, the dogma of the Immaculate Conception, which asserted that the freeing of the Virgin Mary from all taint of original sin had occurred at the moment of her conception. The pope had previously made extensive inquiries among the bishops and other divines and there was little opposition to such an exercise of his undoubted prerogative. When, however, at a gathering of bishops and other dignitaries of the church in Rome in 1862 and again at another in 1867 it had been suggested that the doctrine of infallibility should be authoritatively defined, Dupanloup had led a successful opposition to the project. It was objected that such a definition was inopportune, tending to widen the breach between the church and modern society, and that it would present a one-sided view of the source of authority in the church; for while the pope possessed powers issuing directly from God, so too did the bishops, for instance, whose ordinary jurisdiction arose, not out of their nomination or institution (whether by the state or by the church) but equally from divine origin; so that the pope's powers ought not to be defined without reference to other aspects of the nature of the church. The criticism that must attach to Pius is that he allowed, even encouraged, the council to put aside discussion on the wider issue, which was its original program, in favour of the narrower definition. This was, of course, precisely what the Ultramontane party desired. The Ultramontanes, indeed, undoubtedly possessed the backing of by far the greater part of the church, partly because of the reaction engendered by the political misfortunes of the last decade, partly because of the immense prestige enjoyed by Pius as a result of his long and tragic pontificate, but to a larger extent because of the contemporary movement away from intellectualism and in favour of devotional religion. Pius himself was at the front of this movement and received strong support from the Jesuits, whose influence was most emphatically increasing.

The dilemma of the church

The first Vatican Council opened on Dec. 8, 1869, the day of the feast of the Immaculate Conception, which Pius chose for the majority of the important acts of his pontificate. The opposition, consisting of the German, French, and U.S. bishops, was strong enough to prevent a definition of the doctrines and nature of the church on the lines suggested by the *Syllabus*; but the Ultramontane party, led by Archbishops H.E. Manning (of Westminster) and V.A. Dechamps (of Malines) brought forward the question of infallibility, upon which their position was much stronger. Pius intervened decisively to alter the procedure of the council on Feb. 20, 1870, and again on April 29, and to make known his personal support for Manning and Dechamps. The outcome was to postpone all deliberation except that upon infallibility. The decisive vote came on July 13 when 451 voted for it, 88 against it, and 62 in favour of some amendment. Thereupon the minority left Rome and the final definition was carried on July 18 by 533 votes to 2 (see also OLD CATHOLIC CHURCHES). Infallibility was confined to those occasions upon which the pope made pronouncements *ex cathedra*; but these pronouncements were to take their authority from God and not from the consent of the church, and no mention was made of any other divine authority existing in the church except that of the pope. There is no doubt that this act secured the triumph of the Ultramontane party or that the decisive element in it was provided by Pius himself.

Definition of papal infallibility

Last years. Pius reigned for another eight years, during which he became further estranged from the Italian government and witnessed a general outbreak of anticlericalism in western Europe. In Germany this culminated in Bismarck's *Kulturkampf*, which Pius condemned in the encyclical *Quod Nunquam* of Feb. 5, 1875, leaving the solution of the problem to his successor. Pius died on Feb. 7, 1878, having seen in his long pontificate the creation of the modern papacy.

Evaluation of his pontificate. The exact responsibility of Pius for the events of his pontificate is still a matter for controversy. The result is clearer. Church and state were finally separated, authority in the church was centralized in Rome, and the church was ranged in opposition to the dominant political forces. On that score Pius was condemned by many contemporaries; the 20th century, however, was to see Italian nationalism develop into Fascism and revolutionary principles into Communism. Whether the separation of church and state reinforced anticlericalism or inoculated the church against its effects is less certain.

But perhaps the most damaging criticism of Pius is that he assumed almost the sole direction of events while never appreciating their significance. The great events of his pontificate occurred despite him rather than because of him. Thus he condemned in the *Syllabus* that very separation of church and state which it was the work of the first Vatican Council—to which for the first time the secular Catholic rulers were not invited—to secure. At one time or another Pius was served by the leaders of all the important movements of his century, by liberal cardinals like Pasquale Gizzi, by ecclesiastical statesmen like Antonelli, and by the ascetic Manning, a convert from Protestantism. Yet it is unfair to say that he changed his political principles; rather he never possessed any, save in his personal religious life, that were not rather naïve, and so he

was always open to be guided by experience. In view, however, of the benefit that has accrued to the church from the loss of the temporal power, it should at least be asked whether it was not Pius' determination to maintain it at all costs that contributed most to the temporal power's being lost.

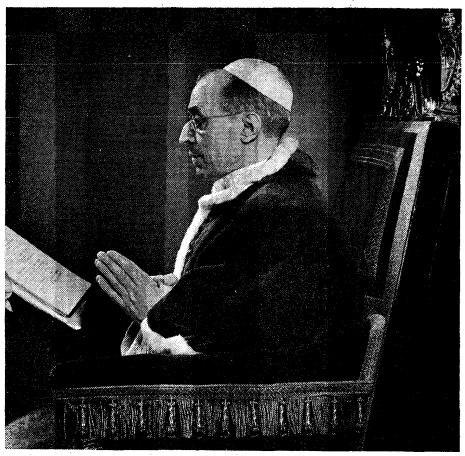
BIBLIOGRAPHY. The best biography in English, *Pio Nono: A Study in European Politics and Religion in the Nineteenth Century* by E.E.Y. HALES (1954), was written avowedly to explain events from the viewpoint of Pius IX. It is a scholarly work, enriched with a bibliography covering not only the pontificate but the wider Italian and European context. FERNAND HAYWARD, *Pie IX et son temps* (1948), is the work of a French scholar, favourable to the Pope but showing much research. For a less favourable view, see *The Roman Question: Extracts from the Despatches of Odo Russell from Rome*, 1858–1870, ed. by NOEL BLAKISTON (1962). A.C. JEMOLO, *Chiesa e Stato in Italia negli ultimi cento anni* (1948), is a massive study, particularly strong in the study of ideas. ROGER AUBERT, *Le Pontificat de Pie IX* (1952), is thoroughly reliable, as is the same author's *Vatican I* (1964). ABBOT CUTHBERT BUTLER, *The Vatican Council*, 2 vol. (1930; new ed. by CHRISTOPHER BUTLER, 1 vol., 1962), based on correspondence with the English bishop Ullathorne, gives a good account of Pius' activity at the council. More specialized studies are *The Syllabus of Errors* by the American scholar, DAMIAN MCELRATH (1964); and *Pio IX e Leopoldo II* by GIACOMO MARTINA (1967).

(I.F.B./D.Wo.)

Pius XII, Pope

Elected in March 1939, Pius XII (Eugenio Pacelli) was pope during World War II and during the years of postwar reconstruction. For 20 years he was a copious speaker and writer, and his ideas on international structures and reforms in liturgy and biblical studies were important, but his conduct during World War II has been the subject of controversy.

© Karsh—Rapho Guillumette



Pius XII.

Early life and career

Eugenio Pacelli was born in Rome on March 2, 1876. His family came from Tuscany and had supplied Vatican lawyers since 1819. Eugenio was educated at Roman day schools and studied for the priesthood at the Gregorian University there, though for reasons of health he soon lived at home, not at his college of Capranica. He was ordained in 1899. He was a successful student with a quick mind and an exceptional memory, although the theological formation then available at the Gregorian University was limited. It was in the subject of law that he made his mark and so passed into the Vatican Secretariat of State. There he rose rapidly and in 1917 was made archbishop and sent as nuncio—a diplomat—to the Bavarian Court to negotiate a concordat (a Vatican-state agreement). This done, he passed (1925) to Berlin with the same aim and remained there until he was named cardinal, by Pius XI, and recalled to Rome at the end of 1929 to become secretary of state. After ten years in this office, during which he travelled on papal missions to North and South America and to France, he was elected pope in the shortest conclave since 1623.

These experiences are crucial in assessing Pacelli's pontificate. He grew up in a home of deep piety and of devotion to the papacy but also in a Rome smarting from the loss of the papal states, a Rome still clerical, resentful, and suspicious of the secular world. He inherited a diplomatic tradition and technique centred on the principle of the concordat, which aims to preserve the church's privileges and freedom of action, even under regimes irreconcilable with Christian principles.

Fascism came to power in Italy (1922) during Pacelli's absence in Germany, and his brother, a lawyer, helped to fashion the concordat with the dictator Benito Mussolini—to achieve which the Catholic Popular Party and any chance of ousting Mussolini were sacrificed. This settlement of the Roman Question, which created the Vatican city state (1929), satisfied Pius XI's ambition and that of many Italians but created many problems for the next 30 years in Pacelli's life, both in relations with the increasingly hostile dictatorship and with the uneasy democracy that succeeded it after the war.

Twelve years in Germany had made Pacelli fluent in German and had given him great love for the Germans. A tireless and methodical worker, he admired such virtues in the Germans and had many German helpers in the Vatican. But his deep knowledge of German affairs precluded illusions about Nazism, and the concordat with Hitler's Germany (1933), largely his work, was (he often said) a calculated risk, aimed at preserving a platform for Catholic life and ministry in a hostile German society.

His part in the anti-Nazi encyclical *Mit brennender Sorge* (1937; "*With Burning Sorrow*") his dozens of other notes and memoranda to the German Government as secretary of state, his frigid and outspoken reception of the Nazi foreign minister Joachim von Ribbentrop in Rome, his bitter reproach of the Austrian Cardinal Theodor Innitzer's weakness in face of the Austro-German *Anschluss*—the union of Austria and Germany—in 1938, and finally the unfavourable German reception of his election as pope, all showed his true estimate of the German tyranny.

During the few months between his election and the outbreak of war, Pius XII turned his diplomatic gifts to preventing the catastrophe, but not in the spirit of appeasement—the Vatican had disliked the Munich Agreement (1938) by which Czechoslovakia was sacrificed to expanding German power by Britain and France. Pius especially strove to keep Italy neutral and was deeply saddened when he failed.

Significance of papal reign

The demands of war harmonized with Pius' austerity and immense capacity for work; they enhanced his exalted, almost dramatic, conception of his office. His untiring efforts to humanize war and to relieve suffering, anxiety, and grief, though they did not always receive Axis cooperation, have been generally applauded. In a notable series of Christmas broadcasts (1939–48), he looked firmly toward the postwar future and to the opportunities it would offer for a new order in which the international community would be given real expression, replacing the egoistic nationalism that had bred war.

There has been less agreement about Pius' conduct and utterance in relation to the war itself. Deeply conscious of paternal responsibilities that transcended nationalism, pledged to neutrality, obliged to condemn unchristian conduct and yet fearful of provoking reprisals on the defenseless—the Pope clearly suffered many agonies of indecision, aggravated by his dislike of strife. He has been charged with neglecting, in the interests of a calculated neutrality, to raise an authoritative voice in defense of the persecuted, Christian or Jewish. Others have claimed that protests would have aggravated Nazi policy without achieving any good result. Both approval and censure of his choices during these difficult years have often been facile and superficial: the deeper question may be asked whether the nuances, the hints, the oblique allusions of traditional diplomacy, which came naturally to him and which often seemed to him much clearer than they seemed to others, were any longer appropriate to the savagery and cynicism of a death struggle for world domination and of deliberate genocide. The problems of

a neutrality that is not completely silent are never wholly surmountable, since causes are not morally neutral.

Though Pius refused to be drawn into any crusade against international Communism, he enacted severe measures against Catholics collaborating with Communists (1946); and the precarious balance between Christian Democracy and the extreme left in postwar Italy led him to encourage the Catholic Action leader Luigi Gedda, whose meddling in politics embarrassed the parliamentary Christian Democratic Party. Premier Alcide De Gasperi's successful protest against this (1949) earned him disfavour at the Vatican. When Pius' failing health left power in the hands of a "bureaucracy" of cardinals, including the autocratic head of the Holy Office, Alfredo Ottaviani, clerical interference in Italian public life reached a high pitch.

Many of Pius' innumerable addresses to specialist audiences dealt with urgent moral problems; by the end of his life he was moving toward outright rejection of modern war and doubted the justification of a war deterrent policy. In the field of marital relations, he made a number of pronouncements about the "safe period" without carrying discussion of birth regulation substantially beyond the position of his predecessor. Nor did he advance Catholic thinking or policy in the ecumenical field, though slight concessions were made in 1949.

Perhaps Pius' greatest utterance was the encyclical *Divino Afflante Spiritu* (1943; "With the Help of the Divine Spirit"), which gave fresh impetus and direction to Catholic biblical studies. These studies had been hampered by the inquisitorial atmosphere persisting since the days of crisis under Pius X over Modernism—a movement that sought, among other things, to apply modern methods of historical criticism to both the Bible and the life of Christ, but was vitiated by arbitrary philosophical assumptions. The encyclical welcomed most of the techniques of modern biblical scholarship.

Mediator Dei (1947; "Mediator of God") was a similar charter for the liturgical movement; he aimed at restoring a community character to public worship but mistrusted extravagant experiment. More conservative and suspicious views prevailed in *Humani Generis* (1950; "Of the Human Race"), touching on theological tendencies.

Pius' later ill health accentuated his solitary bent, increased the power of his conservative curia—the papal administrative bureaucracy—and aroused misgivings in some countries, not least in France, where Roman handling of the worker-priest problem—by sending worker-priests, who had been sharing the work and living conditions of labourers in order to extend their ministry, back to traditional religious tasks—caused deep resentment. When he died, at Castel Gandolfo in 1958, respectful mourning was worldwide, but Rome, showed few signs of the revolution in church life around the corner. Though he was gifted and deeply spiritual, history is unlikely to place him among the great creative popes.

BIBLIOGRAPHY. J.O. SMIT, *Pope Pius XII* (Eng. trans. 1951); OSCAR HALECKI and J.F. MURRAY, *Pius XII: Eugene Pacelli, Pope of Peace*, rev. ed. (1954); and N. PADELLARO, *Portrait of Pius XII* (Eng. trans. 1956), are conventionally laudatory and are works written in Pius' lifetime. English translations of two works of C. FALCONI, *The Popes in the 20th Century from Pius X to John XXIII* (1968), and *The Silence of Pius XII* (1970), assume an impartial manner but are basically hostile to Pius. R. HOCHHUTH, *The Representative* (Eng. trans. 1963), is a bitterly hostile caricature. Other works include: S. FRIEDLANDER, *Pius XII and the Third Reich: A Documentation* (Eng. trans. 1966); J. NOBECOURT, *"Le Vicaire" et l'histoire* (1964); D. TARDINI, *Pio XII* (1960); and W.A. PURDY, *The Church on the Move: The Characters and Policies of Pius XII and John XXIII* (1966).

(W.A.P.)

Pizarro, Francisco

Francisco Pizarro, the Spanish conqueror of the Inca empire of Peru, a man personifying an epoch that saw the extension of Spanish power to the greater part of the Caribbean and Central and South America, possessed in full the qualities of physical strength, valour, cunning, and leadership that were essential to conquest. With only 180 men, he landed in Peru, ascended the Andes, and captured the seats of Inca power in the cities of Cajamarca and Cuzco.



Francisco Pizarro, engraving by Crispijn van de Passe (1560–1637).
By courtesy of the Metropolitan Museum of Art, New York, Whittelsey Fund, 1951

Early life

Pizarro was born about 1475 in Trujillo, Spain, a small town some 130 miles (209 kilometres) southwest of Madrid. The illegitimate son of Captain Gonzalo Pizarro and Francisca González, a young girl of humble birth, he spent much of his early life in the home of his grandparents. According to legend he was for a time a swineherd, a not unlikely possibility since this was a common occupation of boys in that region. He doubtless participated in local manorial wars and, when these were ended, very probably went to fight in Italy. Certainly in 1502 he went to Hispaniola (modern Haiti and Dominican Republic) with the new governor of the Spanish colony.

Pizarro had little inclination toward the settled life of the colonizer, and in 1510 he enrolled in an expedition of the explorer Alonso de Ojeda to Urabá in Colombia. He appears to have been marked out as a hard, silent, and apparently unambitious man who could be trusted in difficult situations. Three years later, acting as captain, he participated in an expedition led by the explorer Vasco Núñez de Balboa that was credited with the discovery of the Pacific. From 1519 to 1523 he was mayor and magistrate of the newly founded town of Panamá, accumulating a small fortune.

Discovery of Peru. It was not until 1523, when he was some 48 years old, that Pizarro embarked upon the adventure that was to lead to his lasting fame. In partnership with a soldier, Diego de Almagro, and a priest, Hernando de Luque, he made preparations for a voyage of discovery and conquest down the west coast of South America. The initial period of exploration took place from 1524 to 1527, during which time Pizarro and his men endured many hardships. After eventually sighting a raft occupied by what appeared to be civilized Indians dressed in manufactured cloth, they decided to send Almagro back to Panama for reinforcements. The new governor of Panama, however, sent back orders that the expedition be abandoned in order that no more lives be lost. At this point Pizarro is reputed to have drawn a line on the ground with his sword, inviting those who desired wealth and glory to cross it. The "famous thirteen" who did cross the line continued their exploration of the coast as far as 9° S, obtaining distinct accounts of a great Indian empire as well as many Inca artifacts. They christened the new land Peru, probably a corruption of the name of a coastal river, the Virú.

Finding the Governor of Panama still opposed to their now promising enterprise, the explorers decided that Pizarro should go to Spain to ask Charles in person for permission to undertake conquest. Sailing in the spring of 1528, Pizarro was in Seville at the same time as Hernán Cortés, conqueror of Mexico, and was able to win Charles over to his scheme. He was decorated, granted a

coat of arms, and, in July 1529, made governor and captain general of the province of New Castile for a distance 600 miles south of Panama along the newly discovered coast. Pizarro was invested with all the authority and prerogatives of a viceroy, and Almagro and Luque were left in subordinate positions. All the famous thirteen received substantial rights and privileges in the new territories.

The conquest. Joined by four of his brothers, Pizarro sailed for Panama in January 1530 and by January of the following year was ready to set off for Peru. He set sail with one ship, 180 men, and 37 horses, being joined later by two more ships. By April they had made contact with emissaries of Atahuallpa, emperor of the Incas, who was residing near the city of Cajamarca with an army of about 30,000 men. Somewhat scornful of Pizarro's small force, the Inca accepted a proposal that the two leaders meet in that city.

Capture of Atahuallpa

Arriving on November 15, Pizarro immediately set up his artillery and sent his brother Hernando and another Spaniard to request an interview. After a day of tense waiting, Atahuallpa, borne on a litter, entered the great square of Cajamarca with an escort of between 3,000 and 4,000 men, who were either unarmed or carrying short clubs and slings beneath their tunics. Pizarro sent out a priest, Vicente de Valverde, to exhort the Inca to accept Christianity and Charles V as his master. Atahuallpa disputed both the religion and the sovereignty of the Spaniards and, after examining a Bible offered by the priest, flung the book to the ground. Valverde reported these events to Pizarro, who immediately ordered an attack. The astonished Incas were cut down from all sides, Pizarro himself seizing Atahuallpa.

Pizarro was now faced with the problem of maintaining his position until reinforcements could arrive from Panama. Atahuallpa was held as hostage and failed to win his release though he fulfilled a promise to fill the chamber in which he was held with gold and silver. Accused of ordering the execution of his brother Huascar, a rival for the title of Inca, and of plotting to overthrow the Spaniards, Atahuallpa was put to death by strangulation on August 29, 1533. With news of Atahuallpa's death, the Inca armies surrounding Cajamarca retreated; and Pizarro was able to progress towards Cuzco, the royal capital, which was occupied without a struggle in November 1533. The Spaniards declared Manco Capac, Huascar's brother, as Inca.

For the remainder of his life, Pizarro was engaged in consolidating the Spanish hold on Peru and in defending his and his brothers' share of the spoils. A certain enmity and rivalry developed between him and Almagro as a result of Pizarro's overriding powers from the king of Spain. This contravened a solemn agreement between the original three partners that the spoils of the expedition should be shared equally. Almagro at one stage seized Cuzco but was persuaded by Pizarro to depart for Chile, over which he had been granted extensive powers by the King. Disappointed by the poverty of that country, however, he returned to Peru, where he was made prisoner and later executed by Pizarro's brother, Hernando.

Pizarro himself was in Lima, a city that he had founded in 1535 and to which he devoted the last two years of his life. Almagro's former adherents had grouped around Almagro's son in Lima, where they were confined and watched. Suspecting that they were to be eliminated, they decided to move first, attacking Pizarro's palace on June 26, 1541. Pizarro died that day a protracted death, drawing a cross of his own blood on the ground, kissing it, and crying "Jesus" as he fell.

BIBLIOGRAPHY. Although not strictly concerned with Pizarro's actual life story, WILLIAM H. PRESCOTT, *History of the Conquest of Peru*, 2 vol. (1847); and SIR CLEMENTS R. MARKHAM, *A History of Peru* (1892), are two classic accounts of the Pizarro brothers' exploits. Modern appraisals of the Pizarros' feat include HAMMOND INNES, *The Conquistadors* (1969); and the magazine HORIZON, *Pizarro and the Conquest of Peru* (1970), both lavishly illustrated; MANUEL BALLESTEROS GAIBROIS, *Francisco Pizarro* (1940) and *Descubrimiento y conquista del Peru* (1963); JOHN HEMMING, *The Conquest of the Incas* (1970); LOUIS BAUDIN, *La Vie de François Pizare* (1930); and CLIFFORD SMYTH, *Francisco Pizarro and the Conquest of Peru* (1931), a narrative.

(M.B.G.)

Placodermi

The Placodermi are an extinct class of fishes characterized by an armour consisting of a head shield and a trunk shield, the two often connected by a paired joint in the neck region. Placoderms existed throughout the Devonian Period (345,000,000–395,000,000 years ago), but only two species persisted into the succeeding Carboniferous Period which began about 345,000,000 years ago. During the Devonian they were a dominant group, occurring in all continents except South America in a variety of marine and freshwater sediments. Most were small or moderate in size, but a few may have reached a length of four metres (13 feet). Their origin is unknown, though it is possible that they shared a common ancestor with the sharks, skates, and rays. It is also possible that they died out leaving no descendants; it has been argued that one order, the Ptyctodontida, gave rise to the modern Holocephali (chimaeras and their allies).

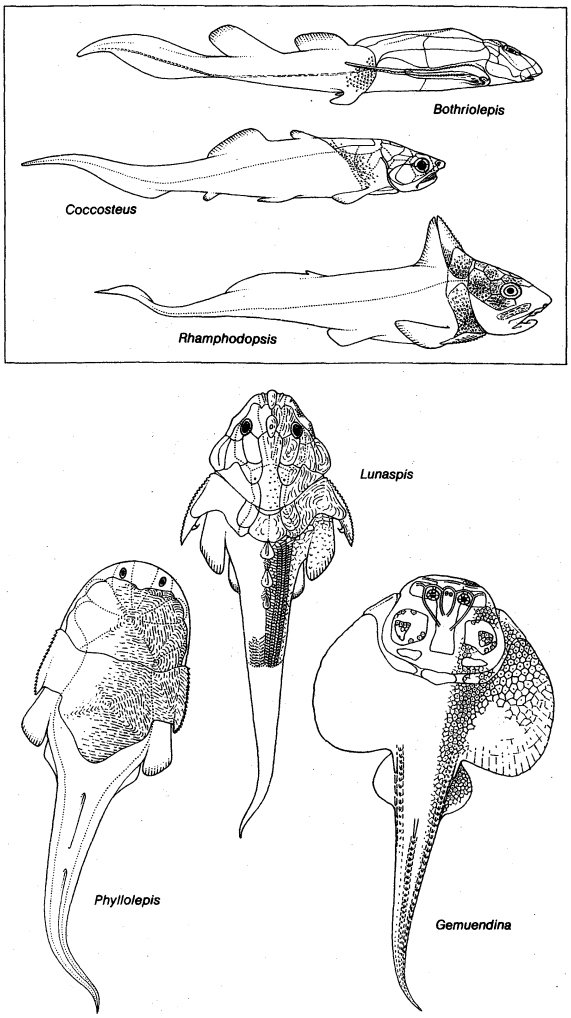
Natural history. Fossils of the earliest known Placodermi, the Arctolepida, are found in freshwater stream deposits. Later members of this suborder were mainly freshwater inhabitants, though a number lived in the sea. The other suborder of Arthrodira, the Brachythoraci, were mostly marine forms, especially in the Upper Devonian. The Petalichthyida and Rhenanida were exclusively marine, the Phyllolepida only freshwater. Although the other two orders occurred in both habitats, the Ptyctodontida lived mostly in the seas and the Antiarchi mostly in freshwaters. One antiarch fossil shows evidence of both lungs and gills, probably an adaptation to life in poorly oxygenated freshwaters.

The earliest placoderms were heavily armoured, bottom-dwelling (benthic) fishes; they were poor swimmers. Many later forms became highly specialized for this type of life, as is well shown in some Arctolepida, which had a lengthened trunk shield and spinal plates, and in the Antiarchi, in which the shield was lengthened and flattened, the eyes and nostrils were in a dorsal position, and the pectoral fins, attached to the anterior end of the trunk shield, just behind the head, were jointed appendages adapted for crawling. The most striking benthic adaptation occurred in Gemuendinida, which superficially resembled modern rays; it is doubtful, however, that they swam by undulating their enlarged pectoral fins, as do the modern rays. The broad, flat-bodied Phyllolepida were adapted for life on stream bottoms. Petalichthyida and Ptyctodontida probably also lived near the bottom. Certain Brachythoraci became adapted for fast swimming between the surface and the bottom (*i.e.*, were nektonic).

Feeding habits

Sluggish, benthic placoderms with small ventrally (*i.e.*, on the lower side) placed mouths (Arctolepida, Petalichthyida) presumably fed on bottom detritus and small invertebrates. One of the Antiarchi, *Bothriolepis*, was a mud-grubber, as is revealed by the preserved contents of its intestine. The Gemuendinida had an anteriorly placed mouth suitable for feeding on small benthic invertebrates. Many Ptyctodontida had heavy, blunt jaw plates adapted for crushing hard-shelled invertebrates, and many Brachythoraci must have been predators—that their wide gape enabled them to swallow smaller fishes has been proved by the stomach content of some fossils. Other Brachythoraci had sharp jaw plates suitable for slicing pieces from larger prey. Some placoderms may have been herbivorous (plant eaters), but this has not been demonstrated.

Form and function. The most striking feature of Placodermi gives them their name; that is, their armour of dermal, or skin, bones, which were formed into a head shield and a trunk shield. The arrangement of bones is so different from that of modern fishes with bony skeletons that it is unlikely that the bones of the two groups have similar origins (*i.e.*, are homologous). For this reason the bones of placoderms have been given distinct names. Although the arrangement of bones differs in each order, it is possible that all placoderms were derived from a common ancestral type. The trunk shield, sometimes



Representative placoderms.
By courtesy of E.A. Stensio

called the dermal shoulder girdle, primitively had a median (*i.e.*, along the body midline) dorsal bone, two median ventral bones, and anterior and posterior pairs of lateral (*i.e.*, toward the sides) and ventral bones, joined by characteristic paired spinal plates. In many placoderms a hinged joint developed between the head and trunk shields, permitting considerable elevation of the head. The presence of an ossified, or bony, endocranium (*i.e.*, the skeleton surrounding the brain and sense organs) in some placoderms has provided detailed information of its structure in some placoderm types. The skeleton of the jaws and gills is not well-known in any placoderm. An enlarged hyomandibular bone supported the jaw articulation of Rhenanida and Ptyctodontida, but this bone is unknown in Arthrodira, in which the jaw hinge is fused to the so-called postsuborbital bone. Structures known as the palatoquadrates and Meckel's cartilages supported bones with teeth, crushing plates, or slicing edges. The gills, situated under the posterior part of the endocranium, were covered laterally by opercula (gill covers) and opened on either side by a slit behind the head shield.

In most placoderms the vertebral column lacked centra (an ossification of the central portion). Neural (dorsal) and hemal (ventral) arches and spines were commonly present on the vertebrae; vertebrae near the head were often fused to form a synarcual element that articulated with the endocranium.

The body behind the trunk shield was probably covered by scales in primitive placoderms; in most known forms these have been reduced or lost, and placoderms with fully scaled posterior regions have been found only in a few Rhenanida, Petalichthyida, and Antiarchi. The tail fin was tapering and whip-like in some, in others was upturned (heterocercal). Arthrodira, Ptyctodontida, and Antiarchi had one or two dorsal fins; Rhenanida and Petalichthyida had only enlarged dorsal scales. The presence of an anal fin in some Brachythoraci is suggested by an internal skeletal plate.

Fin structure

Paired pectoral fins are known in very few Placodermi; their structure can often be inferred from information provided by supporting skeletal elements. Some authorities consider long-based pectoral fins as primitive, but within the Placodermi such fins are found only in otherwise highly specialized Brachythoraci and Gemuendinida. The pectoral fins of the former had as many as 25 basal elements (supportive structures at the base); the latter had very large, raylike pectoral fins, each with an elongated basal plate and three series of radials, or supportive bones in the fins. More primitive Arthrodira (Arctolepida), as well as Ptyctodontida, Stensioellida, and probably Petalichthyida, had narrow-based pectoral fins with only three basal elements in some cases. Within the Arthrodira the long-based pectoral fin evolved from a narrow-based one. The most specialized pectoral fins, which occurred in the Antiarchi, were long, oarlike appendages covered with small plates, and had a complex articulation at the shoulder region and usually at midlength. These fins may have evolved from spinal plates such as occur in most other placoderms.

Paired pelvic fins occur in most Placodermi, but their presence is often indicated only by an ossified pelvic girdle. Rhenanida had relatively small, long-based, flaplike pelvic fins. The antiarch, *Bothriolepis,* had paired "frills" at the posterior end of the trunk shield; these may have been modified pelvic fins. In the Ptyctodontida, the pelvic fins differed in the two sexes; the females had large, scale-bearing pelvic fins, and the males had claspers that probably were used for internal fertilization.

Classification. *Distinguishing taxonomic features.* The major groups of placoderms are distinguished by a number of characters, the most important of which are the pattern of the dermal bones of the head and trunk shields, and the presence or absence of tesserae (minute, superficial plates) between the bones. The proportions of the head and trunk shields are significant, as are the positions of the eye orbits and the nasal openings. The structures of the pectoral fins and of related parts of the trunk shield are also important.

Annotated classification. In 1966 A.S. Romer, an American vertebrate paleontologist, published a classification of Placodermi on which the classification below is based. The dagger (†) is used to indicate the fact that the entire class is extinct and known only from fossils.

†CLASS PLACODERMI

Devonian to Lower Carboniferous (340,000,000–395,000,000 years ago). Head and trunk (or shoulder) region covered by a shield of dermal bones; the 2 shields often connected by a pair of joints in the neck region; a pair of spinal plates typically connect the lateral and ventral parts of the trunk shield; shield formed of bone, usually ornamented by tubercles (knobby prominences) or ridges of dentine (the predominant substance of teeth) or semidentine.

Order Arthrodira (Euarthrodira)

Lower Devonian to Lower Carboniferous; dermal bones of head and trunk shields, including cheek and gill covers, well developed and typically ornamented with tubercles; cranium not elongated posteriorly; 1 pair of paranuchal plates (head bones); lateral and ventral trunk shield with anterior and posterior series of bones; orbits lateral, nasal openings anterior; 2 pairs of upper and 1 pair of lower dermal jaw bones; marine and freshwater.

Suborder Arctolepida

Lower Devonian to Lower Carboniferous; lateral trunk shield not reduced; spinal plates usually large; pectoral fins narrow-based.

Suborder Brachythoraci

Lower to Upper Devonian; lateral part of trunk shield reduced; spinal plates reduced or absent; pectoral fins moderate to long-based.

Order Petalichthyida

Lower to Upper Devonian; roof of cranium greatly elongated posteriorly, with 2 pairs of paranuchal plates; orbits dorsal; cheek region covered with scales; trunk shield short, due largely to reduction in size of posterior plates; caudal fin unknown; no dorsal or anal fins; marine only.

Order Phyllolepida
Upper Devonian; body dorso-ventrally flattened; cranial roof short and broad with a very large median bone nearly surrounded by 5 pairs of small marginal elements; trunk shield broad and moderately long, lacking posterior bones on the sides; bones ornamented with more or less concentrically arranged ridges; freshwater only.

Order Ptyctodontida
Middle Devonian to Lower Carboniferous; cranial roof short, cheek bones reduced or absent, gill covers scaled; 1 pair of upper and 1 pair of lower shearing or crushing tooth plates; trunk shield short, lacking posterior bones; pectoral fins narrow-based, pelvic fins sexually dimorphic, with those of males having claspers; 2 dorsal fins and a long tail that tapers to a point; marine and freshwater.

Order Antiarchi
Lower to Upper Devonian; strongly armoured; adapted to bottom living; shields flattened ventrally; head short with distinctive dermal bone pattern; mouth ventral with small transversely placed jaws; orbits close together on top of head and nasal openings slightly in front of them; trunk shield long with 2 median dorsal bones; pectoral fins long, oarlike appendages articulated in shoulder region and usually jointed at midlength; body rarely covered with scales; 1 to 2 dorsal fins; tail with upper lobe larger than lower; almost worldwide in freshwater and sometimes in marine deposits.

Order Rhenanida
Lower to Upper Devonian; larger skull plates typically separated by fields of tesserae; trunk shield short; marine only.

Suborder Radotinida
Lower Devonian; a well-developed rostrum (blunt beak) anterior to the nasal openings, which are near the midline dorsally; eye orbits lateral.

Suborder Gemuendinida
Lower to Upper Devonian; flat-bodied, raylike fishes with much enlarged pectoral fins; orbits dorsal and nasal openings between them near midline; mouth terminal.

Suborder Stensioellida
Lower Devonian; poorly known forms with resemblances to Rhenanida; nasal openings probably ventral.

Critical appraisal. There is considerable difference of opinion among authorities about the interrelationships of the various orders and suborders. This may be attributed in part to incomplete knowledge of their structure, and in part to different emphases given to characters in the various classifications. The classification above is conservative in that it gives equal ordinal rank to each major group, and it implies no special close relationship among them. In another widely used classification, the Rhenanida and Antiarchi are considered to be so divergent that they are regarded as superorders; the remaining orders are grouped together as the superorder Arthrodira, and the order Euarthrodira is used to unite the Arctolepida and Brachythoraci. A different classification, based primarily on the structure and presumed evolutionary development of the pectoral fins, shoulder girdle, and spinal plates, is founded on the "fin-fold" theory of the origin of paired fins—according to which a long-based pectoral fin is primitive, and a narrow-based fin is specialized. This classification has not been widely accepted because it is completely opposed by the manner in which placoderm fossils are distributed in the geological strata, or layers.

Future studies may demonstrate special relationships among some of the following orders: Arthrodira, Petalichthyida, Ptyctodontida, and Rhenanida. The Phyllolepida and Stensioellida are too poorly known to permit satisfactory conclusions about their position in the classification.

BIBLIOGRAPHY. R.S. MILES, "Features of Placoderm Diversification and the Evolution of the Arthrodire Feeding Mechanism," *Trans. R. Soc. Edinb.*, 68:123–170 (1969), a valuable discussion of placoderm evolution and adaptation; D.V. OBRUCHEV, "Class Placodermi," in Y.A. ORLOV (ed.), *Fundamentals of Palaeontology*, 11:168–259 (1967), a largely systematic treatment, translated from the original (1964) Russian edition for the Smithsonian Institution; A.S. ROMER, *Vertebrate Paleontology*, 3rd ed. (1966), an excellent, college-level textbook; ERIK STENSIO, "Placodermata, Arthrodires," in *Traité de paléontologie*, pt. 4, vol. 2, pp. 71–692 (1969), a profusely illustrated, detailed description of the structure of Placodermi (in French).

(R.H.D.)

Planck, Max

The beginning of the 20th century saw the development of two theories that brought about fundamental changes in physics and have gained significance far beyond this discipline: the quantum theory and the theory of relativity. The first of these, the quantum theory, was established at the turn of the century, on December 14, 1900, when a German physicist, Max Karl Ernst Ludwig Planck, at a meeting of the German Physical Society in Berlin, announced his sensational theoretical research into radiation and absorption by a black body, defined as a body that absorbs all the radiation that falls on it. In order to explain the distribution of radiant energy emitted from a black body at a given temperature, it was necessary for him to postulate that energy can exist only as quanta in multiples of an elementary quantity, now designated as Planck's constant (h). From that time on, a predominant role has been played in an ever-increasing number of fields by the new fundamental constant h, the quantum of action, introduced into physics by Planck, and it has proved to be of vital importance to the theory of structure of atomic dimensions. Thus Planck's discovery was to influence physics and the related sciences to a much higher degree than he himself could ever imagine.

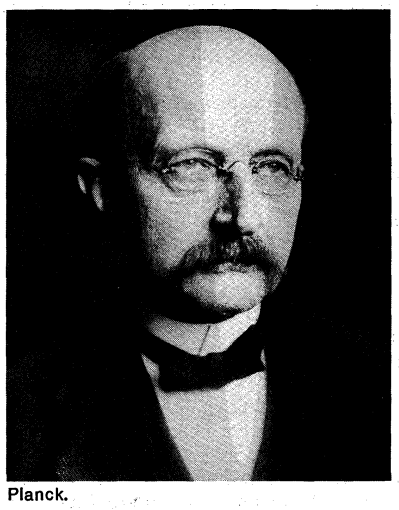
An especially interesting point about the way in which Planck's new concept came into being is that Max Planck was the furthest from having a revolutionary mind. He was one of the conservative scientists of the classical epoch, and in thermodynamics he did not even recognize the probability interpretation, advocated at that time chiefly by an Austrian physicist, Ludwig Eduard Boltzmann. He felt compelled to use this particular statistical method only when he saw no other possibility of providing a theoretical basis for the correct radiation formula that he intuitively conceived.

Early life

Max Planck was born on April 23, 1858, in Kiel, Schleswig, West Germany, but he was descended from a South German family. He spent his school days and his first years of studies at the University of Munich. There he came into close contact with the many aspects of artistic life and developed an inclination toward music that was so strong that it made him hesitate about whether or not to devote himself completely to it.

Among Planck's teachers in his later years of study in Berlin were the famous physicists Hermann Ludwig Ferdinand von Helmholtz and Gustav Robert Kirchhoff. He made extraordinarily rapid progress in his studies and, at the age of 21, now back in Munich once more, he was awarded his doctor's degree in physics *summa cum laude*. The following year he became a lecturer at the University of Munich. His subject, theoretical physics, was at that time not well represented, and he first had to struggle for its authorization.

EB Inc.



Planck.

In 1885 Planck was called as extraordinary professor to the University of Kiel, where he married his first wife, Marie Merck, the daughter of a Munich banker. After the death of his wife, he married her niece, Marga von Hoesslin, the daughter of the Munich painter Georg von Hoesslin, in 1911, and this union lasted until his death.

When Kirchhoff died, Max Planck was called in 1889 as his successor to the University of Berlin, where—from 1892 onward as an ordinary professor—he spent 37 years at the Friedrich Wilhelm University delivering his brilliant lectures, and where he carried out his most important scientific activities.

In Berlin, Planck was able to devote himself unreservedly to many unsolved problems of physics. His preference was for thermodynamics, and here again he attempted to explain how energy is distributed according to wavelength in the so-called black-body radiation of a cavity. At that time there was a formula introduced by the German physicist Wilhelm Wien, for the distribution of energy, which corresponded with experiments in the short wavelength region. Berlin scientists had carried out new measurements of the black-body radiation with longer wavelengths and had found agreement with a theory that Lord Rayleigh, an English physicist, had prepared in 1900, which is still known as Rayleigh's radiation formula.

Planck succeeded first of all in obtaining, by skilful combination of the formulas of both Wien and Rayleigh, a new relationship in full agreement with the measurements taken over all wavelengths of the electromagnetic spectrum. This formula, which has since become famous as Planck's radiation formula, was announced by Planck on October 19, 1900, at a meeting of the German Physical Society, although he was not then able to give a theoretical basis for the formula. Then, in the short space of two months before the end of 1900, he succeeded in making a complete theoretical deduction, at the cost, however, of renouncing fundamental principles of classical physics and of introducing quanta of energy.

Planck's formula

Because of this break with classical physics, the quantum theory was at first rejected by many physicists. Most of them considered it to be an ad hoc hypothesis invented only to prove the correct radiation formula. It was only with the triumph of Niels Bohr, a Danish physicist, who in 1913 calculated for the first time the position of the special lines of a spectrum with the aid of the quantum theory, that the theory became a spectacular success.

Planck took little part in the further rapid development of the quantum theory that reached a pinnacle in the '20s by the efforts of the physicists Werner Karl Heisenberg, Erwin Schrödinger, Paul Adrien Maurice Dirac, and others in formulating quantum mechanics. Planck, on the contrary, frequently spoke out against the new ideas to which he, of course, had personally given the incentive, but which had now become much more revolutionary. Thus, Planck, to the end of his life, belonged to the group of older physicists who recognized the successes of the younger generation but who refused to see anything conclusive in their theoretical constructions.

In the meantime, Planck had taken on countless new fields of activity and honorary offices. In 1912 he became Permanent Secretary of the Section of Mathematics and Natural Science of the Prussian Academy of Science, of which he had been a member since 1894. He was also foreign member of the Royal Society of London and of other societies. In recognition of his services to the advancement of physics by the discovery of energy quanta, he was awarded the Nobel Prize for Physics in 1918. From 1930 he was president of the Kaiser Wilhelm Society of Berlin, which has borne his name since the end of World War II as the Max Planck Society.

Awarded Nobel Prize

But Planck also showed wide interests in fields other than physics. He occupied himself particularly with philosophical and religious problems, to which he devoted several publications.

During the National Socialist regime in Germany, Planck considered it his duty to stay in his native country, and he endeavoured repeatedly to prevent injustice to others. In 1944 his Berlin home was destroyed by bombing; and in the same year, July 20, his son Erwin was arrested for his implication in the plot to assassinate Hitler and was executed six months later. After World War II, Max Planck spent the last years of his life in Göttingen. He died there on October 4, 1947, at the age of 89.

BIBLIOGRAPHY. MAX PLANCK, *Scientific Autobiography and Other Papers* (Eng. trans. 1949); M.J. KLEIN, "Planck, Entropy and Quanta, 1901–1906," *The Natural Philosopher*, vol. 1 (1963); H. HARTMANN, *Max Planck als Mensch und Denker*, 3rd rev. ed. (1953), biographies with emphasis on Planck's scientific work; see also N.H. DE V. HEATHCOTE, *Nobel Prize Winners in Physics, 1901–1950* (1954); and MAX PLANCK, *Akademie-Ansprachen* (1948), for a list of all of Planck's printed papers of 1879–1948.

(W.Br.)

Planet, Minor

The minor planets, sometimes called asteroids, form a distinct and numerous class of very small objects in the solar system. They revolve around the Sun in orbits that nearly all lie between Mars and Jupiter. The existence of these tiny bodies, most of which are only a few kilometres across, challenges every theory of the origin of the solar system. Their discovery illustrates the scientific progress of the early 19th century. J.D. Titius in 1766 proposed and J.E. Bode in 1772 popularized what is known in astronomy as the Titius-Bode Law, for which the physical basis is still not understood. One takes the sequence of numbers (0, 3, 6, 12, 24, 48, 96, 192, 384, 768, etc.), adds 4 to each, and divides each by 10; the result represents the average distance of each of the then-known major planets from the Sun relative to the Earth's distance (one astronomical unit, abbreviated a.u.) quite well; namely, Mercury 0.4, Venus 0.7, Earth 1.0, Mars 1.6, (missing planet 2.8), Jupiter 5.2, Saturn 10.0, and a supposed next planet 19.6. Sir William Herschel's discovery of Uranus in 1781 strongly supported belief in this "law," because the planet's actual distance agreed quite well at 19.2 a.u. with the predicted value. The later discoveries of Neptune in 1846 at an actual distance of 30.1 a.u. and of Pluto in 1930 at 39.5 a.u. did not fit the so-called law, and one also notices that 0 is not one-half of 3.

Titius-Bode law

Nonetheless, in the last years of the 18th century in the hope of finding the missing planet at 2.8 a.u., Baron von Zach organized a group of 24 astronomers to search the sky. This task required that hand-drawn sky charts be prepared for all stars down to an agreed limit of faintness. Then on subsequent nights a region of the actual sky must be observed and compared in every detail with its chart, in the hope that an object would appear to view that had not been there before. The expected motion of such a planet would be about 30 seconds of arc per hour, so that an actual discovery (as contrasted with a missing star on the chart) could be readily established.

History of discoveries. Ironically, the first discovery was not made by a member of the organized group but by Giuseppe Piazzi at Palermo, Sicily, on the night of Jan. 1, 1801. He continued his observations until February 11, when illness interrupted. No one else was aware of his discovery, and so the new planet was "lost" as it receded behind the Sun. C.F. Gauss (1777–1855) brilliantly attacked the problem of computing an orbit from a few observations and laid down principles for tracking a newly discovered object, which have not been improved upon to this day. Using Gauss's predictions, Zach and Heinrich W.M. Olbers rediscovered the little planet, which had been named Ceres by Piazzi after the goddess of fertility, who was also the patron goddess of Sicily. The size of the orbit agreed exactly with Bode's Law. The object was disappointingly small, and no observable diameter could be discerned. Hence the term asteroid (meaning starlike) came into use, but it is a misnomer since these objects are not stars, and they do not remain in constellations but wander around the sky as the planets do.

About three months later Olbers discovered another small planet, which he named Pallas; in 1804 another, named Juno, was discovered by Karl Harding at Lilienthal, Germany; and finally in 1807 Olbers found a fourth member of this group, which he called Vesta. All the first four had similar orbits, and this immediately led to the theory that

the planet that was "required" by Bode's Law at this distance from the Sun either had been broken into fragments or had never completely accreted into one body, according to the nebular hypothesis of Laplace (1749–1827). As a group, these four new orbits had larger eccentricities and inclinations than the major planets, the averages being 0.165 and 16.4° as compared with 0.071 and 3.2°. It was at first thought that the four fragments completed the group, and the search seems to have been abandoned until 1830 when K.L. Hencke renewed it. In 1845 he discovered a fifth minor planet and named it Astraea. After that the annual rate of discovery increased until it reached about 10 by 1890, by which time 300 were known.

Role of photography. The introduction of astronomical photography in 1891 provided a turning point, and the invention of the "blink" comparator proved to be of enormous value in the search for more. Two photographic plates are taken of exactly the same region of the sky, with the exposures separated in time by a few hours or a day. The plates are mounted in the comparator so that an optical path from the same part of either plate may be reflected into the observing eyepiece by flipping a small mirror back and forth. When properly aligned, the background stars appear to be identical on both plates, but an object that has moved will appear in nearby but different positions, and it will "blink" at the observer as he intermittently observes one plate or the other. Afterward the rectangular positions on the plates must be measured accurately with respect to those surrounding stars that can be found in star catalogs. The process of computing the position of the minor planet on the sky from these measures is called the plate-reduction.

"Blink" comparator

The rate of discovery thus increased greatly, although many of the objects were observed over too short a period of time for their orbits to be well determined. The German astronomers participated extensively in this work. After Gauss's work in the computation of orbits, that of J.F. Encke, P.A. Hansen, J. Bauschinger, and others followed. By 1930 the work was concentrated under G. Stracke at the Astronomisches Rechen-Institut in Berlin-Dahlem, where all observations were collected, orbital elements were computed, and predictions were prepared for the observers. M. Wolf, at the Königstuhl Observatory at Heidelberg, exploited the new photographic method, using the 41-centimetre Bruce camera built especially for this work and donated by Catherine Bruce of California. He made 582 new discoveries during his lifetime, and his successor, K. Reinmuth, made 980. Other leading discoverers included C. Jackson at Johannesburg (392), G. Newjmin in the Crimea (379), and E. Delporte at Brussels (214). By 1940 about 7,500 discoveries had been made; 1,564 planets had received permanent numbers; and the predictions of their favourable observing positions were published each year. It is estimated that there are about 50,000 minor planets in the solar system that could be observed with existing telescopes.

Identification and rediscovery. There is no simple way to identify any particular minor planet except by the agreement of its position and motion with that predicted for it from its orbital elements, and by the brightness that is predicted on the basis of previous observations and relative distances. Thus the required amount of computations continually increased. To have accurate predictions after many years, it is necessary to compute the gravitational attraction of each of the major planets upon each minor planet and the perturbing effects of this upon each set of orbital elements. There have been serious proposals that astronomers should abandon studying the minor planets completely. Some astronomers have referred to them as "the vermin of the sky." But the discovery of planet (433) Eros in 1898 showed the importance of further studies and the work has gone on.

In 1900–01 and again in 1930–31, Eros approached within 23,000,000 kilometres of the Earth and thus provided an opportunity to determine a more accurate value of the solar parallax and the astronomical unit, the distance from the Earth to the Sun. The possibility of finding other objects of scientific interest was the basis of continued observations.

Notable groups and minor planets. The discovery of planet (588) Achilles in 1908 represented the first of at least 30 known Trojan planets. They remain in positions that form nearly an equilateral triangle with respect to the Sun and the largest major planet, Jupiter. They are examples of the stable triangular solution of the three-body problem (see MECHANICS, CELESTIAL). The discovery of (1221) Amor in 1932 marked the first of a series of "close-approach" planets, which may pass even nearer to the Earth than Eros and nearer to the Sun than the Earth. Then followed Adonis in 1936, Hermes in 1937, (1566) Icarus in 1949, (1620) Geographos in 1951, and others. On June 12, 1968, Icarus approached within 6,800,000 kilometres of the Earth, and it provided the first opportunity for observations by radar. Recently J. Schubart has determined the solar parallax from the observations of (1221) Amor, as it has been observed during five close passages near the Earth since its discovery. Icarus has the most eccentric elliptic orbit ($e = 0.83$) of all the known minor planets, the smallest semimajor axis ($a = 1.08$), and it passes closest to the Sun (28,000,000 kilometres). Hidalgo (944) has the largest semimajor axis (5.82), the highest inclination (42.5°), and it ranges in distance from 2.0 to 9.5 astronomical units from the Sun. Both were discovered by W. Baade.

First radar observation

Results from modern survey. World War II produced a hiatus of nearly ten years in minor planet research. In 1947 the Minor Planet Center was established at the Cincinnati Observatory in Ohio; the work had previously been coordinated at Berlin and later at Heidelberg in Germany. An intensive observing program was instituted at the Goethe Link Observatory of Indiana University, Brooklyn, Indiana, for the purpose of reobserving those minor planets that had deviated widely from their computed positions over the intervening years. Nearly 5,000 sky photographs were obtained between 1950 and 1965, and more than 19,000 positions of minor planets were observed. Simultaneously there were rapid developments in electronic computers, and these were fully exploited at Cincinnati and elsewhere to carry out the necessary computations of orbital elements, perturbations, and predictions. An annual volume containing the ephemeris of each numbered planet near its opposition from the Sun is published by the Institute of Theoretical Astronomy in Leningrad.

G.P. Kuiper and his collaborators conducted a survey of the sky with the 25-centimetre telescope at the McDonald Observatory during 1950–52 to photograph systematically all minor planets to the limit of 16.5 magnitude. They obtained 3,255 observations, only 25 percent belonging to new objects. In 1960 the 122-centimetre Schmidt telescope of Palomar Observatory (now part of Hale Observatories) was used to photograph a statistical sample of the sky, 18° by 12°, reaching the 20.5 magnitude. These 96 large plates were blinked and measured by C.J. Van Houten and I. Van Houten-Groeneveld, yielding 12,000 observations of 2,100 objects. The results of this statistical sample have been extrapolated to provide an estimate for the whole sky.

The naming of minor planets. The prerogative to choose the name of a new minor planet has been traditionally assigned to the discoverer. Consecutive numbers are now assigned to the planets whose orbits are accurately known, and the names of modern discoveries are submitted to the International Astronomical Union for approval.

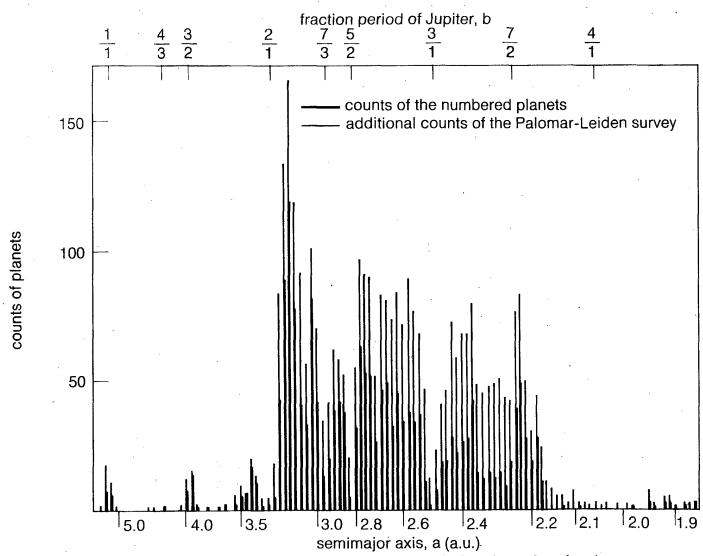
The feminine mythological names adopted for the early discoveries have been supplemented by other classes of names, usually feminine, occasionally with a contrived feminine ending. The names usually honour some individual or place.

Minor planet (12) was named Victoria in 1850 after the queen of England; (250) Bettina after Baroness Rothschild who subsidized the discovery; (323) Brucia was the first minor planet discovered with the telescope donated by Miss Bruce; (1486) Marilyn was named after Paul Herget's daughter. There is a long list of curious hybrids, such as the names of famous astronomers with attached feminine suffixes, (784) Pickeringia, (1000) Piazzia, (1001)

Gaussia. There are acronyms; for example, initials from Reinmuth's eight discoveries spell G. Stracke, honouring a fellow astronomer; (849) Ara, discovered in the U.S.S.R., commemorates the American Relief Administration's help during the famine of 1922–23. Many planets are named after flowers.

Exceptions to the rule of femininity include certain groups or single planets with unusual orbital characteristics such as the Trojan and Apollo groups, and (1620) Geographos was named in recognition of the National Geographic Society subsidy of the Palomar Observatory's photographic sky survey. Fittingly, the planet known to come closest to the Sun is (1566) Icarus.

Statistics of minor planets. The inclinations of minor planets are clustered around a maximum near 5°, but they are fairly well distributed from 0° to 16°, with about 10 percent still above this range. The eccentricities have a maximum near 0.17, and few exceed 0.30. The dominant feature of the distribution of the mean distance from the Sun is the existence of gaps, discovered by Daniel Kirkwood in 1866 when only 88 planets were known. The most prominent one is at the point where the period of revolution is in the exact ratio of 2:1 with the period of Jupiter; others are at 3:1, 5:2, and other simple integral ratios. This phenomenon is believed to be due to the perturbations produced by Jupiter when the minor planet is in resonance with it, although no rigorous mathematical or statistical proof has ever been given. Curiously, at the ratios 1:1 and 3:2 there are concentrations of minor planets. The formula $\log N = 0.39m - 3.3$ closely represents the number of minor planets, within an interval of one magnitude centred on m, that have this mean opposition magnitude. To eliminate the effects of varying distances, one computes the absolute magnitude, g, that the planet would have if it were one unit from the Sun and the Earth. This brightness then depends upon the true diameter and the albedo or reflecting efficiency of the surface. The latter is found to be about 7 percent, the same as the Moon, although for Vesta it is four times that amount. The diameters can be observed for only a few of the largest objects: Ceres is 785 km, Pallas is 490 km, Juno is 190 km, and Vesta is 400 km. The diameter in kilometres for fainter objects may be computed from the formula $\log d\ (\text{km}) = 3.7 - 0.2\ g$. If $g = 20$, the object is only 0.5 km in diameter. The total mass of all the minor planets is probably less than one-thousandth of the mass of the Earth, but the density is not accurately known. The number of planets having a given diameter shows a discontinuity around $g = 11$ or a diameter of 30 km.



Distribution of minor planet distances from Sun. Semimajor axis, a, is average distance of planet from Sun; b stands for the ratio of the minor planet orbit period to the Jupiter period (see text).

Precise photoelectric observations of the brightness of some minor planets show variations that may be due to rotation, irregular shape, and/or differing albedo over the surface. In 1931 Eros was observed to be very elongated, with a maximum diameter of 30 km. The Trojan planet (624) Hector is believed to be nearly cylindrical, with a length of 110 km, a diameter of 40 km, and a rotation period of 6.9 hours. Such series of observations are difficult to obtain, but the known cases indicate that the axes of rotation are distributed at random in space. The known periods of rotation lie mostly between five and eight hours. Eugene Rabe has shown that this is consistent with the periods of revolution of infinitesimal particles about each other and with their gradual accretion.

Rabe has also shown that if the planet Jupiter was formerly about 20 times as massive as it now is, then its former satellites could have escaped the gravitational field of Jupiter and wandered into the present positions of the Trojans. The known Trojans range in size from 100 km to 3 km. Analysis of the three-body problem indicates that they can remain in these stable periodic orbits with slow swings back and forth along it (librational amplitudes) of nearly 60° from the triangular points for an indefinitely long time.

Origins of minor planets. In order to study the past history of the minor planets, Hirayama Seiji devised in 1918 an approximate method for removing the long-term oscillations of the elements that are induced by the major planets. The results of his calculations allowed him to identify certain families of minor planets, each of which appears to have come from some single point in space at a former time, presumably as the product of the fracturing of a larger parent body. Such an analysis can be carried backward with validity for only a few tens of millions of years, or less than 1 percent of the age of the solar system. The same method has been extended by D. Brouwer (1951), James Arnold (1969), and the Van Houtens (1970). There is not complete agreement on the details, but in general the family members have smaller diameters than the average of all the minor planets taken together, and they constitute about 30–40 percent of the total population. Once a larger body is broken into fragments, the probability of future collisions is increased because of the greater total cross-sectional area that the particles present. Thus the process would continue indefinitely, and in part this could lead to the formation of the meteors. Edward Anders (1965) has undertaken to "reconstitute" the parent bodies from the present identifiable families. He conjectures that if this process could be carried out a second time at a much earlier and now unrecognizable state of the solar system, then only a few relatively large minor planets would have been present.

Hirayama families

In spite of the evidence in support of such a theory of the past history of minor planets, G.P. Kuiper, H. Alfvén, the late E.W. Brown, and others have been inclined to a theory that the process of accretion is continuously at work and that there is a dynamical "focussing" mechanism that produces such similarity of elements, in combination with the Kirkwood gaps. The average density of material per unit of volume in the region of the minor planets is about 10^6 lower than in the region of the interior major planets. Thus the process of accretion is far from being completed. Considering the discontinuity that has been found in the count according to diameters, one is led to conclude that the accretion process is still in operation for the larger objects and that the collisional process is responsible for the smaller ones. But the effects of magnetohydrodynamic forces and the solar wind may preclude the possibility of any precise description of the past conditions of the minor planets on the basis of universal gravitation alone.

BIBLIOGRAPHY. H. ALFVEN, "Asteroidal Jet Streams," *Astrophys. Space Sci.*, 4:84–102 (1969), a study and tabulation of asteroidal data with the view to supporting the hypothesis that the process of accretion is still in effect, as opposed to the hypothesis of collision and fragmentation; E. ANDERS, "Fragmentation History of Asteroids," *Icarus*, 4:398–408 (1965), assumes the collision hypothesis and studies the relative masses involved and the amount of fragmentation that has occurred; J.R. ARNOLD, "Asteroid Families and Jet Streams," *Astr. J.* 74:1235–1242 (1969), uses sophisticated computer techniques to analyze the Hirayama families and examines

the data as they support each different hypothesis; D. BROUWER, "Secular Variations of the Orbital Elements of Minor Planets," *Astr. J.* 56:9–32 (1951), a revision and extension of the study of Hirayama's families; J.L. DUNLAP and T. GEHRELS, "Lightcurves of a Trojan Asteroid," *Astr. J.* 74:796–803 (1969); G.P. KUIPER, "Note on the Origin of the Asteroids," *Proc. Natn. Acad. Sci. U.S.A.* 39:1159–1161 (1953), a revision of the author's previous theory, and with others "Survey of Asteroids," *Astrophy. J. Suppl. Ser.* 3:289–427 (1958), the results of a comprehensive observational survey of all the minor planets, with an analysis; E.K. RABE, "On the Origin of the Kirkwood Gaps and the Minor Planet Families," Z. *Astrophys.* 40:107-119 (1956), analyzes the effects on the minor planets of a mass loss of proto-Jupiter; v. HOUTEN *et al.*, "The Palomar-Leiden Survey of Faint Minor Planets," *Astronomy and Astrophysics* (1970), the results of an observational sampling of faint minor planets, with an analysis; F.G. WATSON, *Between the Planets*, rev. ed. (1956), an engaging, popular book.

(P.H.)

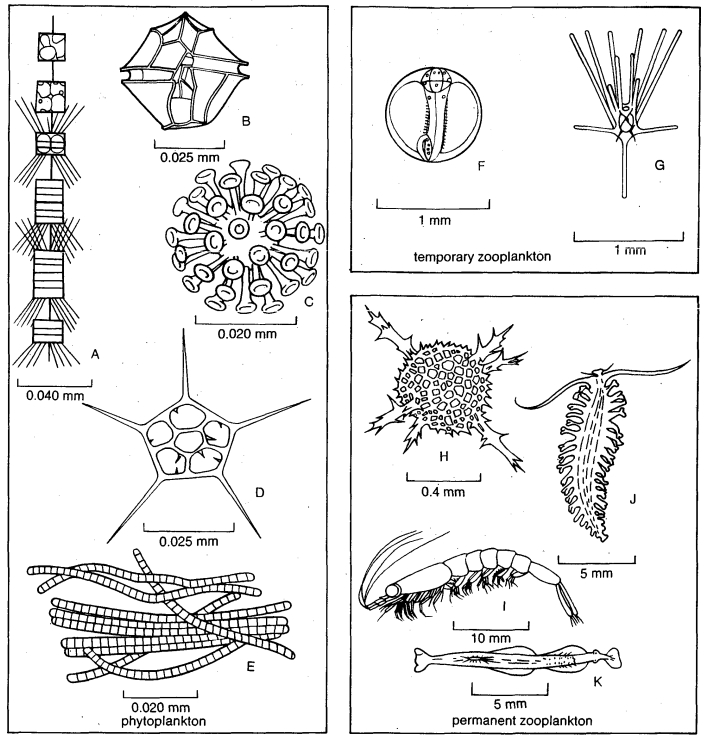
Plankton

Plankton is a collective term for organisms that live suspended in the sea and in bodies of fresh water around the world. These organisms drift about either because they are nonmotile or because they are too small or too weak to swim effectively against the currents. Plankton is rich in algae, bacteria, and protozoans but includes also crustaceans, mollusks, and coelenterates, as well as representatives from almost every other phylum of animals. Plankton is distinguished from the nekton, which is composed of strong-swimming animals, and the benthos, which includes sessile, creeping, and burrowing organisms on the sea floor. Large floating seaweeds (for example, *Sargassum*, which constitutes the Sargasso Sea) and various related multicellular algae do not belong to the plankton but to the pleuston. Organisms resting or swimming on the surface film of the water are called neuston (*e.g.*, the alga *Ochromonas*).

Plankton is the productive base of both marine and fresh water ecosystems, providing food for larger animals and indirectly for man, whose fisheries depend upon plankton. The vast resource of plankton is only beginning to be developed and exploited by man. It may in time be the chief food supply of the world, in view of its high biological productivity and wide extent. It has been demonstrated on several occasions that large-scale cultures of algae are technically feasible. The unicellular green alga *Chorella* has been used particularly in this connection. Through ample culture conditions production is directed toward protein content greater than 50 percent. Although this protein has a suitable balance of essential amino acids, its low degree of digestibility prevents practical utilization. Phytoplankton may become increasingly important in the space age as a source for food and for gas exchange. The carbon dioxide released during respiration of the crew would be transformed into organic substances by the algae, while the oxygen liberated during this process would support human respiration.

The kinds of plankton. The plantlike community of plankton is called phytoplankton, and the animallike community is zooplankton. This convenient division is not without fault, for strictly speaking, many planktonic organisms are neither clearly plant nor animal but are better described as protists. Using size as a criterion, plankton can be subdivided into macroplankton, microplankton, and nannoplankton, though no sharp lines can be drawn between them. Macroplankton can be collected with a coarse net, and morphological details of individual organisms are easily discernible. These forms, one millimetre or more in length, ordinarily do not include phytoplankton. The microplankton (also called net plankton) is composed of organisms between 0.05 and 1 mm in size and is a mixture of phytoplankton and zooplankton. The lower limit of its size range is fixed by the aperture of the finest cloth used for plankton nets. The nannoplankton (dwarf plankton) passes through all nets and consists of forms of a size less than 0.05 mm. Phytoplanktonic organisms dominate the nannoplankton.

Phytoplankton. The chief components of marine phytoplankton are found within the algal groups and include diatoms, dinoflagellates, and coccolithophorids. Silicoflagellates, cryptomonads, and green algae are found in most plankton samples. Fresh-water phytoplankton, usually rich in green algae, also includes diatoms, blue-green algae, and true flagellates (see ALGAE).



Representative Types of marine plankton.
(A) Diatom *(Thalassiosira)*. (B) Dinoflagellate *(Gonyaulax)*. (C) Coccolithophorid *(Discosphaera)*. (D) Silicoflagellate *(Distephanus)*. (E) Blue-green alga *(Trichodesmium)*. (F) Flounder eggs with embryo. (G) Pluteus larva of sea urchin. (H) Radiolarian. (I) Crustacean *(Euphausia)*. (J) Annelid *(Tomopteris)*. (K) Arrowworm *(Sagitta)*.

Important marine phytoplankton

The diatoms range in size from two μ ($1\ \mu = 1/1{,}000$ mm)—*e.g.*, small *Nitzschia* and *Fragilaria* species—to large representatives 1 or 2 mm long—*e.g.*, *Thalassiothrix longissima, Rhizosolenia styliformis.* Some live as individuals (*e.g., Corethron, Coscinodiscus*), but the majority form chains (*e.g., Chaetoceros, Skeletonema, Thalassiosira, Asterionella*). Some of the diatoms (*e.g., Corethron, Chaetoceros*) have their shells drawn into long hairs that greatly improve floating. Most diatoms are nonmotile; but many, particularly benthic or bottom-dwelling forms, are capable of gliding motions. The diatom shell, or frustule, is composed of hydrated amorphous silica. Such shells have formed great fossil deposits known as diatomaceous earth.

The dinoflagellates range in size from a few μ (*e.g.*, small *Gyrodinium*) to 0.2 mm (*e.g., Ceratium*). Most of the species are photosynthetic, but many nonphotosynthetic forms are known (*e.g., Peridinium decipiens, Noctiluca*). Dinoflagellates are motile organisms with two flagella. One drives the organism forward, while the other is responsible for a rotating motion. A dinoflagellate directs its swimming according to both light intensity and direction (phototaxis). Under natural conditions in quiet bodies of water, it may perform diurnal vertical migrations that serve to adjust its position toward light and a greater rate of photosynthesis. Swimming also helps the organism stay above the light (or photosynthetic) zone. Improved flotation has been achieved in some species through the development of spectacular arrangements of "wings." Such features are particularly noticeable in species found in tropical waters. In temperate regions dinoflagellates are typical members of summer and autumn phytoplankton. They produce most of the bioluminescence seen during that time. One of the best known bioluminescent organisms is *Noctiluca*, a nonphotosynthetic dinoflagellate that migrates to the surface at night. Other bioluminescent dinoflagellates are *Ceratium* and *Gonyaulax*. Off the west

coast of California, *Gonyaulax polyedra* causes the "red tide" and the pronounced bioluminescence accompanying it. The flash seen at night around footprints close to the water during "red tides" is caused by the same organism. *Gonyaulax* bioluminescence, more intense during the night than in daylight, has a diurnal rhythm with a period of about 24 hours. Dinoflagellate bioluminescence is most pronounced in warmer regions, where it has a spectacular intensity in some areas; for example, in the offshore waters of Puerto Rico (see BIOLUMINESCENCE). Several species of dinoflagellates are poisonous. Off the west coast of Florida and in the Gulf of Mexico, red water, caused by *Gymnodinium brevis*, has been associated on occasion with enormous numbers of dead fish. In many other areas extensive losses of fish have occurred in connection with red tide. Dinoflagellates of several genera (*Gymnodinium, Exuviaella, Gonyaulax, Prorocentrum*) are responsible. Shellfish poisoning off the west coast of California is caused by the dinoflagellate *Gonyaulax catenella* and by another *Gonyaulax* species in the Bay of Fundy (Nova Scotia). Poisons isolated from these dinoflagellates are among the most powerful known in nature. Typically, dinoflagellates exist in marine environments, although vast numbers of *Glenodinium sanguineum* may cause red discoloration of inland lakes, especially in alpine regions.

Coccolithophorids vary from 5 to 50 μ in size and thus are part of the nannoplankton. These algae are covered by variable numbers of calcified plates (coccoliths) of different shapes. Most of them are ellipsoidal; but coccoliths in the shape of rhombuses, trumpets, or spines are not uncommon. The coccolithophorids are classified according to the shape and the elaborate microstructure of their coccoliths. In many cases the electron microscope must be used for identification. Coccoliths are found in enormous numbers in marine bottom sediments. The coccolithophorids are motile, with two flagella, and typically are found in tropical and temperate regions of the open sea. One species (*Coccolithus huxleyi*), however, is common in coastal areas, particularly fjords, where it occurs sporadically in vast numbers (as high as hundreds of millions per litre), giving the water a milky greenish appearance (green water). A few coccolithophorids also occur in fresh water.

Important fresh-water phytoplankton

Green algae, common in fresh-water plankton, are found more rarely in marine environments. Some are motile (*e.g., Chlamydomonas, Haematococcus*), and many of the motile forms occur in colonies (*e.g., Gonium*). Several nonswimming green algae have developed structures that enable them to float on the surface of the water; these include branching and embedding of organisms in a gelatinous mass of considerable size (*e.g., Staurastrum*). Blue-green algae are present in most samples of fresh-water phytoplankton. At times tremendous populations of *Aphanizomenon flos-aquae* float on the surfaces of lakes, completely covering the water (water bloom). Although blue-green algae are usually absent from marine waters in temperate and polar areas, one genus, *Trichodesmium*, frequently appears in tropical regions. The reddish colour of *Trichodesmium erythraeum* is caused by a photosynthetically active pigment, phycoerythrin. This blue-green alga floats on the surface of the water, aided by a mechanism involving its gas vacuoles, and sometimes causes red discoloration of the water. The Red Sea owes its name to this alga.

Other microflagellate representatives of the phytoplankton undoubtedly play a significant role in the primary production of oceans and inland waters. True flagellates (*e.g., Euglena*) are common in fresh-water plankton. Vast, nearly pure populations of *Euglena* sometimes occur in highly polluted areas.

Temporary and permanent zooplankton

Zooplankton. The zooplankton is divided into two groups. Temporary plankton consists of planktonic eggs and larvae of members of the benthos and nekton; permanent plankton includes all animals that live their complete life cycles in a floating state. The temporary plankton, particularly abundant in coastal areas, is characteristically seasonal in occurrence, though variations in spawning time of different species ensure its presence in all seasons. Representatives from nearly every phylum of the animal kingdom are found in the permanent plankton. Among the protozoans, planktonic foraminiferans and radiolarians are so abundant and widespread that their skeletons form the bulk of bottom sediments over wide ocean areas. They are absent in fresh water. The ciliate protozoans are represented mainly by the tintinnids, which are between 20 and 640 μ in size and sometimes occur in vast numbers. Among the planktonic coelenterates are the beautiful siphonophores (*e.g., Physalia*, the "Portuguese man-of-war"), and the jellyfishes. Planktonic ctenophores, called "combjellies" or "sea walnuts," are also common. Fresh-water rotifers may be present in plankton in vast numbers during the warmer seasons. A group of organisms found at all latitudes, in surface water and at great depths, are the marine arrowworms (*e.g., Sagitta*), important planktonic predators. Oysters, mussels, other marine bivalves, and snails begin life as planktonic larvae. The wing snails (Pteropoda) spend their entire life cycles as plankton.

Crustaceans are the most important members of the zooplankton. They are the marine counterparts of insects on land; noteworthy is the fact that on land, as in the sea, the arthropods are the most diverse and numerous of all animal phyla. The copepod *Calanus finmarchicus* is important as food for the herring, and the euphausid *Euphausia superba* is the main food source for blue and fin whales in the Antarctic Ocean. Whalebone whales migrate to waters where spawning of these crustaceans occurs; and the rapid growth of these large mammals, feeding entirely on plankton, is impressive.

There is a pronounced tendency for zooplankton to perform diurnal vertical migrations in both lakes and the sea. This migratory behaviour varies with stages in the life cycle, seasons of the year, latitude, hydrographic structure, and meteorological conditions. Generally, the animals ascend toward the surface at sunset from daytime depths. At midnight, if there is no optical stimulus (*e.g.*, moon, artificial light), some of the animals return to the daytime depths, then approach the surface once again just before dawn. As the sun rises, all descend to their daytime level.

Bacteria and fungi. Bacteria and fungi found in water belong by definition to plankton, but because of special techniques required for sampling and identification, they usually are considered separately. These organisms are important in the transformation of dead organic materials to inorganic plant nutrients. Some of these marine and fresh-water micro-organisms (including blue-green algae) fix molecular nitrogen from water containing dissolved air, forming ammonia or related nutrients important for phytoplankton growth. Although little is known about the extent of nitrogen fixation, bacteria and fungi always are found in water samples. A peculiar situation exists in the Black Sea where water below 130–180 metres contains hydrogen sulfide and no oxygen. Under these conditions only bacteria are found.

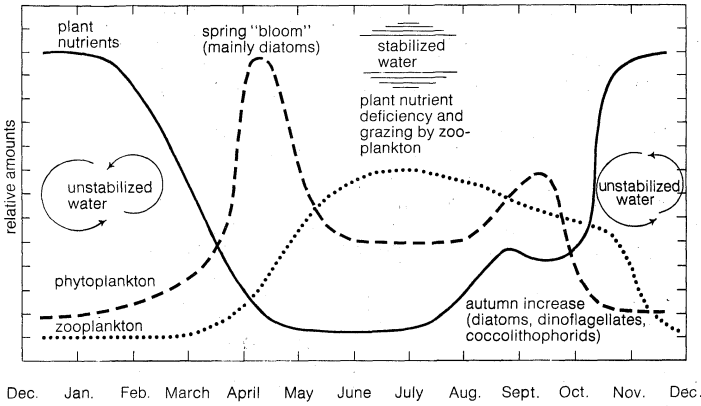
Productivity of plankton. *General features.* The productivity of an area is dependent upon the availability of nutrients and water stability conditions. Currents that flow near continents are important to plankton production in an area. The California Current (a continuation of the Kuroshio Drift from Japan) causes an outland transport of water and combines with a compensating nutrient-rich current along the coast of California to make this area highly productive. The same situation exists along the west coast of southern Africa, which is influenced by the Benguela Current, and off the west coast of South America, influenced by the Peru Current. One of the poorest oceanic areas, the Sargasso Sea, is supplied with water through the North Equatorial Current. Originating at the coast of Africa, the waters are depleted of nutrients as they flow toward the Sargasso Sea. Unstabilized water masses in shallow areas (*e.g.*, Bay of Fundy, Nova Scotia), where phytoplankton is always in the light zone, have high primary production.

Dependence on availability of nutrients and sunlight

In the sea an adequate supply of nutrients, including carbon dioxide, enables phytoplankton and benthic algae to transform the light energy of the sun into energy-rich chemical components through photosynthesis. The bottom-dwelling algae are responsible for about 2 percent of the primary production in the ocean; the remaining 98

percent is attributable to phytoplankton. Most of the phytoplankton serves as food for zooplankton, but some of it is carried below the light zone. After death, this phytoplankton undergoes chemical mineralization, bacterial breakdown, or transformation into sediments. Phytoplankton production usually is greatest from five to ten metres below the surface of the water. Too high light intensity and/or the lack of nutrient in the upper regions may be the causes for this sub-optimal photosynthesis. Although bacteria are found at all depths, they are most abundant either immediately below great phytoplankton populations or just above the bottom. Zooplankton is used directly as food by fish (*e.g.*, herring) or mammals (*e.g.*, whalebone whales); however, several food chains usually have been passed before plankton is available as food for human consumption. Since each step in the chain is calculated as a 90 percent loss of the original energy obtained from the sun, there is considerable energy lost as the number of components in the food chain increases (see BIOSPHERE).

The oceanic cycle. The conditions in northern temperate regions serve as an example of an annual plankton cycle. The growth of phytoplankton is dependent upon



Annual plankton cycle in temperate regions of the Northern Hemisphere.

adequate amounts of required nutrients and light. Temperature is not a critical factor because species with low temperature optima prevail during cold seasons and in low temperature areas. During the cold winter months, the nutrient situation is excellent in northern temperate oceans, but low light levels and complete mixing of water masses (*i.e.*, unstabilized water), which in the North Atlantic prevails down to several hundred metres, combine to keep the phytoplankton population below the light zone. As a result, no photosynthesis occurs and primary production is impossible. During spring the upper water masses are warmed by the sun; this results in lighter water layers on top of colder, heavier ones (*i.e.*, stabilized water). The phytoplankton remain within the light zone; nutrients are available; and a burst in phytoplankton population, consisting of diatoms (*i.e.*, spring diatom bloom), occurs. The high population level falls quickly because the phytoplankton are consumed by zooplankton whose numbers now increase. As dead organisms and feces sink, the upper water masses become depleted of plant nutrients, resulting in moderate production during the summer months. During the autumn months, cooling of nutrient-poor surface water layers makes them heavier and results in their mixing with deeper water rich in nutrients; an autumn increase in phytoplankton of short duration occurs. As the mixing of water masses extends far below the light zone during winter, primary production again diminishes. In polar regions only one summer maximum occurs, while in tropical areas the production is either uniform or irregular throughout the year.

BIBLIOGRAPHY. J.S. BURLEW (ed.), *Algal Culture from Laboratory to Pilot Plant* (1953), an analysis of algal growth conditions with the goal of using algae as human food; A.C. HARDY, *The Open Sea: Its Natural History Part 1, The World of Plankton* (1958), an outstanding general work; D.F. JACKSON (ed.), *Algae and Man* (1964), collection of conference papers on the theme of algae as food for man; J.E.G. RAYMONT, *Plankton and Productivity in the Oceans* (1963) a readable account of this subject; H.U. SVERDRUP, M.W. JOHNSON, and R.H. FLEMMING, *The Oceans: Their Physics, Chemistry, and General Biology* (1942), an advanced text in oceanography, with chapters on plankton; R.S. WIMPENNY, *The Plankton of the Sea* (1966), an excellent work with a physiological approach to the subject; WOODS HOLE OCEANOGRAPHIC INSTITUTION, *Oceanus* (quarterly), a journal containing general articles, including plankton.

(P.H.H.H.)

Plant Breeding

In its modern form, plant breeding involves the application of genetic principles to produce plants that are more useful to man. This is accomplished by selecting plants that man finds desirable, economically or aesthetically, by controlling the mating of the selected individuals and by selecting desirable individuals among the progeny. These processes, repeated over many generations, can change the hereditary makeup and value of a plant population far beyond the natural limits of previously existing populations. This article emphasizes the application of genetic principles to the improvement of plants; the biological factors underlying plant breeding are dealt with in the article HEREDITY.

Historical background

Plant breeding is an ancient activity of man, dating back to the very beginnings of agriculture. Probably soon after the earliest domestications of cereal grains, man began to recognize degrees of excellence among the plants in his fields and saved seed from the best plants for planting new crops. Such tentative selective methods were the forerunners of early plant-breeding procedures.

The results of early plant-breeding procedures were conspicuous. Most present-day varieties are so modified from their wild progenitors they are unable to survive in nature. Indeed, in some cases, the cultivated forms are so strikingly different from existing wild relatives that it is even difficult to identify their ancestors. These remarkable transformations were accomplished by early plant breeders in a very short time from an evolutionary point of view, and the rate of change was probably greater than for any other evolutionary event.

Scientific plant breeding dates back hardly more than 50 years. The role of pollination and fertilization in the process of reproduction was not widely appreciated even 100 years ago, and it was not until the early part of the 20th century that the laws of genetic inheritance were recognized and a beginning was made toward applying them to the improvement of plants. One of the major facts that has emerged during the short history of scientific breeding is that an enormous wealth of genetic variability exists in the plants of the world and that only a start has been made in tapping its potential.

PLANT-BREEDING GOALS

The plant breeder usually has in mind an ideal plant that combines a maximum number of desirable characteristics. These characteristics may include resistance to diseases and insects; tolerance to heat and frost; appropriate size, shape, and time to maturity; and many other general and specific traits that contribute to improved adaptation to the environment, ease in growing and handling, greater yield, and better quality. The breeder of fancy show plants must also consider aesthetic appeal. Thus the breeder can rarely focus attention on any one characteristic but must take into account the manifold traits that make the plant more useful in fulfilling the purpose for which it is grown.

Increase of yield. One of the aims of virtually every breeding project is to increase yield. This can often be brought about by selecting obvious morphological variants. One example is the selection of dwarf, early maturing varieties of rice. These dwarf varieties are sturdy and give a greater yield of grain. Furthermore, their early maturity frees the land quickly, often allowing an additional planting of rice or other crop the same year.

Another way of increasing yield is to develop varieties resistant to diseases and insects. In many cases resistant varieties have been the only practical method of pest con-

trol. Perhaps the most important feature of resistant varieties is the stabilizing effect they have on production and hence on steady food supplies. Varieties tolerant to drought, heat, or cold provide the same benefit.

Modifications of range and constitution. Another common goal of plant breeding is to extend the area of production of a crop species. A good example is the modification of grain sorghum since its introduction to the United States about 100 years ago. Of tropical origin, grain sorghum was originally confined to the southern Plains area and the Southwest, but earlier maturing varieties were developed until grain sorghum is now an important crop as far north as North Dakota.

The influence of mechanized agriculture

Development of crop varieties suitable for mechanized agriculture has become a major goal of plant breeding in recent years. Uniformity of plant characters is very important in mechanized agriculture because field operations are much easier when the individuals of a variety are similar in time of germination, growth rate, size of fruit, and so on. Uniformity in maturity is, of course, essential when crops such as tomatoes and peas are harvested mechanically.

The nutritional quality of plants can be greatly improved by breeding. For example, it is possible to breed varieties of corn (maize) much higher in lysine than previously existing varieties. Breeding high-lysine maize varieties for those areas of the world where maize is the major source of this nutritionally essential amino acid has become a major goal in plant breeding.

In breeding ornamentals, attention is paid to such factors as longer blooming periods, improved keeping qualities of flowers, general thriftiness, and other features that contribute to usefulness and aesthetic appeal. Novelty itself is often a virtue in ornamentals, and the spectacular, even the bizarre, is often sought.

EVALUATION OF PLANTS

The appraisal of the value of plants so that the breeder can decide which individuals should be discarded and which allowed to produce the next generation is a much more difficult task with some traits than with others.

Qualitative characters. The easiest characters, or traits, to deal with are those involving discontinuous, or qualitative, differences that are governed by one or a few major genes. Many such inherited differences exist, and they frequently have profound effects on plant value and utilization. Examples are starchy versus sugary kernels (characteristic of field and sweet corn, respectively) and determinant versus indeterminant habit of growth in green beans (determinant varieties are adapted to mechanical harvesting). Such differences can be seen easily and evaluated quickly, and the expression of the traits remains the same regardless of the environment in which the plant grows. Traits of this type are termed highly heritable.

Quantitative characters. In other cases, however, plant traits grade gradually from one extreme to another in a continuous series, and classification into discrete classes is not possible. Such variability is termed quantitative. Many traits of economic importance are of this type; *e.g.*, height, cold and drought tolerance, time to maturity, and, in particular, yield. These traits are governed by many genes, each having a small effect. Although the distinction between the two types of traits is not absolute, it is nevertheless convenient to designate qualitative characters as those involving discrete differences and quantitative characters as those involving a graded series.

Quantitative characters are much more difficult for the breeder to control, for three main reasons: (1) the sheer numbers of the genes involved make hereditary change slow and difficult to assess; (2) the variations of the traits involved are generally detectable only through measurement and exacting statistical analyses; and (3) most of the variations are due to the environment rather than to genetic endowment; for example, the heritability of certain traits is less than 5%, meaning that 5% of the observed variation is caused by genes and 95% is caused by environmental influences.

It follows that carefully designed experiments are required to distinguish plants that are superior because they carry desirable genes from those that are superior because they happen to grow in a favourable site.

METHODS OF PLANT BREEDING

Mating systems. Plant mating systems devolve about the type of pollination, or transferal of pollen from flower to flower (see POLLINATION). A flower is self-pollinated (a "selfer") if pollen is transferred to it from any flower of the same plant and cross-pollinated (an "outcrosser" or "outbreeder") if the pollen comes from a flower on a different plant. About half of the more important cultivated plants are naturally cross-pollinated, and their reproductive systems include various devices that encourage cross-pollination; *e.g.*, protandry (pollen shed before the ovules are mature, as in the carrot and walnut), dioecy (stamens and pistils borne on different plants, as in the date palm, asparagus, and hops), and genetically determined self-incompatibility (inability of pollen to grow on the stigma of the same plant, as in white clover, cabbage, and many other species).

Other plant species, including a high proportion of the most important cultivated plants such as wheat, barley, rice, peas, beans, and tomatoes, are predominantly self-pollinating. There are relatively few reproductive mechanisms that promote self-pollination; the most positive of which is failure of the flowers to open (cleistogamy), as in certain violets. In barley, wheat, and lettuce the pollen is shed before or just as the flowers open; and in the tomato pollination follows opening of the flower, but the stamens form a cone around the stigma. In such species there is always a risk of unwanted cross-pollination.

Control of breeding

In controlled breeding procedures it is imperative that pollen from the desired male parent, and no other pollen, reaches the stigma of the female parent. When stamens and pistils occur in the same flower, the anthers must be removed from flowers selected as females before pollen is shed. This is usually done with forceps or scissors. Protection must also be provided from "foreign" pollen. The most common method is to cover the flower with a plastic or paper bag. When the stigma of the female parent becomes receptive, pollen from the desired male parent is transferred to it, often by breaking an anther over the stigma, and the protective bag is replaced. The production of certain hybrids is, therefore, tedious and expensive because it often requires a series of delicate, exacting, and properly timed hand operations. When male and female parts occur in separate flowers, as in maize, controlled breeding is easier.

A cross-pollinated plant, which has two parents, each of which is likely to differ in many genes, produces a diverse population of plants hybrid (heterozygous) for many traits. A self-pollinated plant, which has only one parent, produces a more uniform population of plants pure breeding (homozygous) for many traits. Thus, in contrast to outbreeders, self-breeders are likely to be highly homozygous and hence true breeding for a specified trait.

Breeding self-pollinated species. The breeding methods that have proved successful with self-pollinated species are (1) mass selection; (2) pure-line selection; (3) hybridization, with the segregating generations handled by the pedigree method, the bulk method, or by the backcross method; and (4) development of hybrid varieties.

Mass selection. In mass selection, seeds are collected from (usually a few dozen to a few hundred) desirable appearing individuals in a population, and the next generation is sown from the stock of mixed seed. This procedure, sometimes referred to as phenotypic selection, is based on how each individual looks. Mass selection has been used widely to improve old "land" varieties, varieties that have been passed down from one generation of farmers to the next over long periods.

An alternative approach that has no doubt been practiced for thousands of years is simply to eliminate undesirable types by destroying them in the field. The results are similar whether superior plants are saved or inferior plants are eliminated: seeds of the better plants become the planting stock for the next season.

A modern refinement of mass selection is to harvest the best plants separately and to grow and compare their progenies. The poorer progenies are destroyed and the seeds of the remainder are harvested. It should be noted that selection is now based not solely on the appearance of the parent plants but also on the appearance and performance of their progeny. Progeny selection is usually more effective than phenotypic selection when dealing with quantitative characters of low heritability. It should be noted, however, that progeny testing requires an extra generation; hence gain per cycle of selection must be double that of simple phenotypic selection to achieve the same rate of gain per unit time.

Mass selection, with or without progeny test, is perhaps the simplest and least expensive of plant-breeding procedures. It finds wide use in the breeding of certain forage species, which are not important enough economically to justify more detailed attention.

Pure-line selection. Pure-line selection generally involves three more or less distinct steps: (1) numerous superior appearing plants are selected from a genetically variable population; (2) progenies of the individual plant selections are grown and evaluated by simple observation, frequently over a period of several years; and (3) when selection can no longer be made on the basis of observation alone, extensive trials are undertaken, involving careful measurements to determine whether the remaining selections are superior in yielding ability and other aspects of performance. Any progeny superior to an existing variety is then released as a new "pure-line" variety. Much of the success of this method during the early 1900s depended on the existence of genetically variable land varieties that were waiting to be exploited. They provided a rich source of superior pure-line varieties, some of which are still represented among commercial varieties. In recent years the pure-line method as outlined above has decreased in importance in the breeding of major cultivated species; however, the method is still widely used with the less important species that have not yet been heavily selected.

Exploitation of sports

A variation of the pure-line selection method that dates back centuries is the selection of single-chance variants, mutations or "sports" in the original variety. A very large number of varieties that differ from the original strain in characteristics such as colour, lack of thorns or barbs, dwarfness, and disease resistance have originated in this fashion.

Hybridization. During the 20th century planned hybridization between carefully selected parents has become dominant in the breeding of self-pollinated species. The object of hybridization is to combine desirable genes found in two or more different varieties and to produce pure-breeding progeny superior in many respects to the parental types.

Genes, however, are always in the company of other genes in a collection called a genotype. The plant breeder's problem is largely one of efficiently managing the enormous numbers of genotypes that occur in the generations following hybridization. As an example of the power of hybridization in creating variability, a cross between hypothetical wheat varieties differing by only 21 genes is capable of producing more than 10,000,000,000 different genotypes in the second generation. At spacings normally used by farmers, more than 50,000,000 acres would be required to grow a population large enough to permit every genotype to occur in its expected frequency. While the great majority of these second generation genotypes are hybrid (heterozygous) for one or more traits, it is statistically possible that 2,097,152 different pure-breeding (homozygous) genotypes can occur, each potentially a new pure-line variety. These numbers illustrate the importance of efficient techniques in managing hybrid populations, for which purpose the pedigree procedure is most widely used.

Pedigree breeding starts with the crossing of two genotypes, each of which have one or more desirable characters lacked by the other. If the two original parents do not provide all of the desired characters, a third parent can be included by crossing it to one of the hybrid progeny of the first generation (F_1). In the pedigree method superior types are selected in successive generations, and a record is maintained of parent–progeny relationships.

The F_2 generation (progeny of the crossing of two F_1 individuals) affords the first opportunity for selection in pedigree programs. In this generation the emphasis is on the elimination of individuals carrying undesirable major genes. In the succeeding generations the hybrid condition gives way to pure breeding as a result of natural self-pollination, and families derived from different F_2 plants begin to display their unique character. Usually one or two superior plants are selected within each superior family in these generations. By the F_5 generation the pure-breeding condition (homozygosity) is extensive, and emphasis shifts almost entirely to selection between families. The pedigree record is useful in making these eliminations. At this stage each selected family is usually harvested in mass to obtain the larger amounts of seed needed to evaluate families for quantitative characters. This evaluation is usually carried out in plots grown under conditions that simulate commercial planting practice as closely as possible. When the number of families has been reduced to manageable proportions by visual selection, usually by the F_7 or F_8 generation, precise evaluation for performance and quality begins. The final evaluation of promising strains involves (1) observation, usually in a number of years and locations, to detect weaknesses that may not have appeared previously; (2) precise yield testing; and (3) quality testing. Many plant breeders test for five years at five representative locations before releasing a new variety for commercial production.

Bulk-population breeding

The bulk-population method of breeding differs from the pedigree method primarily in the handling of generations following hybridization. The F_2 generation is sown at normal commercial planting rates in a large plot. At maturity the crop is harvested in mass, and the seeds are used to establish the next generation in a similar plot. No record of ancestry is kept. During the period of bulk propagation natural selection tends to eliminate plants having poor survival value. Two types of artificial selection also are often applied: (1) destruction of plants that carry undesirable major genes and (2) mass techniques such as harvesting when only part of the seeds are mature to select for early maturing plants or the use of screens to select for increased seed size. Single plant selections are then made and evaluated in the same way as in the pedigree method of breeding. The chief advantage of the bulk population method is that it allows the breeder to handle very large numbers of individuals inexpensively.

Backcross breeding

Often an outstanding variety can be improved by transferring to it some specific desirable character that it lacks. This can be accomplished by first crossing a plant of the superior variety to a plant of the donor variety, which carries the trait in question, and then mating the progeny back to a plant having the genotype of the superior parent. This process is called backcrossing. After five or six backcrosses the progeny will be hybrid for the character being transferred but like the superior parent for all other genes. Selfing the last backcross generation, coupled with selection, will give some progeny pure breeding for the genes being transferred. The advantages of the backcross method are its rapidity, the small number of plants required, and the predictability of the outcome. A serious disadvantage is that the procedure diminishes the occurrence of chance combinations of genes, which sometimes leads to striking improvements in performance.

Hybrid varieties. The development of hybrid varieties differs from hybridization in that no attempt is made to produce a pure-breeding population; only the F_1 hybrid plants are sought. The F_1 hybrid of crosses between different genotypes is often much more vigorous than its parents. This hybrid vigour, or heterosis, can be manifested in many ways, including increased rate of growth, greater uniformity, earlier flowering, and increased yield, the last being of greatest importance in agriculture.

By far the greatest development of hybrid varieties has been in maize, primarily because its male flowers (tassels) and female flowers (incipient ears) are separate and easy to handle, thus proving economical for the production of

hybrid seed. The production of hand-produced F_1 hybrid seed of other plants, including ornamental flowers, has been economical only because greenhouse growers and home gardeners have been willing to pay high prices for hybrid seed.

Recently, however, a built-in cellular system of pollination control has made hybrid varieties possible in a wide range of plants, including many that are self-pollinating, such as sorghums. This system, called cytoplasmic male sterility, or cytosterility (Figure 1), prevents normal

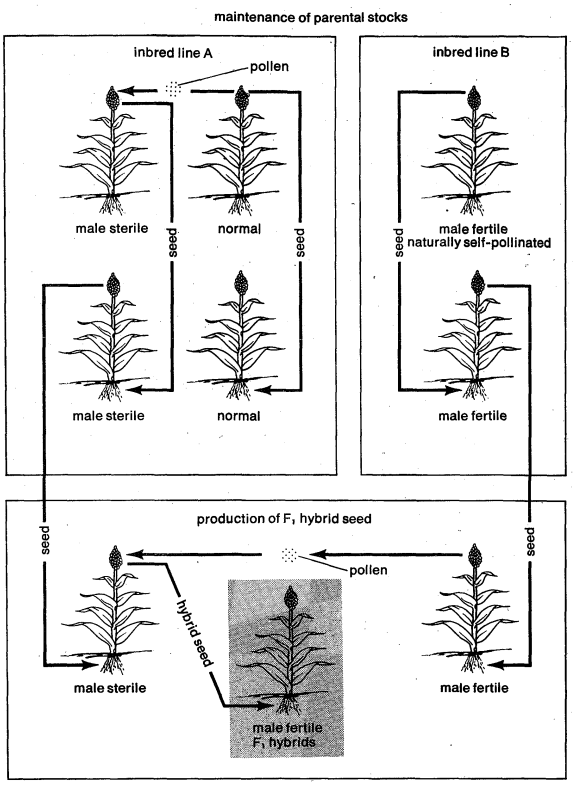


Figure 1: Production of hybrid seed in sorghum using cytoplasmic male sterility.

maturation or function of the male sex organs (stamens) and results in defective pollen or none at all. It obviates the need for removing the stamens either by hand or by machine. Cytosterility depends on the interaction between male sterile genes $(R+r)$ and factors found in the cytoplasm of the female sex cell. The genes are derived from each parent in the normal Mendelian fashion, but the cytoplasm (and its factors) is provided by the egg only; therefore, the inheritance of cytosterility is determined by the female parent. All plants with fertile cytoplasm produce viable pollen, as do plants with sterile cytoplasm but at least one *R* gene; plants with sterile cytoplasm and two *r* genes are male sterile (produce defective pollen).

The production of F_1 hybrid seed between two strains is accomplished by interplanting a sterile version of one strain (say A) in an isolated field with a fertile version of another strain (B). Since strain A produces no viable pollen, it will be pollinated by strain B, and all seeds produced on strain A plants must therefore be F_1 hybrids between the strains. The F_1 hybrid seeds are then planted to produce the commercial crop. Much of the breeder's work in this process is in developing the pure-breeding sterile and fertile strains to begin the hybrid seed production.

Breeding cross-pollinated species. The most important methods of breeding cross-pollinated species are (1) mass selection; (2) development of hybrid varieties; and (3) development of synthetic varieties. Since cross-pollinated species are naturally hybrid (heterozygous) for many traits and lose vigour as they become purebred (homozygous), a goal of each of these breeding methods is to preserve or restore heterozygosity.

Mass selection. Mass selection in cross-pollinated species takes the same form as in self-pollinated species; *i.e.*, a large number of superior appearing plants are selected and harvested in bulk and the seed used to produce the next generation. Mass selection has proved to be very effective in improving qualitative characters, and, applied over many generations, it is also capable of improving quantitative characters, including yield, despite the low heritability of such characters. Mass selection has long been the most important method of breeding cross-pollinated species and still is a major method, especially in the economically less important species.

Hybrid varieties. The outstanding example of the exploitation of hybrid vigour through the use of F_1 hybrid varieties has been with maize. The production of a hybrid maize variety (Figure 2) involves three steps: (1) the selection of superior plants; (2) selfing for several generations to produce a series of inbred lines, which although

Production of hybrid corn

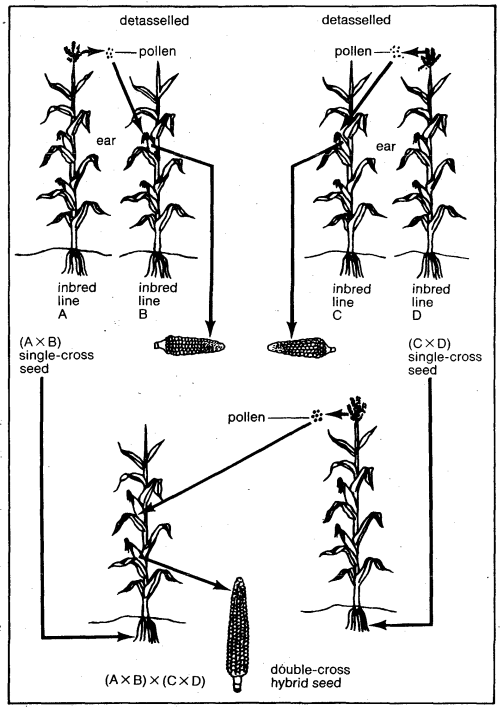


Figure 2: The crossing of inbred lines of maize (corn) to produce single crosses and the crossing of single crosses to produce double-cross hybrid seed.

different from each other are each pure breeding and highly uniform; and (3) crossing selected inbred lines. During the inbreeding process the vigour of the lines decreases drastically, usually to less than half that of field-pollinated varieties. Vigour is restored, however, when any two unrelated inbred lines are crossed, and in some cases the F_1 hybrids between inbred lines are much superior to open-pollinated varieties. An important consequence of the homozygosity of the inbred lines is that the hybrid between any two inbreds will always be the same. Once the inbreds that give the best hybrids have been identified, any desired amount of hybrid seed can be produced.

Pollination in maize is by wind, which blows pollen from the tassels to the styles (silks) that protrude from the tops of the ears. Thus controlled cross-pollination on a field scale can be accomplished economically by interplanting two or three rows of the seed parent inbred with one row of the pollinator inbred and detasselling the former before it sheds pollen. In practice most hybrid corn is produced from "double crosses," in which four inbred lines are first crossed in pairs (A × B and C × D) and then the two F_1 hybrids are crossed again (A × B) × (C × D). The double-cross procedure has the advantage that the commercial F_1 seed is produced on the highly productive single cross A × B rather than on a poor-yielding inbred, thus reducing seed costs. In recent years cytoplasmic male

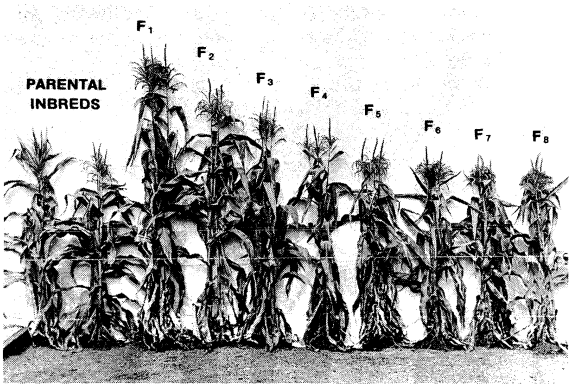


Figure 3: *Heterosis followed by decline in vigour in maize.* Representative plants of the two parental inbreds are at left. To their right are vigorous F_1 hybrids and the F_2 through F_8 selfed generations. The decline in vigour from F_1 through F_8 is similar to that observed when plants from open-pollinated varieties are self-pollinated.
D.F. Jones, *Genetics*, no. 9 (1924)

sterility, described earlier, has been used to eliminate detasselling of the seed parent, thus providing further economies in producing hybrid seed.

Much of the hybrid vigour exhibited by F_1 hybrid varieties is lost in the next generation. Consequently, seed from hybrid varieties is not used for planting stock but the farmer purchases new seed each year from seed companies.

Perhaps no other development in the biological sciences has had greater impact on increasing food supplies of the world than hybrid maize. Hybrid varieties in other crops, made possible through the use of male sterility, have also been dramatically successful and it seems likely that use of hybrid varieties will continue to expand in the future.

Synthetic varieties. A synthetic variety is developed by intercrossing a number of genotypes of known superior combining ability—*i.e.*, genotypes that are known to give superior hybrid performance when crossed in all combinations. (By contrast, a variety developed by mass selection is made up of genotypes bulked together without previous testing of performance in hybrid combination.) Synthetic varieties are known for their hybrid vigour and for their ability to produce usable seed for succeeding seasons. Because of these advantages, synthetic varieties have become increasingly favoured in the growing of many species, such as the forage crops, in which expense prohibits the development or use of hybrid varieties.

Distribution and maintenance of new varieties. The benefits of superior new varieties obviously cannot be realized until sufficient seed has been produced to permit commercial production. Although the primary function of the plant breeder is to develop new varieties, he usually also carries out an initial small-scale seed increase. Such seed is called breeders seed. The next stage is the multiplication of breeders seed to produce foundation seed. Production of foundation seed is usually carried out by governmentally regulated seed associations or institutes. The third step is the production of certified seed, the progeny of foundation seed, produced on a large scale by specialized seed growers for general sale to farmers and gardeners. Certified seed must be produced and handled in such a way as to meet the standards set by the certifying agency (usually a seed association). Seed associations are also usually responsible for maintaining the purity of new varieties once they have been released for commercial production.

The distribution of new varieties developed by commercial plant-breeding companies is often through seed associations, but many reputable companies market their products without following the official certification process. In some countries, particularly in Europe, new varieties can be patented for periods up to 15 years or more, during which time the breeder has an exclusive right to reproduce and sell the variety.

BIBLIOGRAPHY. J.B. HUTCHINSON (ed.), *Essays on Crop Plant Evolution* (1965); W.J.C. LAWRENCE, *Plant Breeding* (1968), a popular account; F.N. BRIGGS and P.F. KNOWLES, *Introduction to Plant Breeding* (1967), an introductory college textbook; J.L. BREWBAKER, *Agricultural Genetics* (1964), a college-level textbook; R.W. ALLARD, *Principles of Plant Breeding* (1960), a college-level textbook emphasizing genetic principles basic to plant breeding; W.D. HANSON and H.F. ROBINSON (eds.), *Statistical Genetics and Plant Breeding* (1963), essays at an advanced level on quantitative aspects of plant breeding; W.H. LEONARD, A.G. CLARK, and E.L. LE CLERG, *Field Plot Technique*, 2nd ed. (1962), a textbook of field plot methods used in plant breeding.

(R.W.A.)

Plant Internal Transport

Movement of water and nutrients takes place in plants, just as the bloodstream distributes nutrients throughout the body of animals. This internal movement, usually called transport, is necessary in all plants, even the most primitive ones, unicellular algae. Nutrients are taken up through the cell membrane, and certain substances may in turn be released into the surrounding water. Transport across membranes is fundamental and may be regarded as the basis for long distance transport (see MEMBRANE, BIOLOGICAL). Molecules are moved actively—*i.e.*, with the expenditure of energy—across a cytoplasmic membrane, from one compartment (*e.g.*, one cell) to another. Active transport

The importance of transport processes in plants increased as multicellular plants evolved and their tissues acquired specialized functions. In large marine algae, for example, carbohydrates (such as sugars) are transported over relatively long distances, from their site of formation in well-illuminated plant parts near the water surface to parts near the sea floor that receive only small amounts of light. As land plants developed, long-distance transport assumed a still more important role; in this case, not only are carbohydrates transported from the organs in which they are formed (the leaves) to other parts—such as reproductive organs (flowers and fruits), stems, and roots—but water and minerals must be transported to leaves, which are not submerged in water (as are those of most primitive plants) but are in a relatively dry air environment. Highly developed land plants, such as trees, have two types of tissues specialized for long-distance transport: the xylem and the phloem. Water and dissolved mineral nutrients ascend in the xylem (the wood of a tree, such as an oak or a pine) and products of photosynthesis, mostly sugars, move from leaves to other plant parts in the phloem (the inner bark of a plant).

Evolving land plants faced not only the problem of transport but also the problem of supporting their weight. Aquatic plants are supported by their buoyancy in water. Even large seaweed do not need a rigid stem; flotation devices such as gas-filled bladders hold them upright and enable them to grow toward the water surface and obtain sufficient sunlight for photosynthesis. On land, a rigid, self-supporting structure is necessary for plants; this structure, the xylem, consists of tiny rigid tubes through which water and dissolved mineral nutrients can move. The rigidity of the tubes within a stem is sufficient to make it self-supporting.

THE TWO MAIN TRANSPORT SYSTEMS, XYLEM AND PHLOEM

Land plants take up water from the soil through the roots; the few rare exceptions, such as some plants of the Peruvian desert that grow in dry soil, rely on adaptations that enable them to obtain water from the air. In most plants, then, water ascends through the xylem, the tiny capillaries of the woody stem tissue, into all plant parts but primarily into the leaves, from which it is transpired (evaporated) into the air. In this way, the mineral nutrients are transferred from the soil to all above-ground plant parts.

Plants living in humid habitats, such as the small and primitive mosses and liverworts, do not have a well-developed xylem. Similarly, water plants that have returned from land to an aquatic habitat during evolution have a reduced xylem; such plants, which have readapted to an aquatic environment, are not woody, because they need neither water-conducting tissues nor a self-supporting structure. On the other hand, tall land plants such as trees and vines (lianas) have the most highly developed long-

distance transport systems. Vines differ from trees in that their xylem serves primarily for water conduction; they depend, for the most part, on other plants for support. Certain vines are of great length (several hundred feet) and have extremely highly developed tissues for transporting water and nutrients.

Most of the material that comprises a plant's dry weight is a consequence of photosynthesis, in which light energy is converted into chemical energy used to synthesize organic substances. Carbon dioxide from the air and water, which the plant takes from the soil, are utilized during photosynthesis, which occurs mostly in green plant parts —especially the leaves (see PHOTOSYNTHESIS). Since plants shed their leaves either continuously or periodically but still increase in size, it is clear that many photosynthetic products must be transported out of the leaves and carried to all other plant parts; this process takes place primarily in the phloem.

The discovery of the functions of xylem and phloem (wood and bark) was made following that of the circulation of blood in the 17th century. By the early 19th century, it had been established that water ascends from roots into leaves through xylem and that photosynthetic products descend through phloem. Experiments now called girdling experiments were performed, in which a ring of bark is removed from a woody plant. Girdling or ringing does not immediately interfere with upward movement of water in the xylem, but it does interrupt phloem movement. In some plants surgical removal of phloem is difficult; in this case phloem may be killed by using steam (steam girdling). Xylem conduction is normally not affected by such treatment, and movement in the two transport tissues can thus easily be distinguished. Girdling experiments, however, are not entirely foolproof. The question as to whether or not mineral nutrients can ascend in the phloem illustrates the kinds of difficulties that may be encountered. Much smaller amounts of mineral nutrients reach the leaves in girdled plants than in ungirdled ones. From this observation it might be concluded that some nutrients ascend in the phloem of ungirdled trees; girdling, however, interrupts the flow of sugars into roots. Roots are thereby starved and take up fewer mineral nutrients; the reduced flow of mineral nutrients to the leaves of girdled plants can thus be explained as a secondary effect.

Girdling experiments

It was once widely believed that flogging a fruit tree increases the crop. The rational explanation of this is simple: flogging a tree damages the phloem; *i.e.*, it girdles the tree, perhaps partially. Girdling increases the sugar concentration of the leaves by restricting downward transport to the roots and may thus directly or indirectly stimulate flowering or, if done after fruit-set, may increase fruit size.

STRUCTURAL BASIS OF TRANSPORT

Cell-to-cell transport. Two features of plant cells differ conspicuously from those of animal cells. In plant cells the protoplast, or living material of the cell, contains one or more vacuoles, which are vesicles containing aqueous cell sap. Plant cells are also surrounded by a relatively tough but elastic wall. Water entering the vacuole by osmosis (*i.e.*, movement of water across a membrane into regions of lower water concentration, such as cell interiors, which normally contain dissolved substances) expands the protoplast and consequently the cell wall until the internal pressure is balanced by the elastic counterpressure of the wall. Spaces between and within cell walls are sufficiently large to permit water to flow around all cells. The space available for free water flow is called apoplast. Water in apoplast originates from the roots and contains nutrients taken up by them. Nutrients enter a cell by crossing the outer cytoplasmic membrane.

Because mature cell walls are dead structures, metabolic activities—the chemical reactions of living systems—occur within protoplasts. Substances can enter a protoplast by cytoplasmic connections between neighbouring cells (plasmodesmata) or by active transport mechanisms requiring energy and a group of enzymes called permeases. Plasmodesmata penetrate neighbouring cell walls at areas called pits. Some substances pass out of cells into the apoplast and are transported by energy-requiring processes into the protoplast of another cell.

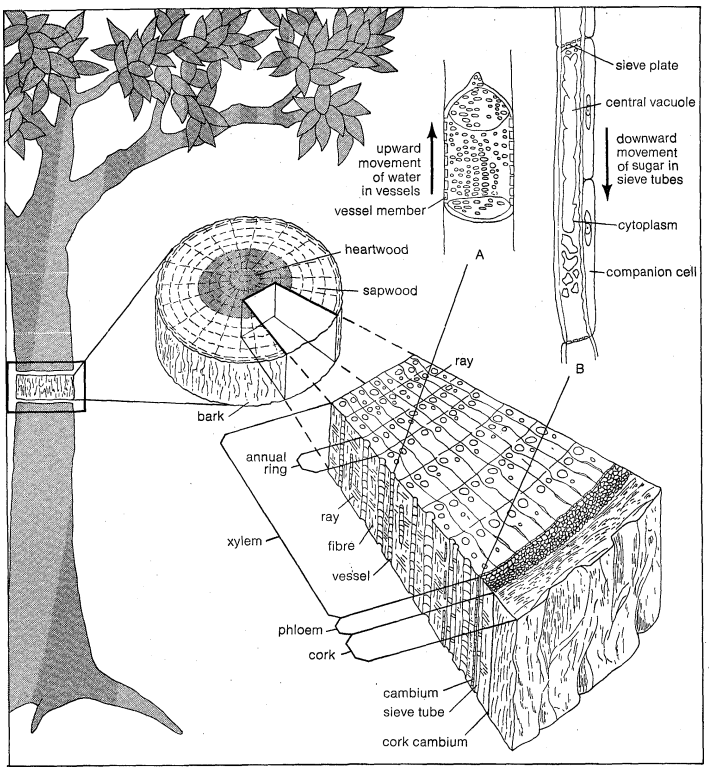
Structure of xylem and phloem. *The xylem.* Cell-to-cell transport takes place in all multicellular plants, but it is a slow process; the higher plants evolved the specialized tissues, xylem and phloem, for rapid long-distance transport. The woody tissue, xylem, contains highly specialized cells for water conduction. The cells are long and reinforced by strong, woody walls; their protoplast breaks down and dissolves after wall growth is completed, so that the entire inside of the cell becomes available for rapid water conduction. In other words, the water-conducting cells of xylem are dead when functional. In the more primitive conifers the xylem consists largely of spindle-shaped cells called tracheids, which have a diameter around 40 microns (.04 millimetres or about 0.0016 inch) and a length of about three millimetres (about 0.12 inch). Flowering plants have a more highly specialized xylem, in which the mechanical function and the water conduction function have been separated during evolution. Tracheids, the primitive conducting cells, have evolved to form fibres for mechanical strength and vessels for water conduction. Vessel elements are barrel-like cells with widths of up to 0.5 millimetres (or about 0.02 inch) in some plants. Vessel elements are arranged end-to-end; their end walls are partly or wholly dissolved, and rows of such cells thus form long capillaries (tubes), up to several metres in length. These tubes are the vessels.

Conducting units—tracheids in conifers and vessels in flowering plants—have overlapping ends that are laterally connected by means of numerous pits. Pits are "windows" in the walls consisting of a thin wall membrane through which water but not an air–water interface (a bubble) can easily pass from one conducting unit to the next. Numerous vessels of limited length thus provide a certain protection against injury: that is, since water pressures in the xylem are often well below zero (*i.e.*, the water is under tension), air will be sucked into any injured xylem vessel and spread immediately throughout it but cannot pass through the wet pit membranes into the uninjured units. Damage is thus confined to the units that are injured and cannot easily spread; in addition, the smaller the conducting unit, the more confined is the damage. Plants with large, highly efficient vessels are much more vulnerable to injury, as is evident, for example, from the vulnerability of the elm, which has large vessels, to the Dutch elm disease, in which the water-conduction vessels are injured by beetle activity and fungal growth. In general, both the less efficient but safer coniferous wood and the more highly efficient but more vulnerable wood of flowering plants have been successful during evolution. Very tall trees occur in both groups; *e.g.*, *Sequoia* among the conifers and *Eucalyptus* among the flowering plants.

The phloem. The conducting elements of the phloem underwent evolutionary changes somewhat similar to those of the xylem. The conducting elements of conifers, called sieve cells, are similar in shape and dimensions to tracheids. They do not have a woody wall, however, and they are alive at functional maturity even though their cytoplasm may be highly specialized and the cells have usually lost their nucleus during development. In flowering plants the conducting elements in the phloem, called sieve elements, are arranged end-to-end to form sieve tubes, a name derived from the sievelike end walls through which passage of food from one cell to the next occurs. Sieve cells and sieve elements are almost invariably accompanied by special companion cells believed to control, to a certain extent, the metabolism of the nucleus-free conducting cells.

Phloem cells

Vascular tissue. In primary tissues, such as leaves, strands of phloem and xylem run parallel, forming conducting bundles. These form a network of veins that can be seen when a leaf is held against a light. Xylem sap from the stem enters a leaf through the petiole (leaf stalk) and spreads throughout the leaf blade in the xylem part of the conducting bundles, or veins. Photosynthetic products enter the phloem tissues of the veins, and they are



Internal transport system in a tree.
(A) Enlarged xylem vessel; (B) enlarged mature sieve element.
Adapted from (tree trunk wedge) M.H. Zimmerman, "How Sap Moves in Trees." Copyright © 1963 by Scientific American, Inc. All rights reserved; (A, B) K. Esau, *Plant Anatomy*. Copyright © 1965 by John Wiley and Sons, Inc.

transported from the leaf through the phloem of the petiole.

Throughout most of the plant, xylem and phloem channels run more or less in an axial, or longitudinal, direction, although this path may not be precisely parallel with the stem axis. In most stems of monocotyledonous plants —flowering plants with one seed leaf, such as lilies, grasses, palms—the tissue is primary; *i.e.*, it is formed entirely by the growing shoot tip. Because such plants have no cambium tissue in their stems, they have no secondary growth and thus show relatively little increase in stem diameter. The conducting tissue consists of narrow strands, containing both xylem and phloem, distributed throughout the plant. A single crosswise section through the stem of a coconut palm may contain 20,000 vascular bundles, which are connected into a complex network by numerous cross connections; if the stem is injured, alternate pathways of transport are available.

Tree stems of both conifers and dicotyledonous plants (flowering plants with two seed leaves, such as oaks, maples, elms) contain a cell layer called the vascular cambium that gives rise to secondary xylem and phloem. With each season's secondary growth, the stems increase in diameter. Xylem is formed as a woody cylinder on the inner side of the cambium layer near the centre of the stem. The vessels are deposited by the cambium as a complex network in such a way that an injury to the stem interrupts the local flow of water, but water continues to flow around the injury. Phloem is formed toward the outside of the stem from the cambium layer, forming the inner, conducting bark. The conducting phloem layer remains very thin, less than 0.5 millimetre, and is renewed each year. Older phloem serves only for food storage. Conducting and storing phloem comprise the living bark.

Both xylem and phloem have limited longevity. The oldest phloem layers are the outermost—the dead bark of the stem surface. The yearly amounts of xylem visible in cross sections of stems as distinct rings are known as annual rings. The oldest xylem layers (*i.e.*, the oldest annual rings) are in the dead central core, or heartwood, of the woody stem, which can often be recognized by its darker coloration. The lighter coloured sapwood is living and functions as storage tissue and, especially in the outermost sapwood, as conducting tissue; the younger annual rings comprise the sapwood. In some highly specialized tree species with large vessels (such as some oaks, ashes, and others) only the very outermost growth ring functions in water conduction. Conducting tissues seldom run straight along a tree stem; usually they are arranged in a helical or spiral pattern, sometimes called the spiral grain of a tree. The angle of the spiral arrangement usually changes from year to year; the path of water up a tree stem may therefore be very complicated if more than one growth layer acts as a conducting tissue. Functionally, the effect of the variable spiral grain is to distribute water to all parts of the tree from any root.

Ray tissue

An important tissue in the stems of conifers and dicotyledonous trees is the ray tissue. Rays, small strands of radially running tissue, are continuous from the stem centre through the xylem and phloem and serve for radial conduction and storage of starch. No rays are present in monocotyledonous stems, however; more or less radially running bundles that originally connected the stem with leaves, which have fallen from lower stem parts, serve the same purpose.

Gas exchange and transport in leaves. Although some transpiration (evaporation) of water occurs from the surfaces of all leaf cells, most of it takes place through special structures called stomata; these structures are pores whose openings are controlled by a flexible pair of cells called guard cells. The movement of guard cells is controlled so that they open during daylight hours and close at night or under conditions of excessive water stress. The stomatal opening not only permits the escape of water vapour but also allows carbon dioxide to enter leaves for photosynthesis. Photosynthesis ceases as night falls, and excessive water loss is prevented by stomatal closure.

Most water, which is taken up by roots and ascends into leaves, is lost by transpiration. Transpiration is necessary for two reasons. One is the necessity of keeping stomata open throughout the daylight hours for the entry of carbon dioxide. The second is the fact that xylem sap is very dilute; *i.e.*, it contains much water. In order for sufficient quantities of soil nutrients to be accumulated in leaf cells, large amounts of water must be transpired.

Stomata are very efficient in transpiring water. Water molecules that emerge from such a small pore can escape without interference in all radial directions. (From an open surface water molecules can escape only in a straight line because of mutual interference of neighbouring molecules.) Thus, water can evaporate through a multiperforate membrane (with many pores) almost as efficiently as from an open water surface of equal size even though the total area of all the pores is extremely small. The advantage of the multiperforate membrane is that the size of the pore opening can be controlled.

Xylem sap entering leaves from the stem spreads out into the xylem tissues of all veins and from vein ends into the walls of other leaf cells. Thus surrounded by xylem sap, leaf cells live in an environment not unlike that of algal cells in water. Solutes—dissolved nutrients—brought up from roots can be taken up from the xylem sap into the cell interior. Any solute not actually used by the leaf cells can either be transported to the leaf surface or transferred into the phloem to be returned to the plant body for use in other plant parts; certain substances of no value to the plant are deposited in the form of crystals, sometimes in special crystal cells along the phloem path.

Movement of gases in leaves

Leaf tissue is very spongy—*i.e.*, the cells are surrounded by an abundance of communicating gas spaces. Water evaporates from cell walls into the gas space, diffuses into the stomatal cavity, and then moves into the air through the stomatal pore between the guard cells. Within the gas spaces of the leaf, gases presumably move by diffusion and perhaps, as leaves are moved about by wind, by convection.

The bark of the stem also contains stomata when young and secondary structures of similar function—openings in the bark called lenticels—when older. Very young stems are usually green and function somewhat like

leaves; *i.e.*, they photosynthesize and transpire. Photosynthesis in older stems is negligible although transpiration continues. Transpiration through lenticels takes place in the summer, but closure of lenticels largely prevents transpiration in the winter.

Uptake of water and mineral nutrients from the soil. Water uptake from the soil by root cells is passive, in that water may be pulled into the root by low xylem pressure and also follows osmotic gradients caused by the mineral nutrients, which are taken up actively (*i.e.*, with the expenditure of metabolic energy) across root cell membranes. As the mineral nutrients—the ions (charged components) of inorganic salts—are taken up, they are largely incorporated into organic molecules. Thus, the solutes in xylem sap are mostly complex organic substances, sometimes of a specific nature; for example, nicotine synthesis takes place in the roots of tobacco plants, where nitrogen is incorporated into compounds that have moved to the roots through the phloem as sugars. If a tomato shoot is grafted onto a tobacco rootstock, nicotine-containing tomato leaves are formed. On the other hand, a tobacco shoot grafted onto a tomato rootstock results in a plant with nicotine-free tobacco leaves. Many other specific nitrogen-containing substances originate in the roots; in most plants, however, nitrogen is transported to the leaves from the roots in the form of compounds known as amino acids and amides.

The major chemical elements needed by a plant are: carbon, hydrogen, oxygen, phosphorus, potassium, nitrogen, calcium, iron, and magnesium; in addition, many others are required in very small amounts. A lack of any element may result in deficiency diseases. A few elements taken up by plants are of no nutritive value and usually are eliminated (*e.g.*, sodium chloride) or crystallized (*e.g.*, silica), sometimes by deposition in special cells.

The plant is able to control to some extent the substances that enter. If equal amounts of sodium and potassium are available to roots of plants, and the amount of the two elements inside the plant is analyzed, less sodium is likely to be found than potassium, which is a vital nutrient. The structural basis for the control of uptake of substances into roots is the so-called Casparian strip, a conspicuously thickened wall area of one cell layer around primary roots; it prevents excess soil solution from being pulled directly into the central part of the root where the xylem is located. As a result, the soil solution has to pass through a cell barrier in which uptake can metabolically be controlled. After nutrients are inside living root cells and have been converted to appropriate compounds, the latter are released into the xylem and move to above-ground parts.

Casparian strips

XYLEM TRANSPORT

Quantitative aspects. The sum of the cross-sectional areas of all the branches of a tree approximately equals that of the stem; this means that the total amount of conducting tissue remains about the same from roots to leaves. In terms of water movement, the velocity of movement might be expected to be uniform throughout the entire axial system of stem, branches, and twigs. Because some trees (*e.g.*, oaks) have thick twigs, however, the velocity of water movement is greater in the stem than in the twigs at any time. Similarly, in tree species with slender branches (such as birches), the reverse is true. Normally the proportion of xylem to leaves supplied by that xylem is greater in plants growing in dry habitats than in plants found in wet ones and may be as much as 700 times greater in certain desert plants than in aquatic plants and herbs of relatively humid forest floors. The leaves of dry-habitat plants thus are more richly supplied with water-conducting xylem tissue than are those of moist habitats.

The velocity of sap movement in trees varies throughout a 24-hour period. During the night, especially a rainy night, sap flow may stop; velocity increases with daylight, peak rates being found in the early afternoon. Peak velocities correlate with vessel size; the rate of sap flow in trees with small vessels is about two metres (seven feet) per hour; that in trees with large vessels, about 50 metres (160 feet) per hour. The energy required to lift water in both cases is comparable; in large-porous trees water simply moves faster through fewer and larger vessels.

Mechanism of sap ascent. It was demonstrated about 1900 that living cells of the stem are not responsible for water movement. It is now generally recognized that water in the xylem moves passively along a gradient of decreasing pressures. Under certain special conditions water is pushed up the stem by root pressure. This may be the case with herbaceous (nonwoody) plants in the greenhouse under conditions of ample water supply and little transpiration. In nature, these conditions may be met in early spring before the leaves emerge, when the soil is wet and transpiration is low. Under such conditions, water movement is caused by active uptake of ions (charged particles) and by the entry of water from the soil into the roots. Most of the time, however, water is pulled into the leaves by transpiration. A gradient of decreasing pressures from the base to the top of a tree can be measured, even though pressures are low.

A vacuum pump cannot pull water to a height of more than ten metres (about 33 feet). Since many trees are far taller than 33 feet, the mechanism by which they move water to their crowns has been sought. Is it possible for trees to pull water into their crowns along a decreasing pressure gradient or do they employ some other mechanism? If trees pull water, that in the xylem would have to be held on the tracheid and vessel walls by adhesion, and water molecules would have to hold together by cohesion. The hypothesis that water is pulled upward along a pressure gradient during transpiration has been called the cohesion theory. Two critical requirements of the cohesion mechanism of sap ascent are: (1) sufficient cohesive strength of water and (2) existence of tensions (*i.e.*, pressures below zero) and tension gradients in stems of transpiring trees.

The cohesion theory of sap ascent

Although the tensile strength of water is very high, an excessive pull exerted on a water column will break it. The tallest trees are about 330 feet (100 metres) high. A nonmoving water column at an atmospheric pressure of one atmosphere at the base of the tree is exposed to a pressure of minus nine atmospheres (*i.e.*, a tension of nine atmospheres) at the top. Calculations show that, under conditions of peak flow at midday, this gradient increases by about 50 percent; in other words, a transpiring sequoia would have a pressure in the xylem of minus 14 atmospheres at the top if the basal pressure is one atmosphere. If the pressure at the base is minus ten atmospheres because of dry soil, however, the pressure at the top drops to minus 25 atmospheres.

It has now been shown that negative pressures and gradients of negative pressures do exist in trees with an ingeniously simple device called the pressure bomb. A small twig is inserted in a container (the pressure bomb), its cut stump emerging from a tightly sealed hole. As pressure is applied to the container and gradually increased, water from the xylem emerges from the cut end as soon as the pressure being applied is equal to the xylem tension that existed when the twig was cut. This method has been used to measure gradients of negative pressures in trees; the cohesion theory thus would seem to explain the sap-ascent mechanism.

PHLOEM TRANSPORT

Quantitative aspects. Products of photosynthesis, primarily sugars, move through phloem from leaves to growing tissues and storage organs. The rate at which these substances are transported can be measured in various ways; *e.g.*, as velocities in distance traveled per unit time or as mass transfer in (dry) weight transported per unit time. Velocities appear to be graded—*i.e.*, some molecules move faster than others within the same channel. Peak velocities of molecules usually are of the order of 100 to 300 centimetres (40 to 120 inches) per hour. Average velocities, more difficult to measure but significant in mass-transfer considerations, are lower.

Velocity of phloem transport

Mass transfer can be measured by weighing a storage organ, such as a potato tuber or a fruit, at given time intervals during its growth. Mass transfer per cross-sec-

tional area of conducting tissue is referred to as specific mass transfer and is expressed as grams per hour per square centimetre of phloem or sieve tubes. With a given specific mass transfer, the velocity with which a liquid of a certain concentration flows can be calculated; in dicotyledonous stems, for example, specific mass transfer is between ten and 25 grams per hour per square centimetre of sieve tube tissue at times of peak performance. In certain tree species the sieve tubes can be tapped to obtain an exudate. The concentration of this exudate, multiplied by the measured average velocity, is of the same order of magnitude as specific mass transfer, indicating that liquid movement through sieve tubes could account for transport.

Much of the experimental work on phloem transport now is done with the aid of radioactive substances; radioactive carbon dioxide administered to an illuminated leaf is incorporated into sugar during photosynthesis and carried from the leaf. The velocity of this movement can be measured by determining the arrival of radioactivity at given points along the stem. Whole plants, as long as they are reasonably small, can be pressed against photographic film after the conclusion of a similar experiment, and the photographic image will indicate the portions to which radioactive sugar has moved.

Mechanism of phloem transport. The mechanism of phloem transport has been studied by plant physiologists for many years. A number of hypotheses have been put forth over the past years, but none is entirely satisfactory. One fundamental question is whether sugars and other solutes move en masse as a flowing solution or whether transport takes place independently of the solvent water. The phenomenon of exudation from injured sieve tubes supports the first possibility, which has been further supported by a discovery involving aphids (phloem-feeding insects); when aphids are cut away from plants while feeding, their mouthparts remain embedded in the phloem. Exudate continues to flow through the mouthparts; the magnitude of the rate of this exudation indicates that transport within the sieve tube to the mouthparts occurs as a flow of solution.

Evidence against solution flow is the movement of substances in opposite directions through a section of phloem at any one given time. This, however, has never been convincingly demonstrated in just one sieve tube. On the other hand, unsuccessful attempts have been made to find simultaneous movement of sugars and water along a phloem path in order to demonstrate solution flow.

Mass flow

Mass-flow hypotheses include the pressure-flow hypothesis, which states that flow into sieve tubes at source regions (places of photosynthesis or mobilization of storage products) raises the osmotic pressure in the sieve tube; removal of sugars from sieve tubes in sink regions —*i.e.*, those in which sugars are removed for growth and storage—lowers it. Thus a pressure gradient from source to sink is established in sieve tubes that would allow solution flow. The electroosmotic hypothesis postulates that solution is moved across all sieve plates (areas at which individual sieve elements end) by an electrical potential that is maintained by a circulation of cations (positively charged chemical ions), such as potassium. Transport hypotheses postulating solute movement independent of solvent water include the spreading of solute molecules between two liquid phases and the active movement of molecules by a type of cytoplasmic movement that is frequently referred to as cytoplasmic streaming.

TRANSPORT AND PLANT GROWTH

Storage and circulation. It is important to realize that the plant, with its two transport systems, xylem and phloem, is able to move any substance to virtually any part of its body; the direction of transport is usually opposite in the two systems, and transfer from one system to the other takes place easily. An exception is transport into flowers and certain fruits, in which flow in each system is unidirectional.

Numerous substances move from roots to mature leaves through xylem and are transferred from the leaves, together with sugars, through the phloem to other plant parts. In the autumn months in temperate regions, plants store most of the products resulting from photosynthesis during the summer months in structures such as stems, bulbs, and tubers and mobilize it in the spring when new growth begins. A few plants, such as some tropical monocotyledons (certain palms, for example), store food for many years for use at the time of flowering and fruit-set at the end of their life-span.

Transport of growth-regulating substances. Plant hormones, or growth regulators, are effective in very small amounts; they induce or enhance specific growth phenomena. Because the site of hormone synthesis is different from its place of action, hormones must be transported before they can exert their effects. There are several groups of plant hormones; each has different effects on plant growth. The best known plant hormones are the auxins, the most common of which is called indole acetic acid. Auxins are formed in young, growing organs, such as opening buds, and are transported away from tips of shoots toward the base of the plant, where they stimulate the cells to elongate and sometimes to divide. Responses to gravity and light are also under auxin control. Auxin moves to the lower side of a leaning stem; cells on the lower side then elongate and cause the stem to bend back to a vertical position. Response to gravity in many roots is the opposite of that in shoots; the same mechanism of auxin distribution is responsible, but roots react to different quantities of the hormone than do shoots. Similar auxin distributions are responsible for phototropic responses—*i.e.*, the growth of plant parts such as shoot tips and leaves toward light. In certain cases auxin may be destroyed on the illuminated side.

Auxin transport

Auxin is not normally transported through vascular tissue; moreover, transport is polar—*i.e.*, it takes place along the stem from tip to base, regardless of the stem's position. Velocities of transport are of the order of five to ten millimetres (0.2 to 0.4 inch) per hour, and transport requires the expenditure of metabolic energy. There is evidence that most growth hormones can be transported through xylem or phloem, but, at least in the case of auxin, the transport mechanism is specific.

Hormone transport is also involved in the stimulation of flowering. In some plants, flowering is triggered by short or long days. The receptor of this stimulus is in the leaves. A chemical substance, probably a flowering hormone of as yet unknown nature, then moves to the shoot apex and causes a transformation of the vegetative growing point into a flowering shoot.

Many growth-correlating phenomena are effected by transported hormonal stimuli. A vigorously growing terminal (topmost) shoot may inhibit lateral buds lower down from growing out and may force later branches to bend down. If the terminal shoot is removed, laterals grow out and topmost lateral branches bend upward. In leaning trees with secondary tissue (wood), the cambium produces compression wood on the lower side (in conifers) or tension wood on the upper side (in dicotyledons) in response to a hormone; the stem responds by pushing (in conifers) or pulling (in dicotyledons) itself upright. Transport of growth-regulating substances is thus largely responsible for the characteristic shape of each plant species (see also HORMONE).

BIBLIOGRAPHY. The following three books contain extensive bibliographies through which the reader will find his way into every aspect of the literature: K. ESAU, *Plant Anatomy*, 2nd. ed. (1965), an excellent college-level textbook on anatomy, with chapters on xylem, phloem, leaves, and roots; M.B. WILKINS (ed.), *The Physiology of Plant Growth and Development* (1969), a graduate-level book, containing chapters on transport of nutrients, growth regulators, mineral uptake, photosynthesis, and movement of stomata; M.H. ZIMMERMANN and C.L. BROWN, *Trees: Structure and Function* (1971), a profusely illustrated college-level tree physiology textbook, discussing xylem and phloem transport as well as growth correlations in trees. For continuing reviews, the reader is also referred to the *Annual Review of Plant Physiology* and to supplement volumes to the *Handbuch der Pflanzenphysiologie (Encyclopedia of Plant Physiology)*.

(M.H.Z.)

Plasma State

The plasma state of matter is unique like the solid, liquid, and gaseous states, and is sometimes referred to as the fourth state of matter. A plasma is an electrically conducting medium, in general consisting of negatively charged electrons, positively charged ions (atoms or molecules that have lost one or more of their electrons), and atoms or molecules or both. Its properties depend on the collective behaviour of the particles, as distinct from the individual; and the number of positive charges is almost equal to the number of negative charges. Nearly all the matter in the universe exists in the plasma state, occurring predominantly in this form in the Sun and stars and in interstellar space. Auroras, lightning, and welding arcs are also plasmas; plasmas exist in neon and fluorescent tubes, the crystal structure of metallic solids, and in many other phenomena and objects.

A plasma may be produced in the laboratory by heating a gas to an extremely high temperature, which causes such vigorous collisions between its atoms and molecules that they dissociate into electrons and ions. Plasmas may be partially or completely ionized. A completely ionized hydrogen plasma, consisting solely of electrons and protons (hydrogen nuclei), is less complex than a mixture of atoms or molecules and so is the most elementary aggregation of matter known. In certain circumstances, a plasma obeys the laws of magnetohydrodynamics, the study of the behaviour of electrically conducting fluids in the presence of electric and magnetic fields.

Relation to laws of magneto-hydro-dynamics

DEVELOPMENT OF PLASMA PHYSICS

The modern concept of the plasma state is of recent origin, dating back only to the early 1950s. Its history is interwoven with many disciplines. Three basic fields of study have made unique contributions: electrical discharges, magnetohydrodynamics, and kinetic theory.

Interest in electrical-discharge phenomena may be traced back to the beginning of the 18th century, with three English physicists, Michael Faraday in the 1830s and Joseph John Thomson and John Sealy Edward Townsend at the turn of the 19th century, laying the foundations of the present understanding of the phenomena. Irving Langmuir introduced the terminology plasma in 1923, while investigating electrical discharges. In 1929 he and Levi Tonks, another physicist working in the United States, used the term to designate those regions of such a discharge in which certain oscillations of the negatively charged electrons could occur. They called these oscillations plasma oscillations, their behaviour suggesting that of a jelly-like substance. Not until 1952, however, when two more United States physicists, David Bohm and David Pines, first considered the collective behaviour of electrons in metals as distinct from that in ionized gases was the general applicability of the concept of a plasma fully appreciated.

The collective behaviour of charged particles in magnetic fields is implicit in magnetohydrodynamic studies, the foundations of which were laid in the early and middle 1800s by Faraday and André-Marie Ampère of France. Not until the 1930s, however, when new solar and geophysical phenomena were being discovered, were many of the basic problems of the mutual interaction between ionized gases and magnetic fields considered. In 1942 Hannes Alfvén, a Swedish physicist, introduced the concept of magnetohydrodynamic waves.

Kinetic theory

These two separate approaches—the study of electric discharges and the study of the behaviour of charged particles within magnetic fields—were unified by the introduction of the kinetic theory of the plasma state, which holds that plasma, like gas, consists of particles all in random motion. In 1905 the Dutch physicist Hendrik Antoon Lorentz applied the kinetic equation for atoms (the formulation by the Austrian physicist Ludwig Eduard Boltzmann) to the behaviour of electrons in metals. Various physicists and mathematicians in the 1930s and 1940s developed the plasma kinetic theory further to a high degree of sophistication, and since the early 1950s interest has increasingly focussed on the plasma state itself. Space exploration, the development of electronic devices, a growing awareness of the importance of magnetic fields in astrophysical phenomena, and the quest for controlled thermonuclear (nuclear-fusion) power reactors all have stimulated such interest. Problems remain for the theory of plasma state, especially involving the simultaneous solution of James Clerk Maxwell's electromagnetic equations determining electric and magnetic fields and kinetic equations determining statistical behaviour of plasma particles.

GENERAL CONCEPTS

Although a gas normally acts as an electrical insulator, at sufficiently high temperatures it becomes a conductor. Increased thermal agitation of the molecules and atoms causes them to become ionized as they collide, disintegrating into electrons and positively charged entities called ions. At sufficiently high electron densities an ionized gas becomes a plasma. A basic property of a plasma is its electrical quasi-neutrality; that is, the numbers of negative charges and of positive charges per unit volume of plasma are approximately equal. At sufficiently high electron densities such quasi-neutrality is inevitable. Any excess concentration of either charge would lead to the development of large electrostatic restoring forces resulting from the attraction between opposite charges. Just as a weight on a spring will bob up and down about its rest position, any general displacement of the electrons as a group with respect to the positive ions leads to oscillation of the electrons as a whole about an equilibrium state. These movements were the plasma oscillations studied by Langmuir and Tonks.

Plasma properties and parameters. In general, a plasma is a complex mixture of many different particle types, each distinguishable by its mass and charge and with all particles in rapid thermal motion. The basic properties of a plasma as matter are essentially independent of its chemical properties; they are determined primarily by conservation of energy and momentum laws and by the behaviour of the electrons. The study of the extremely complex interactions among the vast numbers of particles requires the use of statistical methods. Number densities or number per unit volume (n) and kinetic temperatures (T) are defined for each type of particle present. The average kinetic energy per particle is 1.5 times Boltzmann's constant (k) and the kinetic temperature, or $\frac{3}{2}kT$. As in a gas, kinetic pressures are defined as equal to the number density times Boltzmann's constant times temperature, or nkT. In thermodynamic equilibrium, all temperatures are the same as that of the enclosure, and (ignoring particle interactions) the sum of the kinetic pressures is the force per unit area exerted on the enclosure. A basic classification of plasmas can be made in terms of the electron-number density (n_e, the number of electrons per unit volume), the electron temperature (T_e), and the degree of ionization, or the fraction of the plasma ionized. One of the most important parameters associated with a plasma is the Debye length, a concept similar to that introduced in 1923 by the chemists Peter W. Debye and Walter K.F. Hückel in their theory of strong electrolytes (compounds that ionize easily and conduct electricity). In a plasma the Debye length is the maximum distance over which a large difference in the number of positive charges and negative charges can occur. It is equal to a constant times the square root of the ratio of the electron temperature to the electron number density; that is, $h = 69(T_e/n_e)^{1/2}$, in which the Debye length h is given in metres, n_e is the number per cubic metre, and T_e is given in degrees Kelvin. For distances appreciably greater than the Debye length, quasi-neutrality must exist, whereas for distances less than the Debye length an individual positively charged ion can exert a force on an electron. This phenomenon leads to a quantitative definition of a plasma: an ionized gas is a plasma provided the physical dimensions of the gas are much greater than the Debye length. For disturbances on a scale greater than the Debye length, the plasma behaviour is macroscopic, or collective; whereas, on a scale smaller than the Debye length, individual particle behavior can occur.

The basic properties of plasma

When the plasma is not in equilibrium and, especially, when it contains electrical currents (*i.e.*, general motion of the electrons relative to the ions), the particle interactions within a sphere having a Debye radius are best accounted for in terms of equivalent binary collisions. That is, it is possible to add together the many simultaneous interactions between particles within such a sphere in such a way that they may be represented as a series of discrete events between two particles only. The number of such events is called the collision frequency.

Another important measure used in plasma physics is the angular frequency (ω_p) of the plasma oscillations noted by Langmuir and Tonks; it is equal to the average thermal speed of the electron divided by the Debye length, giving a value proportional to the square root of the electron density number, or $\omega_p = 56\sqrt{n_e}$ radians per second.

Elastic and inelastic collisions. In any gas, separate collision frequencies are defined for collisions between all different particle types and all different types of collisions. The total collision frequency for a particular particle type is the sum of these separate frequencies. Two basic types of collisions exist: elastic and inelastic; in an elastic collision, the kinetic energy of colliding particles remains the same after a collision as before; in an inelastic collision, a fraction of the kinetic energy is transferred to the internal energy states of the colliding particles. In an atom, for example, the electrons have certain allowed (discrete) energies and are said to be bound. In a collision a bound electron may be raised from a low to a high energy state, or excited. This, however, can occur only by the expenditure of kinetic energy and only if the kinetic energy exceeds the difference between the two energy states. If the energy is sufficient, a bound electron may become a free electron, and the atom is said to be ionized; the minimum, or threshold, energy required to free an electron is called the ionization energy. Inelastic collisions may also occur with positive ions if electrons are still attached to them. In general, only collisions of electrons and photons (electromagnetic radiation particles) with atoms and ions are significant in these inelastic collisions; ionization by a photon is called photo-ionization.

Collisions defined

A molecule also has discrete energy states, which may be excited by particle collisions. At sufficiently high energies of interaction, the molecule can dissociate into atoms or into atoms and atomic ions. Again, collision of electrons and photons with molecules may cause ionization, producing molecular ions. In general, the reaction rate for inelastic collisions is similar to that of chemical reactions. At sufficiently high temperatures, the atoms are stripped of all electrons and are bare atomic nuclei, and at temperatures of about 1,000,000° K (degrees Kelvin) or greater, nuclear reactions can occur—another form of inelastic collisions. When such reactions lead to the formation of heavier elements the process is called thermonuclear fusion; excess mass is transmuted into kinetic energy (that of motion), and kinetic energy is gained instead of lost.

Radiation from plasmas. Elastic collisions occur between all particles, a fraction of the energy and momentum of one particle being exchanged with the other. The amount of energy exchanged depends on the mass ratio of the colliding particles. On the average, the energy exchanged between one particle type and another is proportional to the temperature difference and the collision frequency. In basically elastic collisions of electrons with ions, however, all of the energy is not exchanged; some is radiated; such radiation is called *bremsstrahlung* (German: "braking radiation") and is associated with the deceleration of electrons in a collision. The power (energy per unit of time) radiated is proportional to the square root of the electron temperature, the radiation being emitted as a continuum of wavelengths.

A similar process occurs in a magnetic field. In such a field an electron gyrates about a magnetic line of force in a helix, a configuration similar to a so-called spiral staircase, the frequency of rotation being called the cyclotron frequency. Charged particles moving in an arc or circle radiate energy, and when an electron moves in a helix about a magnetic line of force, an electromagnetic wave is radiated. At low electron energies an electron radiates at integral multiples of the cyclotron frequency, and at the highest energies the radiation becomes a continuum of frequencies, called synchrotron radiation, a phenomenon associated with similar processes in a synchrotron particle accelerator.

Synchrotron radiation

Other forms of radiation met with in plasma physics include line and recombination radiation. The lifetime of an atom or ion in an excited state is only about 1/100,000,000 of a second. When an electron goes from a high to a low energy state, a photon is emitted, with its frequency proportional to the energy difference between the two states. This emission is called line radiation because the photon would be observed as a line in a spectrometer. In another process, a free electron can be captured by an ion, followed by the emission of a photon; this is called radiative recombination, the associated radiation being a continuum of frequencies with a low frequency cutoff.

Emitted radiation may be absorbed and re-emitted from one atom or molecule to another, and depending on the processes, frequency changes can occur. If the distance travelled after emission and before absorption is small, compared with the dimensions of the plasma, the plasma is said to be optically thick. In this case, the spectrum of the radiation emitted from the surface is characteristic of an ideal blackbody (which emits radiation at the same rate as it absorbs it) and a blackbody temperature may be assigned; the power radiated is proportional to the fourth power of the blackbody temperature. If, however, the absorption distance is large compared with the plasma dimensions, the plasma is said to be optically thin, and the spectrum of the radiation emitted is characteristic of the individual types of radiation—in general, a function of the electron temperature.

BASIC PLASMA PHYSICS

Plasma formation. Apart from solid-state plasmas, such as those in metallic crystals, plasmas do not usually occur naturally at the surface of the earth. For laboratory experiments and technological applications, therefore, plasmas must be produced artificially. Because the atoms of such alkalies as potassium, sodium, and cesium possess low ionization energies, plasmas may be produced from these by the direct application of heat at temperatures around 3,000° K. In most gases, however, before any significant degree of ionization is achieved, temperatures in the neighbourhood of 10,000° K are required. Because all substances melt at temperatures far below that level, so far no container can withstand an external application of the heat necessary; therefore, any heating must be supplied internally by applying an electric field to the gas to accelerate and scatter all of the free electrons, thereby heating the plasma. Because of their small energy exchange in elastic collisions, electrons can be raised to much higher temperatures than other particles. As long as the rate of loss of electrons exceeds their rate of production, however, a plasma cannot take form.

Artificial production of plasmas

For plasma formation a sufficiently high electric field must be applied, its exact value depending on geometry and the gas pressure. An equilibrium state is reached when the electrical power input is balanced by energy losses to the outside. The electric field may be set up via electrodes or by transformer action, which is induced by a changing magnetic field. Laboratory temperatures of about 10,000,000° K with electron densities of about 10^{19} per cubic metre have been achieved by the transformer method. Extremely high temperatures, but relatively low-density plasmas, have been produced by the separate injection of ions and electrons into a mirror system (a plasma device using a particular arrangement of magnetic fields for containment). Other methods have used the high temperatures that develop behind a wave that is moving much faster than sound to produce what is called a shock front; lasers have also been employed.

Effects of magnetic and electric field. *Individual particle description.* The behaviour of a plasma may be described at different levels. If collisions are relatively infre-

quent, it is useful to consider the motions of individual particles. A magnetic field exerts a force on a charged particle only if the particle is moving, the force being at right angles to both the direction of the field and the direction of particle motion. In a uniform magnetic field, (B), a charged particle, gyrates about a line of force with an angular frequency equal to the magnetic field strength times the ratio of the charge of a particle (q) to its mass (m), or, $\omega_c = Bq/m$, charges of different sign rotating in opposite directions; in this equation, ω_c is the rate of gyration, or the cyclotron frequency.

The radius of the orbit is called the Larmor radius (r_L) and is equal to the product of the particle mass and the component (v) of the particle velocity in a plane perpendicular to the magnetic field divided by the product of its charge and the magnetic field—*i.e.*, $r_L = mv/qB$. The centre of the orbit is called the guiding centre. The particle may also have a component of velocity parallel to the magnetic field and so traces out a helix when the motion of the guiding centre is uniform. If a uniform electric field (E) is applied at right angles to the direction of the magnetic field, the guiding centre drifts with a uniform velocity of magnitude equal to the ratio of the electric to the magnetic field (E/B), at right angles to both the electric and magnetic fields. Should the electric field alternate in time, the problem is even more complex. If, however, such an alternating electric field varies at the same frequency as the cyclotron frequency, the guiding centre will remain stationary and the particle will be forced to travel in an ever-expanding orbit. This phenomenon is called cyclotron resonance and is the basis of the cyclotron particle accelerator.

Cyclotron resonance

The motion of a particle about its guiding centre constitutes a circular current. As such, the motion produces a dipole magnetic field not unlike that produced by a simple bar magnet. Thus, a moving charge not only interacts with magnetic fields but also produces them. Therefore, interactions between particles caused by their motions must also be considered. The direction of the magnetic field produced by a moving particle, however, depends both on whether the particle is positively or negatively charged and on the direction of its motion. If the motion of the charged particles is completely random, the net associated magnetic field is zero. On the other hand, if charges of different sign have an average relative velocity (*i.e.*, if an electrical current flows) then a net magnetic field over and above any externally applied field exists. The magnetic interaction between charged particles is therefore of a collective, rather than of an individual, particle nature.

Statistical and fluid descriptions. At a higher level of description than that of the single particle, kinetic equations of the Boltzmann type are used. Such equations essentially describe the behaviour of those particles about a point in a small-volume element, the particle velocities lying within a small range about a given value. The interactions with all other velocity groups, volume elements, and any externally applied electric and magnetic fields are taken into account.

At a still higher level of description, equations of a fluid type may be derived from the kinetic equations; they express the conservation of mass, momentum, and energy per unit volume, with one such set of equations for each particle type. Total conservation equations may be obtained from these by a process of summation from the equations of each particle type. For instance, the derived thermal-energy equation states that the rate gain of thermal energy is equal to the work done by pressure forces plus heating from electrical currents less losses from radiation and heat conduction. The total heat flux —that is, the thermal energy transported across a unit area of the plasma in a unit of time—is the sum of each term for each particle type present.

Determination of plasma variables. The basic variables useful in the study of plasma are number densities, temperatures, electric and magnetic field strengths, and particle velocities. In the laboratory and in interplanetary space, both electrostatic (charged) and magnetic types of sensory devices called probes help determine the magnitudes of such variables. With the electrostatic probe, ion densities, electron and ion temperatures, and electrostatic potential differences can be determined. Small search coils and other types of magnetic probes yield values for the magnetic field; and from Maxwell's electromagnetic equations, the current and charge densities and the induced component of the electric field may be found.

Furthermore, in the laboratory, the absorption, scattering, and excitation of neutral and high-energy ion beams are helpful in determining electron temperatures and densities; in general, the refraction, reflection, absorption, scattering, and interference of electromagnetic waves also provide ways to determine these same variables. As an example, a powerful but difficult technique for such measurements is based on the scattering of a laser beam by plasma electrons. Aside from the above methods, much can be learned from the radiation generated and emitted by the plasma itself; in fact, this is the only means of studying cosmic plasma beyond the solar system. The various spectroscopic techniques covering the entire continuous radiation spectrum determine temperatures and identify such nonthermal sources as those producing synchrotron radiations.

Ion temperatures can be determined by measuring the widths of spectral lines on a spectrographic plate; their widths vary with thermal motions, a phenomenon called Doppler broadening. Velocities of plasmas in space are measured by Doppler shifts; that is, the spectral lines of radiation from receding bodies shift to lower frequencies, and vice versa. From the frequency of line radiation, researchers can identify elements, map the distribution of particles in plasmas, and determine their electron temperatures. Spectral lines are affected also by magnetic fields, and thus magneto-optical effects can determine such fields.

Doppler measurements of plasmas

Fluid-like behaviour. *Plasma waves.* An important parameter in plasma physics is the magnetic Reynolds number, the value of which depends on electrical resistivity of the plasma and the velocity of the particles. The Reynolds number is named for the British physicist Osborne Reynolds, who devised a measure of the critical velocity at which a flowing fluid changes its character. When the magnetic Reynolds number is much greater than one, resistance effects can be ignored and the magnetic lines of force are said to be frozen to (or to move with) the plasma. Because of this phenomenon certain interesting types of waves, called magnetohydrodynamic waves, occur at low frequencies. Generally, in a wave, the plasma particles oscillate about an equilibrium position, and their energy and momentum are transferred from one to another either by collisions or by interaction with electric and magnetic fields. Two fundamental types of wave motion can occur: longitudinal, like a sound wave, in which particle oscillation is in a direction parallel to the direction of wave propagation; and transverse, like a surface water wave, in which particle oscillation is in a plane perpendicular to the direction of wave propagation. In both types a wave may be characterized by a speed of propagation (u), a wavelength (λ), and a frequency (ν) related by an expression in which the velocity is equal to the product of the wavelength and frequency, namely, $u = \lambda\nu$. At low frequencies, on twanging or disturbing a line of force, a transverse wave called an Alfvén wave propagates parallel to the magnetic field with a velocity that depends on the particle density and the magnetic field strength. It is equal to the magnetic flux density (B) divided by the square root of the particle density (ρ) times the permeability of free space (μ_0); that is to say, $B/\sqrt{\mu_0\rho}$. A sonic wave that is an ordinary longitudinal sound wave can also propagate parallel to the magnetic field. Perpendicular to the magnetic field a different type of longitudinal wave called a magneto-sonic wave can occur.

Alfvén waves

In these magnetohydrodynamic transverse and longitudinal waves the plasma behaves as a whole, and the speeds are independent of wave frequency when the frequency is low; as in any fluid, excessive speeds can produce shock waves. At higher frequencies, however, the separate behaviour of ions and electrons causes the wave

velocities to vary with frequency. The Alfvén wave splits into two components, referred to as the fast and slow Alfvén waves, which propagate at different frequency dependent speeds. At still higher frequencies these two waves (called the electron cyclotron and ion cyclotron waves, respectively) cause electron and cyclotron resonances (synchronization) at the appropriate resonance frequencies. Beyond these resonances, transverse wave propagation does not occur until frequencies, called cutoffs, comparable with the plasma frequency, are reached, when the usual kind of electromagnetic wave may propagate.

For propagation of a transverse wave perpendicular to a magnetic field, two transverse modes of wave vibration exist. One corresponds to the electric field oscillating parallel to the magnetic field. This wave behaves as though there were no magnetic field present and is referred to as the ordinary wave. Propagation occurs only if the frequency is greater than the plasma frequency (ω_p); an electromagnetic wave with a lower frequency is reflected by the plasma. The other wave, in which the electric field is perpendicular to the magnetic field, is affected by the magnetic field and is called the extraordinary wave. It has several resonance and cutoff frequencies. At low frequencies it is a fast Alfvén wave, and at high frequencies it is a normal electromagnetic wave.

The speed of an ordinary longitudinal sonic wave at high frequencies also becomes frequency dependent and is referred to as the ion-acoustic wave, a resonance occurring at a frequency determined by electrostatic oscillations of the ions. Beyond this frequency no sonic wave propagates parallel to a magnetic field until the frequency reaches the plasma frequency, above which electroacoustic waves occur. The wavelength of these waves at the critical frequency (ω_p) is infinite, the electron behaviour at this frequency being the plasma oscillations of Langmuir and Tonks. Even without particle collisions, waves shorter than the Debye length are heavily damped; that is, their amplitude is rapidly decreased with time. This phenomenon, called Landau damping (after the Soviet physicist Lev Davidovich Landau), arises from the fact that some electrons always have the same velocity as the wave. As they move with the wave, they are accelerated and thus extract energy from the wave, damping it in the process. In the absence of truly energy-dissipative mechanisms, such as collisions, and in suitable circumstances, after the initial wave has damped away, an echo having related wave characteristics can be detected in the plasma.

Landau damping of wave amplitude

Containment and instabilities. Magnetic fields are used to contain high density, high temperature plasmas, such fields exerting pressures and tensions on the plasma. An equilibrium configuration is reached only when at all points in the plasma these pressures and tensions exactly balance the pressure from the motion of the particles. A well-known example of this is the pinch effect observed in specially designed equipment. If an external electric current is imposed on a cylindrically shaped plasma and flows in parallel to the plasma axis, the magnetic forces cause the plasma to constrict, or pinch. An equilibrium condition is reached for a temperature proportional to the square of the electric current. This phenomenon implies that any temperature may be achieved by making the electric current sufficiently large, the heating resulting from currents and compression. In practice, however, since no plasma can be infinitely long, serious energy losses occur at the ends of the cylinder; and also, major instabilities develop in such a simple configuration.

In general, there are two basic methods of eliminating or minimizing end losses from an artificially created plasma: the production of toroidal plasmas and the use of magnetic mirrors (see NUCLEAR FUSION). A toroidal plasma is essentially one in which a plasma of cylindrical cross section is bent in a circle so as to close on itself. For such plasmas to be in equilibrium and stable, however, requires special magnetic fields, the largest component of which is a circular field parallel to the axis of the plasma; any currents must be produced by transformer action.

A useful way of describing the confinement of a plasma by a magnetic field is by measuring containment time (τ_c), or the average time for a charged particle to diffuse out of the plasma; such time is different for each type of configuration. Various types of instabilities can occur in plasma. These lead to a loss of plasma and a catastrophic decrease in containment time. The most important of these instabilities is called magnetohydrodynamic. Although an equilibrium state may exist, it may not correspond to the lowest possible energy. The plasma, therefore, seeks a state of lower potential energy, just as a ball at rest on top of a hill (representing an equilibrium state) rolls down to the bottom if perturbed; the lower energy state of the plasma corresponds to a ball at the bottom of a valley. In seeking the lower energy state, turbulence develops, leading to enhanced diffusion, increased electrical resistivity, and large heat losses. In toroidal geometry, circular plasma currents must be kept below a critical value called the Kruskal–Shafranov limit, otherwise a particularly violent instability consisting of a series of kinks may occur.

Kruskal–Shafranov limit

Another class of instabilities is called drift wave, or universal. Spatial changes in density, temperature, or magnetic fields permit an abnormal number of particles with speeds equal to those of certain plasma waves. In an inverse manner to Landau damping, such waves grow by absorbing energy from the particles, and thus instabilities develop. A third basic type, dissipative instability, is associated with collisions between particles. Although a completely stable system appears to be virtually impossible, considerable progress has been made in devising systems that eliminate the major instabilities. Temperatures of the order of 10,000,000° K at densities of 10^{19} particles per cubic metre and containment times as high as 1/50 second have been achieved. The containment, however, is basically inefficient for practical use.

Applications. The most important practical applications of plasmas lie in the future, largely in the field of power production. The major method of generating electrical power has been to use heat sources to convert water to steam, which drives turbogenerators. Such heat sources depend on the combustion of fossilized fuels, such as coal, oil, and natural gas, and fission processes in nuclear reactors. A potential source of heat might be supplied by a fusion reactor, with a basic element of deuterium–tritium plasma; nuclear fusion collisions between those isotopes of hydrogen would release large amounts of energy to the kinetic energy of the reaction products (the neutrons and the nuclei of hydrogen and helium atoms). By absorbing those products in a surrounding medium, a powerful heat source could be created. To realize a net power output from such a generating station—allowing for plasma radiation and particle losses and for the somewhat inefficient conversion of heat to electricity—plasma temperatures of about 100,000,000° K and a product of particle density times containment time ($n\tau_c$) of about 10^{20} seconds per cubic metre are necessary. For example, at a density of 10^{20} particles per metre cubed, the containment time must be one second. Such figures have still to be reached, although there has been much progress.

In addition to its generation of power, a fusion reactor might desalinate seawater. Approximately two-thirds of the world's land surface is uninhabited, but only one-half of this because it is arid. The use of both giant fission and fusion reactors in the large-scale evaporation of seawater could make irrigation of such wastes economically feasible.

Another possibility in power production is to eliminate the heat–steam–mechanical energy chain. One suggestion depends on the dynamo effect. If a plasma moves perpendicular to a magnetic field, an electromotive force, according to Faraday's law, is generated in a direction perpendicular to both the direction of flow of the plasma and the magnetic field. This dynamo effect can drive a current in an external circuit connected to electrodes in the plasma, and thus electrical power may be produced without the need for steam-driven rotating machinery. This process is referred to as MHD (magnetohydrodynamic) power generation, and has been proposed as a method of extracting power from certain types of fission reactors.

Current from dynamo effect

The inverse of the dynamo effect, called the motor effect, may be used to accelerate plasma. By pulsing cusp-shaped magnetic fields in a plasma, for example, thrusts proportional to the square of the magnetic field can be achieved. Motors based on such a technique have been proposed for the propulsion of craft in deep space. They have the advantage of being capable of achieving large exhaust velocities, thus minimizing the amount of fuel that must be carried.

Provided the current is not too large, a glow discharge occurs between two electrodes at pressures of one thousandth of an atmosphere or thereabouts. Such glow discharges are responsible for the light given off by neon tubes and such other light sources as fluorescent lamps that operate by virtue of the plasmas that they produce in electric discharge. The degree of ionization in such plasmas is usually low, but electron densities of 10^{16} to 10^{18} electrons per cubic metre can be achieved with an electron temperature of 100,000° K. The electrons responsible for current flow are produced by ionization in a region near the cathode, most of the potential difference between the two electrodes occurring there. This region does not contain a plasma, but the region between it and the anode, or positive electrode, does.

Use in radio switching devices

Other applications of the glow discharge include electronic switching devices; and it and similar plasmas produced by radio frequency techniques can be used to provide ions for particle accelerators and act as generators of laser beams. In increasing the current through a glow discharge, a stage is reached when the energy generated at the cathode is sufficient to provide all the conduction electrons directly from the cathode surface, rather than from gas between the electrodes. Under this condition the large cathode potential difference disappears and the plasma column contracts. This new state of electric discharge is called an arc. Compared with the glow discharge, it is a high-density plasma and will operate over a large range of pressures. Arcs are used as light sources for welding, in electronic switching, for rectification of alternating currents, and in high-temperature chemistry. By running an arc between concentric electrodes and injecting gas into such a region, a hot, high-density plasma mixture is ejected, called a plasma jet; it has many chemical and metallurgical applications.

NATURAL PLASMAS

Extraterrestrial. It has been suggested that the universe originated as a violent explosion about 10,000,000,000 years ago and initially consisted of a fireball of completely ionized hydrogen plasma. Irrespective of the truth of this, there is little matter in the universe now that does not exist in the plasma state. The observed stars are composed of plasmas, as are interstellar and interplanetary media and the outer atmospheres of planets. Man's knowledge of the universe has come primarily from studies of electromagnetic radiation emitted by plasmas and transmitted through them; and, since the 1960s, from space probes within the solar system.

In a star the plasma is bound together by gravitational forces, and the enormous energy that it emits originates in thermonuclear fusion reactions within the interior. Heat is transferred from the interior to the exterior by radiation; in the outer layers, convection or circulation of the plasma is important. In the vicinity of a hot star the interstellar medium consists almost entirely of completely ionized hydrogen, ionized by the star's ultraviolet radiation. Such regions are referred to as H II regions. The greater proportion by far of interstellar medium, however, exists in the form of neutral hydrogen clouds, referred to as H I regions. Because the heavy atoms in such clouds are ionized by ultraviolet radiation (or photoionized), they also are considered to be plasmas, although degree of ionization is probably only one part in 10,000. Other components of the interstellar medium are grains of dust and cosmic rays, the latter consisting of very high-energy atomic nuclei completely stripped of electrons. The almost isotropic velocity distribution of the cosmic rays is believed to stem from interactions with waves of the background thermal plasma.

Throughout this universe of plasma there are magnetic fields. In interstellar space magnetic fields are about 5×10^{-6} gauss (a unit of magnetic field strength), in interplanetary space 5×10^{-5} gauss, whereas in intergalactic space they could be as low as 10^{-9} gauss. These values are exceedingly small compared with the Earth's surface field of about one-half gauss. Although small in an absolute sense, these fields are nevertheless gigantic, considering the scales involved. For example, to simulate interstellar phenomena in the laboratory, fields of about 10^{15} gauss would be necessary. Thus, these fields play a major role in nearly all astrophysical phenomena. On the Sun the average surface field is something like one to two gauss, but magnetic disturbances arise, such as sunspots, in which fields of between 10 gauss and 1,000 gauss occur. Many other stars are also known to have magnetic fields, whereas such fields as 10^{-3} gauss are associated with various extragalactic nebulae, synchrotron radiation being observed.

Magnetic fields in space

Solar–terrestrial. The visible region of the Sun is the photosphere (see Table), with its radiation about the same as the continuum radiation from a blackbody.

Various Natural Plasmas, Their Electron Densities and Temperatures

plasma	n_e (per cu m)	T_e (°K)
Sun		
Centre	10^{31}	1.5×10^7
Photosphere	10^{20}	4,200
Chromosphere	10^{17}–10^{20}	5×10^5
Corona	10^{13}	1.5×10^6
Solar wind (near Earth)	5×10^6	4×10^5
Interstellar space		
H II regions	10^6	10^4
H I regions	10^2	100–125
Intergalactic space	1	3 (?)
Earth		
Outer magnetosphere	10^7–10^6	10^4
Plasmasphere	10^{10}–10^9	10^4
Ionosphere	10^{11}–10^{12}	250–3,000
Metals	10^{28}	10^4

Lying above the photosphere, the chromosphere is characterized by the emission of line radiation from atoms and ions. Outside the chromosphere, the corona expands into the everblowing solar wind (see below), which on passing through the planetary system eventually encounters the interstellar magnetic field.

The Sun's corona is believed to be heated by magnetohydrodynamic or acoustic waves generated in the photosphere. The solar wind, a collisionless plasma made up primarily of electrons and protons, represents an outflow of matter moving at supersonic speed. The wind takes with it frozen-in extensions of the Sun's magnetic field. In the region of the Earth, the wind has an average speed of 400 kilometres (128.5 miles) per second; and on encountering the Earth's magnetic field, a shock front develops, the pressures acting to compress the field on the side toward the Sun and elongate it on the night side (in the Earth's lee away from the Sun). The Earth's magnetic field is, therefore, confined to a cavity called the magnetosphere, into which the direct entry of the solar wind is prohibited. This cavity extends for about ten Earth radii on the Sun's side and about 1,000 Earth radii on the night side.

The magnetosphere contains plasma. At about four Earth radii from the centre of the Earth, electron density markedly increases. Within this region, plasma particles are confined in closed orbits by the Earth's magnetic field, whereas those outside the region may drift to the edge of the magnetosphere. The inner region is called the plasmasphere. To within about 2,500 kilometres of the Earth's surface, the plasma throughout the magnetosphere consists of completely ionized hydrogen. In a somewhat ill-defined region below 2,500 kilometres and above 350 kilometres, oxygen and helium ions predominate. The degree of ionization decreases but the electron density increases with decreasing height. The degree of ionization at 350 kilometres is about 5×10^{-3} per cubic

metre, whereas the electron density reaches a maximum of about 10^{12} cubic metres. Below this level, the electron density slowly decreases, falling away rapidly beneath 100 kilometres.

The region between about 100 and 350 kilometres is normally referred to as the ionosphere, although this term has also been used to describe the entire magnetospheric plasma. In the ionosphere the dominant ions are those of atomic and molecular oxygen and nitric oxide. These are produced by photo-ionization of the upper atmosphere by solar ultraviolet radiation.

High-energy particles also exist in the magnetosphere. At about 1.5 and 3.5 Earth radii from the centre of the Earth, two regions contain high-energy protons. These regions are the Van Allen belts; the protons in them are trapped in mirror systems formed by the Earth's magnetic dipole field. In general, the inner regions of the plasmasphere are stable, but in the outer regions disturbances can lead to the rejection and ejection of particles. These precipitate into the upper atmosphere in the region of the poles, causing auroral displays (see IONOSPHERE).

Terrestrial. Plasma can exist also in the lowest regions of the Earth's atmosphere. A lightning stroke is actually an oxygen–nitrogen plasma heated to approximately 20,000° K, with an ionization of about 20 percent, similar to that of a laboratory arc. Although the bolt or stroke is only a few centimetres thick and lasts only a fraction of a second, tremendous energies are dissipated.

Lightning flashes and ball lightning

A lightning flash between ground and a cloud, on the average, consists of four such strokes. An associated phenomenon is ball lightning. There are authenticated reports of glowing, floating, stable balls of light several tens of centimetres in diameter occurring at times of intense electrical activity in the atmosphere. On contact with an object these balls release large amounts of energy. Although lightning balls are probably plasmas, so far no adequate explanation of them has been given.

Considering the origins of plasma physics and the fact that the universe is little more than a vast sea of plasma, it is ironic that the only naturally occurring plasmas at the surface of the earth besides lightning are those to be found in ordinary matter. The free electrons responsible for electrical conduction in a metal constitute a plasma. Ions are fixed in position at lattice points (crystal sites), and so plasma behaviour in metals is limited to such phenomena as plasma oscillations and electron cyclotron waves (called helicon waves) in which the electron component behaves separately from the ion component. In semiconductors (a special class of conductors), on the other hand, the current carriers are electrons and positive holes, the latter behaving in the material as free positive charges of finite mass. By proper preparation, the number of electrons and holes can be made approximately equal so that the full range of plasma behaviour can be observed.

The importance of magnetic fields in astrophysical phenomena has already been emphasized. It is believed that these fields are produced by self-generating dynamos, although the exact details are still not fully understood. On Earth, differential rotation in its liquid core causes the external magnetic dipole field (manifest as the North and South Poles). Fields inside the core are twisted into circular loops; cyclonic turbulence in the liquid, generated by heat conduction and Coriolis forces (apparent forces accompanying all rotating systems, including the heavenly bodies) then regenerate the original dipole field from these loops.

Similar processes are believed to occur both in the Sun and the Galaxy. In the Sun the circular internal magnetic field is observed by lines of force apparently breaking the solar surface to form exposed loops; entry and departure points are what are observed as sunspots. Although the exterior magnetic field of the Earth is that of a dipole, this is further modified by currents in both the ionosphere and magnetosphere. Lunar and solar tides in the ionosphere lead to motions across the Earth's field that produce currents, like a dynamo, which modify the initial field. Again, changes in the intensity of the solar wind induce or produce other currents in the magnetosphere; such currents constitute the essence of a magnetic storm (see EARTH, MAGNETIC FIELD OF).

BIBLIOGRAPHY. Basic introductory texts include: LYMAN SPITZER, *The Physics of Fully Ionized Gases*, 2nd rev. ed. (1962); and T.G. COWLING, *Magnetohydrodynamics* (1957). More advanced texts are J.L. DELCROIX, *Plasma Physics*, vol. 1 and 2 (1965–68); and WULF B. KUNKEL (ed.), *Plasma Physics in Theory and Application* (1966). Two excellent texts on experimental techniques are W. LOCHTE-HOLTGREVEN (ed.), *Plasma Diagnostics* (1968); and M.A. HEALD and C.B. WHARTON, *Plasma Diagnostics with Microwaves* (1965). A classic discussion of astrophysical problems is HANNES ALFVEN and CARL-GUNNE FALTHAMMAR, *Cosmical Electrodynamics*, 2nd ed. (1963). A classic text in the kinetic theory of ionized gases is SYDNEY CHAPMAN and T.G. COWLING, *The Mathematical Theory of Non-Uniform Gases*, 3rd ed. (1970). An advanced text, with many references, is V.L. GINZBURG, *The Propagation of Electromagnetic Waves in Plasmas* (trans. from the Russian by J.B. SYKES and R.J. TAYLER, 1964).

(B.S.L.)

Plastics and Resins

Plastics, in the modern meaning of the word, are synthetic materials that are capable of being formed into usable products by heating, milling, molding, and similar processes. The term is derived from the Greek *plastikos*, "to form." By some interpretations of the word, rubber and other natural products are said to be plastic, but the modern definition, especially in relation to industry, excludes natural rubber and also such other natural products as wood, leather, and metals. Resins are both natural and synthetic. Natural resins range from pitch and asafetida to frankincense, myrrh, and amber. Synthetic resins are the bases of an enormous range of products the growth of which shows no signs of abating.

The distinction between plastics and resins is at best arbitrary, since many of today's synthetic materials can properly be called both resins and plastics. Historically, it appears that the term resin was applied to those products primarily used as substitutes for the natural product in coating compositions, whereas the term plastic was applied to designate those compositions that involved a molding operation in their fabrication.

By custom, the fibre industry is considered to be outside the plastics industry, although the same raw materials—polyamides (nylon), cellulose, and cellulose acetate—are used by the two industries. For a detailed discussion of these substances and their use in the manufacture of fibres, see FIBRES, MAN-MADE.

Composition of plastics

The characteristic ability of most plastics to soften but not melt when heated, so that their shape can be changed under mechanical stress without losing cohesion, producing a rigid new form upon cooling, is a property resulting from their molecular structure. Plastics are composed of giant molecules, known as polymers, with immensely long chains of repeating units derived from short molecules; in the case of addition polymers the short molecules are known as monomers. The chains are built up by formation of chemical bonds. The assembly of polymeric chains is held together by a number of different types of forces: (a) London, Van der Waals, or dispersion forces, which are relatively weak; (b) for more polar materials, hydrogen bonding or dipole-dipole interactions can be involved and result, in general, in somewhat higher softening points; (c) the long polymer molecules are inevitably entangled with each other and considerable forces are required to disentangle them. When heated, these chains move apart sufficiently to permit them to slide over one another but retain their cohesiveness because of the forces still operating. By contrast, the components in crystals of small molecules move apart sufficiently to melt to mobile liquids.

This ability to slide into new positions, which are retained on cooling, is the basis of plastics fabricating processes, such as molding or extrusion of sheets or other forms from heated granules, vacuum forming of sheets into desired forms, blowing of extruded tubes into bottles, and other processes. At higher temperatures maximum mobility is attained, allowing injection molding (in which the plastic is injected into a hollow casting and

cooled) into an infinite variety of forms. These processes require only moderately high temperatures, usually in the range of 150–250° C (300–480° F), and are very versatile, enabling easy fabrication of a variety of forms. Plastics also have the advantages of toughness, lightness, and of lack of colour, giving rise to an enormous range of coloured products by the addition of dyes and pigments; some plastics are transparent.

Intermeshing of the long-chain molecules of the polymers of which plastics are composed is responsible for their mechanical toughness and ductility, in contrast to the brittleness of low molecular weight organic substances, whether glasslike or crystalline. The hardness and ductility of polymers may be modified by the choice of the constituents that make up the chains. The intensity of the molecular attraction between neighbouring chains and flexibility within the chains (*i.e.*, freedom or restriction of rotation about main chain bonds) are important factors.

All these polymers are known as thermoplastics. Some long-chain materials known as thermosets can also be fabricated; on heating or with chemical treatment during fabrication, the chains of thermosets link with one another to produce a network structure that prevents softening on reheating. Phenol–formaldehyde is a typical thermoset.

This article is divided into the following sections:

I. History of the plastics industry
II. Types of plastic materials: their formation and usage
  Polymerization: the basic reaction
  Thermosetting resins
  Thermoplastic resins
  Other types of resins
III. Plastics manufacture and fabrication
  Natural raw materials
  Chemicals and additives
  Basic techniques for processing plastics
  Fabrication and finishing
IV. Industrial plastics
  Structural foams
  Sheets and films
  Laminates, reinforced plastics, and composites
  Miscellaneous industrial plastics

I. History of the plastics industry

The first plastic was Parkesine, later called Xylonite, exhibited by the English chemist and inventor Alexander Parkes in 1862. Essentially this substance was nitrocellulose softened by vegetable oils and a little camphor. It was John W. Hyatt in the United States who recognized in 1869 the vital plasticizing effect of camphor and named the product Celluloid. Though primarily regarded as a substitute for ivory and tortoiseshell, Celluloid, despite its inflammability, found substantial early uses as carriage and automobile windshields and motion picture film.

Celluloid

The first completely synthetic plastic was commercially produced from phenol and formaldehyde in 1910 by Leo Hendrik Baekeland, a Belgian-born American chemist. A dark thermosetting resin, later called Bakelite, this substance was used in the electrical industry. The first colourless resin, from urea–formaldehyde, though patented in 1918, was not marketed until 1928. Though available in any colour, it was moisture adsorbent (that is, moisture clung to its surface); this problem was overcome in 1939 by using melamine instead of urea.

Meanwhile, cellulose acetates received intensive study. At first, only the triacetate, soluble in toxic solvents, was produced and applied as a waterproof varnish to the fabric on airplane wings; at the end of World War I, the triacetate was manufactured in powdered form for later melting and molding.

During the same period, transparent acetate film was developed and used for packaging. With development of methylene chloride manufacture, triacetate film came into use as a photographic base in 1935.

Three principal factors contributed to the immense progress in plastics that took place after 1920. Of these, the work of the German chemist Hermann Staudinger is outstanding. In May 1922 he published a paper that demonstrated that rubber was composed of a chain of isoprene units, and the word macromolecule first appeared. His views aroused much opposition and this led him to an intensive study of styrene resin. By examination of the viscosities of its solutions, Staudinger showed that styrene resin was an assembly of macromolecules or giant polymers of different lengths that he called polystyrene. Staudinger's discovery of the theoretical background of polymerization coincided with two other factors: the growth of the chemical industry in general and a political impetus in Germany toward independence from many natural materials, leading to the industrial development of processes long known in the laboratory.

Vinyl chloride, discovered in 1835, is formed by the reaction of acetylene with hydrochloric acid. The polymer polyvinyl chloride (PVC) was first produced in 1912. There was a similar background to polymethyl methacrylate, starting with acrylates, which were polymerized in the laboratory in 1880 and came into production in 1927. A great deal of this work was done in Germany, including large-scale production of polystyrene.

Plastics research and manufacture was proceeding on a considerable scale in the United States as well. Outstanding was the study of polymers in the laboratory of E.I. du Pont de Nemours & Company from 1928 onward, which led to the "superpolyamide" now known as nylon.

The next great step forward was the British invention of polyethylene, which came into commercial production in 1939. Originally of unique value in radar, it is now used in enormous quantities in the form of film and in products made by injection-molding techniques. Further developments included polyurethanes in 1937, epoxy resins in 1943, polyethylene terephthalate in 1941, and polycarbonates in 1956.

Invention of polyethylene

Along with the work on polymers, there was a corresponding development of special machinery for plastics fabrication. When Celluloid was invented, the only equipment available for possible fabrication was that used in the rubber industry, such as the crude hydraulically operated "stuffing" machines for extrusion. The first patent for an extrusion machine employing an Archimedean screw was dated 1879; as thermoplastics were developed, more of these machines were used until one was designed especially for thermoplastics in 1935. Later, extrusion machines were produced for manufacture of plastic sheets and film and for blown bottles. The injection-molding machine was introduced in 1921. The first automatic machine was produced in 1937; later, an improved feed mechanism was added.

II. Types of plastic materials: their formation and usage

POLYMERIZATION: THE BASIC REACTION

Principles. The manufacture of plastics depends upon the building of chains and networks either by condensation polymerization or by addition polymerization.

A condensation polymer is formed by a synthesis that involves the stepwise reaction of reactive molecules with one another, with the elimination of small molecules such as water. The reaction gradually slows down as polymers are built up.

An addition polymer forms chains by the linking of small identical units (which may be diversified to a degree) without elimination of small molecules.

The most important concept in condensation polymers is that of "functionality" (f)—*i.e.*, the number of reactive groups in each molecule participating in the chain build-up. Each molecule must have at least two reactive groups, of which hydroxyl ($-OH$), acidic endings ($-COOH$), and amine endings ($-NH_2$) are the simplest.

Hydroxyl is characteristic of alcohol endings, combining with an acid ending to give an ester, the polymer being known as a polyester. Examples are polyethylene terephthalate (by reaction of ethylene glycol containing hydroxyl groups at each end and terephthalic acid containing two acidic groups) and polycarbonate resins. (For a detailed description of polymerization, see POLYMERS.)

Alcohols are a particular class of oxygen-containing chemical compounds with a structure analogous to ethyl alcohol (C_2H_5OH). Amines are various compounds derived from ammonia by replacement of hydrogen by one

or more hydrocarbon radicals (molecular groups that act as a unit). Esters are compounds formed by the reaction between an acid and an alcohol or phenol with the elimination of water.

Processes. Bulk addition polymerization of pure monomers is mainly confined to styrene and methyl methacrylate. The process is highly exothermic, or heat producing. The dissipation of heat (necessary to maintain chain length) is achieved in the case of styrene by intensive stirring of the viscous, partially polymerized mixture, which is then passed down a tower through zones of increasing temperature. Alternatively, polymerization may be completed in containers or molds that are small enough to avoid an excessive temperature rise as a result of the heat released during polymerization.

Methyl methacrylate is also partially polymerized before pouring into molds formed between sheets of plate glass, to produce clear acrylic sheet.

Ethylene polymerization

Ethylene is polymerized in tubular reactors about 100 feet (30 metres) long and less than one inch (25 millimetres) in diameter at pressures of 600–3,000 atmospheres (one atmosphere equals the pressure of air at sea level) to give 10–20 percent conversion to low-density polyethylene. Residual gas is recycled.

Solution polymerization. Polymerization of monomers in solution gives easy temperature control but reduces the molecular weight of polymers formed as a result of chain transfer reactions. (Such reactions involve growing polymer chains reacting with small solvent molecules to give inactive polymers and starting a new chain, rather than allowing the original chain to grow to the high molecular weight which would have been obtained in the absence of such transfer.) Solvent removal from such a solution may also be very difficult. The process can be applied particularly to vinyl acetate and acrylic esters.

Solution polymerization. Polymerization of monomers tion producing beads of plastic is extensively applied to styrene, methyl methacrylate, vinyl chloride, and vinyl acetate. The monomer, in which the initiator or catalyst must be soluble, is maintained in droplet form suspended in water by agitation in the presence of a stabilizer such as gelatin, each monomer droplet behaving as a small bulk polymerization.

Emulsion polymerization. In emulsion polymerization the monomer is dispersed in water by means of a surface-active agent (a substance that reduces the surface tension of a liquid), slightly soluble in water, its bulk aggregating into tiny particles held in suspension (micelles). The monomer enters the hydrocarbon part of the surface-active micelles and is polymerized there by a water-soluble catalyst. This process is particularly useful for the preparation of very high molecular weight polymers.

Radiation polymerization. This process operates by molecular excitation leading to ejection of an electron. It can be used for manufacture of such thermoplastics as polyethylene and polyvinyl chloride.

THERMOSETTING RESINS

These include the phenolics, furan resins, aminoplastics, alkyds, allyls, epoxy resins, polyurethanes, polyesters, and silicones.

Phenolics or phenol–aldehydes. The most important commercial phenolic resin is based on phenol and formaldehyde. The two processes in general use are the one-step process producing resol resins (the first stage in the formation of a phenolic resin) that are either liquid or brittle, soluble, fusible solids, from more than one molecule of formaldehyde per phenol molecule; and the two-step process, using an excess of phenol to produce novolacs, resins that have no reactive methylol groups and must be mixed with an aldehyde to undergo further reaction.

Resol resins thermoset on heating and are used for adhesives. Novolacs require a further source of formaldehyde in the form of hexamethylenetetramine to produce molding powders. Both resins are run out from the reaction vessel, after removal of water by distillation, and ground up, then compounded on heated rolls with fillers that vary from wood flour to mica; for strength and heat resistance fibrous asbestos is used as a filler (hexamethylenetetramine is also added at this stage in the case of the two-step resin). Final grinding produces the molding powders, which on further heat treatment will yield the typical thermoset resin.

Characteristics of phenolic moldings

Phenolic moldings are resistant to heat, chemicals, and moisture and are preferred for wet-dry applications as in washing machines. Their high heat resistance and low heat conductivity recommend them for appliance parts, and their high electrical insulation for electric fittings such as switches, plugs, and distributor caps; resistance to hydraulic fluids has led to their use in automotive parts, and all these applications have been substantially increased by the development of injection molding and extrusion methods not hitherto applied to thermosets. Complex phenols from; *e.g.*, cashew-nut shell liquid are used in manufacture of brake linings.

Furan resins. Furfural is a five-membered ring compound (*i.e.*, the basic molecule has a ring shape and contains five atoms) of four carbon atoms and one oxygen atom, carrying the aldehyde group, – CHO; it reacts like formaldehyde with phenols in the presence of an acid catalyst to give a rigid polymer with high chemical resistance, used for coatings in industry. It can be prepared in semiliquid form with a low viscosity and remarkable penetrating power when applied to porous forms such as foundry sand cores or graphite blocks, being in this respect superior to other liquid resins.

Aminoplastics. Urea resins are made by the condensation in aqueous solution of formaldehyde and urea in the presence of ammonia as an alkaline catalyst, giving a colourless syrupy solution to which cellulose filler is added to yield a molding powder upon drying, which when heated in a mold gives a water-white (transparent) molding unless previously coloured by pigment.

The filler confers considerable strength, so that thin sections such as in cups and tumblers can be molded. Very large quantities of urea–formaldehyde resin are used in closures, kitchen and bathroom hardware details, and electric appliance housings and fittings.

Melamine behaves in the same way as urea, but the product is more moisture resistant, harder, and stronger, leading to wide use for plates and food containers. Melamine moldings are glossy and harder than any other plastic and retain a dust-free surface. Solutions of the thermoplastic stages of urea–formaldehyde resins are widely used as bonding agents for plywood and wood-fibre products.

Alkyds. Alkyds are polyesters, generally of phthalic acid with two acid endings and glycerol, a triol—*i.e.*, an alcohol with three hydroxyl groups. The solid resins are molded at high speed under low pressure, cured quickly, and are used where insulating properties, strength, and dimensional stability over a wide range of voltage, frequency, temperature, and humidity are required, as in vacuum-tube bases and automotive ignition parts and with glass-fibre reinforcement for switchgear and portable tool housings.

DAP and DAIP resins

Polyesters of unsaturated alcohols. Known as DAP and DAIP, these resins are esters of phthalic and isophthalic acid, respectively. They are notable for maintaining rigidity and excellent electrical properties at temperatures up to 230° C (450° F), properties also manifested by allylic resin-impregnated glass cloth, used in aircraft and missile parts. Other advantages are good storage life and absence of gas evolution during polymerization.

The allyl resin allyl diglycol carbonate, optically clear and colourless, is used for making cast objects; fully cured castings are more heat and abrasion resistant than other cast resins.

Epoxy resins. These resins have outstanding mechanical and electrical properties, dimensional stability, resistance to heat and chemicals, and adhesion to other materials. They are used for casting, potting, encapsulation, protective coatings, and adhesives, and for reinforced moldings and laminates of the highest quality. Popular adhesives (epoxy glues) contain the resin components and the curing agent, usually an amine or an anhydride, in separate packages. The two are mixed just before use.

Polyurethanes. Formed by the reaction between diisocyanates and polyols (multihydroxy compounds), polyurethanes are among the most versatile of plastics, ranging from rigid to elastic forms. Their major use is for foams, with properties varying from good flexibility to high rigidity. Thermoplastic polyurethanes that can be extruded as sheet and film of extreme toughness can also be made.

Polyesters of unsaturated acids. Certain esters can be polymerized to an infusible resin and are used on a very large scale in glass-fibre-reinforced plastics. Unsaturated acid (usually maleic acid in the form of its anhydride) is first polymerized to a relatively short polymer chain by condensation with a dihydric alcohol such as propylene glycol, the chain length being determined by the relative quantities of the two ingredients. The resulting syrupy condensation polymer is then diluted with a monomer such as styrene and an initiator for addition polymerization added. This mixture is quite stable at room temperature over a long period. Frequently, a silicone compound is added to promote adhesion to glass fibres, and wax to protect the surface from oxygen inhibition of polymerization. Glass-fibre materials are impregnated with the syrup and polymerization is brought about by raising the temperature. Alternatively, the polymerization can be carried out at room temperature by addition of a polymerization accelerator to the syrup immediately before impregnation. After an induction period, which can be controlled, polymerization takes place, with rapid increase in temperature, to give a glass-fibre-reinforced cross-linked polymer, which is effectively a thermoset type of plastic and very resistant to heat. The properties of the resin are frequently varied by replacing part of the unsaturated maleic anhydride by anhydrides of saturated acids such as sebacic acid.

Silicones. Silicon, unlike carbon, does not form double bonds or long silicon chains. It does, however, form long chains with oxygen such as in siloxanes with hydrocarbon groups attached to the silicon; these result in a wide range of oils, greases, and rubbers. Produced through a series of reactions involving replacement of certain atoms in the chain, silicon resins, or silicones, can be used for high- and low-pressure lamination, with glass-fibre reinforcement and with mineral or short glass-fibre fillers, or for molding powders. The outstanding characteristic of these products is high dielectric strength (that is, they are good insulators at high voltages) with low dissipation over a wide temperature and humidity range. Silicones are not distorted by heat up to 400° C (752° F). They are also physiologically inert and therefore valuable for prostheses (artificial body parts).

THERMOPLASTIC RESINS

These include cellulose derivatives, addition polymers, and condensation polymers.

Cellulose derivatives. Cellulose is a naturally recurring chain of units, each of which carries three hydroxyl groups, which can be esterified by acids. Cellulose nitrate is formed by treating cellulose with a mixture of strong sulfuric and nitric acids; after a purification process, mixing with alcohol and camphor yields a plastic mass that is rolled into sheets. These are pressed into a block, sliced to the required thickness, and then dried and polished by hot pressing against polished metal sheets. The sheets made by slicing still contain alcohol and can be cut up and repressed in various ways to give patterned material or, if cut into dice, can be extruded as plain sheet. Nitrocellulose cannot be injection molded because of its inflammability. Otherwise it is tough and moisture resistant and is used for eyeglass frames and table tennis balls.

Cellulose acetates

If cellulose is pretreated with acetic acid, followed by more acid and its anhydride, acetate esters are formed, giving cellulose triacetate. Complete acetylation would produce an acetic acid content of 62.5 percent; in practice about 61 percent is usual, giving a moisture-resistant, dimensionally stable material with a crystalline structure, soluble in a limited range of solvents, preferably methylene chloride with 10 percent methanol. It cannot be worked into a plastic with plasticizers, but plasticizers can be incorporated in a 15–16 percent solution from which film is cast as a photographic base. This is the type of solution referred to as "dope" in Figure 5.

Further treatment of the triacetate in solution in the presence of sulfuric acid splits off some acetic acid giving diacetate, soluble in acetone, and compatible with a range of plasticizers that can be incorporated in a rugged type of mixer without solvent to yield molding powders especially suited to injection molding.

Tool handles are a common use of cellulose acetate. It can also be treated by the slicing process described above, but its largest use is as packaging film, made either by extrusion or by a casting process, which gives a superior product.

If butyric or propionic acids are included in the esterifying mixture, acetate butyrates and acetate propionates are formed. These require less plasticizer and are tougher and more moisture resistant. An important use is for automobile steering wheels.

Ethyl cellulose is formed from cellulose and ethyl chloride. Very tough and requiring little plasticization, it is used for athletic helmets, flashlight cases, and fire-extinguisher components. It is also compatible with a very wide range of plasticizers and waxes and, when mixed with wax, can be used for protective coatings.

Low-density and high-density polyethylene

Addition polymers. Polyethylene is made in two principal qualities: low density, polymerized at pressures from 9,000 to 45,000 pounds per square inch (630 to 3,150 kilograms per square centimetre); and high density, made with special catalysts at 250 to 500 pounds per square inch (18 to 35 kilograms per square centimetre).

Low-density polyethylene has a 20° C (68° F) lower softening point than high-density polyethylene, which is more crystalline and consequently stiffer. The rigidity characteristics and surface of high-density polyethylene are comparable with polystyrene. It resembles nylon in feel, has a bursting strength three times that of low-density polyethylene, and withstands repeated exposure to 120° C (250° F) so that it can be sterilized.

Low-density polyethylene is principally used in the form of film for bags of all sizes, with thickness ranging from 0.00025 inch (0.00635 millimetre) to 0.007 inch (0.178 millimetre), for agricultural mulch, building-site protection, lamination, all kinds of packaging, and electrical-cable insulation.

Because of its stiffness, high-density polyethylene is used for large drums, gasoline tanks, bottle crates, domestic appliances that can be injection molded to large sizes, and squeeze bottles. It can also be produced in the form of foam.

Polyethylene can be chlorinated either in solution in carbon tetrachloride or in suspension, both in the presence of a catalyst. Below 55–60 percent chlorine it is more stable than polyvinyl chloride and is compatible with many polymers, especially polyvinyl chloride, to which it gives increased impact strength.

The low pressure process can be used to copolymerize polyethylene with propylene and butylene to increase resistance to stress cracking.

Copolymerization with vinyl acetate at high pressures increases flexibility, resistance to stress cracking, and also sealability, which is especially valuable in the food industry.

Ionomers are ethylene polymers in which carboxyl groups are located along the ethylene chain. These acid groups are converted into salts of sodium, potassium, magnesium, or zinc; this stiffens and toughens the plastic without destroying its thermoplastic character. As a result, unplasticized sheet has drawing characteristics superior to those of any other resin.

Ionomers are highly transparent, particularly resistant to organic solvents, and maintain toughness at low temperatures. They are used in composite films with polyethylene and nylon because of their puncture resistance; they have strong natural adhesion to aluminum. Their main applications are in food packaging and as injection moldings in the shoe industry.

Polypropylene. Polypropylene is made by the polymerization of propylene. Stronger and more rigid than poly-

ethylene, polypropylene has a harder surface, softens 30° to 40° C (54° to 72° F) higher, and is the lightest common plastic. It is more prone to oxidation than polyethylene unless appropriate stabilizers and antioxidants are added.

Injection molding affords its greatest outlet over a very wide range of applications including beer crates, luggage, and automotive panels and trims. An outstanding use is for integral hinges, because of its fatigue resistance. (An integral hinge is part of an object as it is manufactured, in contrast to most hinges, which are fastened to the object later.)

Polypropylene is being used in increasing quantities for fibres and for film. The latter is extruded and develops its greatest strength when oriented by stretching biaxially—that is, along two axes at right angles. Film clarity is obtained by minimizing the size of crystals and by various additive agents. It is made in thicknesses down to 0.0002 inch (0.005 millimetre). Copolymers of propylene with minor amounts of ethylene are also used for making film.

Methylpentene polymer. Methylpentene polymer is the lightest plastic, has a high crystalline melting point of 240° C (464° F) retaining form up to 200° C (392° F), has a light transparency of 92 percent, and electrical properties similar to polytetrafluoroethylene (described later). Its impact strength is greater than that of polystyrene and polymethyl methacrylate. It is also resistant to alkalies, weak acids, and solvents (except chlorinated solvents) and is excellent for laboratory ware and medical equipment (because it can be sterilized) made by injection molding, and for food packaging and catering, since food can be cooked in methylpentene polymer containers in conventional ovens, providing a nonstick surface resistant to hot fat.

Vinyls. Vinyl chloride is made from ethylene and chlorine, though acetylene can also be used (see Figure 1). The polymer (called polyvinyl chloride) is mainly processed in a highly plasticized form, with varying degrees

Drawing by D. Meighan

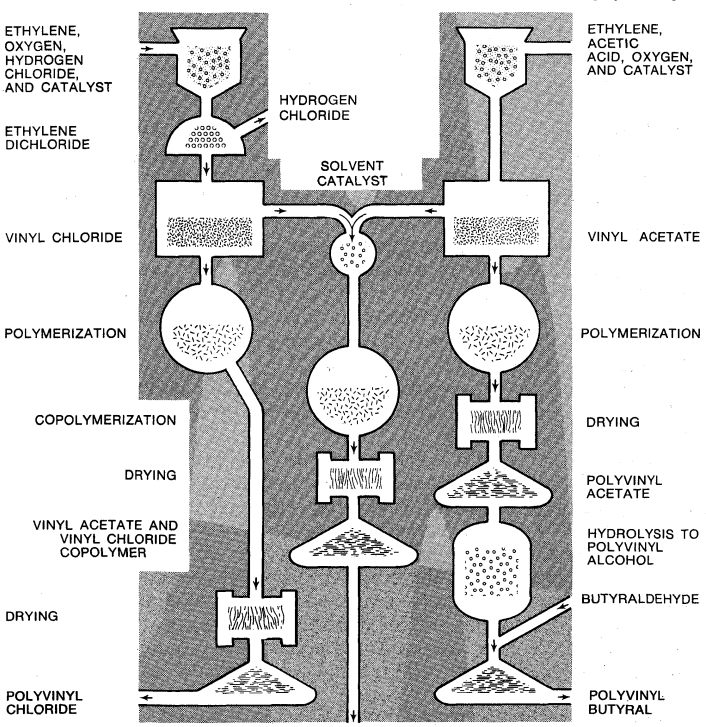


Figure 1: Principal steps in the production of vinyl resins.

of flexibility, by calendering (pressing between heavy rollers), extrusion, and molding, often from "dry blends," mixtures made below fusion temperature from polymer, plasticizer, and pigments. Plasticizers are chosen to maintain flexibility at low temperatures; the most commonly used plasticizer is dioctyl phthalate, but adipates may be used for special resistance to temperature change.

The range of applications of flexible polyvinyl chloride is enormous, covering clothing, flooring, wire insulation, home furnishings, piping, toys, and steel sheet covering. Extrusion, calendering, and molding of rigid polyvinyl chloride (*i.e.*, without plasticizer) is only possible with additions of stabilizers to prevent decomposition. Such stabilizers are mainly metallic (tin or cadmium) salts aided by epoxy resins, which also act as plasticizers, and lubricants such as stearates. Processing also may be aided by blending with such polymers as styrene acrylonitrile, polymethyl methacrylate, and chlorinated polyethylene.

Rigid polyvinyl chloride applications include bottle blowing, ducting in chemical plants, rainwater goods, piping for water and drainage mains, and light roofing. It has one of the highest combinations of stiffness and impact strength.

Plastisols and organosols

Plastisols and organosols are made by dispersing a very fine emulsion of polymerized resin in a plasticizer, usually in excess of the amount in flexible polyvinyl chloride (about 50 percent), together with stabilizers and gellants such as a calcium soap. Solvents and diluents are also used, and if they amount to more than 10 percent of the compound the product is known as an organosol.

The formulation can be difficult, as the flow characteristics can be varied greatly by choice by plasticizer and pigment (if used) and must be varied to suit the end use. Products are made by spreading, as on fabric, by dipping or by filling hot molds and pouring off excess material after gelling the outer layer at about 170° C (338° F). Plastisols can also be cast in molds; when maximum rigidity is required, diluents are used to permit lower plasticizer content. Plastisols can also be foamed.

Copolymers of vinyl chloride containing 5 to 40 percent vinyl acetate are made by the inclusion of vinyl acetate in the polymerization process; they have lower softening points and flow more easily than polyvinyl chloride at temperatures below decomposition points. They are thus suitable for making phonograph records, rigid clear sheeting, and molding powders. They are soluble in ketones, such as acetone, and certain esters, and they can be cast as film from solutions.

Copolymerization with vinylidene chloride, in quite small quantities, has the same effect on flow as vinyl-acetate and also increases the tensile strength of the plastic.

The chlorination of polyvinyl chloride, originally done in solution, is now carried out in a water–chloroform suspension. Extruded tube can be used in soil and waste systems and for central heating since its maximum service temperature is 100–105° C (212–221° F).

Vinylidene chloride can be made directly from ethylene and chlorine or by the further chlorination of vinyl chloride with subsequent removal of hydrochloric acid by alkali treatment. Copolymers of vinylidene chloride with minor proportions of vinyl chloride or acrylonitrile are used for making synthetic fibres, coating resins, and extruded thin films. The outstanding property of vinylidene chloride polymers is their low permeability for water vapour and gases. Coatings of the polymer are used on cellophane and other films to make them impervious to moisture.

Vinyl acetate is polymerized to a tough resin (polyvinyl acetate) by addition polymerization in solution to give an adhesive with a very high degree of tack (stickiness) and consequently of good adhesion; its main use, however, is as a basis of emulsion paints, in which case it is emulsion polymerized and not separated from the reaction mixture. Its use as a copolymer with ethylene and vinyl chloride has already been mentioned.

Polyvinyl alcohol is made by the saponification of polyvinyl acetate—*i.e.*, the removal of the acetic acid by alkali. In practice, a small amount of combined acetic acid is left in the molecule. The product is a water-soluble film-forming material used in dental molding and in the form of fibres for reinforcement (*e.g.*, of polyester), notably for injection-molded rail-tie retainers. If reacted with butyraldehyde, it gives polyvinyl butyral, an extremely tough but soft plastic used in sandwich-type safety glass.

Acrylics. Acrylic is a generic name applied to derivatives of acrylic acid, of which methyl methacrylate is

the most important. The process of polymerization is directed toward control of chain length, which for cast resins is from 1,800 to 3,000 monomer units; for molding powders only about 500 units are required to facilitate flow into a mold. A small amount of plasticizer such as dibutyl phthalate may be added before bulk polymerization to assist in deep molding.

Properties of transparency

The outstanding property of polymethyl methacrylate is its light transmission of 92 percent and water-whiteness (transparency), a property maintained over years of exposure to weather. It is also resistant to ultraviolet radiation from fluorescent lamps and is readily fashioned by deep drawing (*e.g.*, for airplane canopies) and by injection molding (*e.g.*, for light fittings and automotive details, including signal lights).

A most successful use is for internally illuminated signs. Any colour can be introduced, and structural uses that take advantage of its transparency and toughness are numerous and varied. Its dimensional stability recommends it for many optical uses.

Sheets can be cast in cells made from two sheets of glass plates; the cells are filled with a viscous solution of partially polymerized material. Other shapes are cast by the use of appropriate cell forms. Sheet is also made by an extrusion process, the emerging sheet being drawn through a calender with a highly polished chromium finish heated to 100° C (212° F). Acrylic dentures are made by stirring the powdered polymer into methyl methacrylate monomer containing a catalyst, and completing the polymerization in a mold formed from a dental impression.

Acrylics are also modified by copolymerization to improve impact strength, though they lose their extreme transparency and are usually pigmented, as for piano keys. An example is acrylic-modified polyvinyl chloride sheet, which is tougher than acrylonitrile–butadiene–styrene and polycarbonate and is suitable for corrosion-resistant parts, aircraft components, and materials-handling equipment.

Fluorine-carbon resins. Polytetrafluoroethylene and fluorinated ethylene–propylene are the only resins composed wholly of fluorine and carbon. The monomers are polymerized in water either to a granular solid or a dispersion. In the polymers the fluorine atoms surround the carbon chain as a sheath, giving a chemically inert and relatively dense product with immensely strong carbon–fluorine bonds. Polytetrafluoroethylene cannot be molded in the ordinary way but must be preformed at a very high pressure, followed by a sintering (heating under pressure) process and, in certain cases, by further processing to obtain an accurate form. Fluorinated ethylene–propylene can, however, be injection molded and extruded as thin film. Both plastics have exceptional heat resistance and very low friction coefficients. This leads them to be used for such mechanical applications as seal rings, nonlubricated bearings, and packings.

Resins for electrical use

For electrical applications, polytetrafluoroethylene and, to a slightly less degree, fluorinated ethylene–propylene resins are exceptional in that electrical and other properties are not adversely affected by long exposure to high temperatures. Polytetrafluoroethylene can be used in temperatures up to 260° C (500° F), and neither it nor fluorinated ethylene–propylene shows any embrittlement; they are also flexible at low temperatures and are standard insulation in aircraft and spacecraft. Their excellent electrical insulating properties lead to their use in very high frequency, ultrahigh-frequency, and other electronic applications. Polytetrafluoroethylene is also available in emulsion form to give adherent coatings to metals. Its complete inertness and low friction coefficient lead to its use in surgical prostheses, especially in joints.

Chlorotrifluoroethylene is a thermoplastic resin produced in varying formulations from combinations of fluorine-containing monomers. It has inherent flexibility, toughness, and clarity, differing in this last respect from the waxlike polytetrafluoroethylene. It is chemically resistant to all inorganic corrosive liquids and to most organic solvents, has zero moisture absorption, and a very wide range of service temperatures.

Its flexibility, combined with high insulating and thermal properties and facilitated by its extrusion properties, recommends it for wire and cable covering; its thermoplasticity, which can be augmented by plasticizer addition, or copolymerization with vinylidene fluoride, makes it an excellent material for injection molding of valve parts, closures for corrosive liquids, and for medical supplies.

Available with surfaces prepared for adhesion, polyvinyl fluoride transparent extruded film is unequalled for weatherability; it is used for external surfaces as a laminate. Polyvinylidene fluoride, a transparent injection-moldable plastic stronger than other fluoroplastics, is slightly less inert but particularly resistant to oxidizing agents found in the chemical industry.

Styrenes. Styrene is polymerized to a clear transparent water-white resin called polystyrene on a very large scale by a mass-block or continuous-mass process or by suspension polymerization in a water medium. In the mass-block method, polymerization takes place in a system of cells alternately filled with monomer and a cooling medium; continuous polymerization takes place in a tower, with cooling zones and a heated base from which the polymer is extruded. Polystyrene is an ideal injection material, hard, brilliantly transparent, and a good insulator; but it softens in boiling water, is brittle, and burns readily.

In very thin moldings for food containers, the polymer becomes oriented by flow, which toughens it; extruded film can be vacuum formed to give containers suitable for fresh food because of their high gas permeability and good water-vapour transmission. Polystyrene is readily foamed by blowing agents, flame retardants being added to the polymer.

High-impact polystyrene is made by dissolving styrene–butadiene rubber in styrene and then polymerizing the styrene. Polymerization of the mixture may be carried out by bulk or suspension processes.

Acrylonitrile–butadiene–styrene polymers are essentially similar in physical structure but with acrylonitrile added in the polymerization to harden the polymer. Minute rubber particles diminish clarity but act as stress-relief centres, making acrylonitrile–butadiene–styrene in particular an outstanding polymer, especially for large shapes (luggage, car bodies, etc.). Special grades are also made for automotive parts (softening point above 100° C, or 212° F) and for chrome plating. It is regarded as a good engineering plastic, can be foamed, injection molded, blown, and alloyed with other plastics.

Polymer for large shapes

Other addition polymers. Formaldehyde polymerizes readily to form paraformaldehyde. Under controlled conditions, a stable polymer of high molecular weight may be obtained. Polyformaldehyde is one of the strongest and stiffest thermoplastics, having a tensile strength of 10,000 pounds per square inch (700 kilograms per square centimetre). It can be injection molded and extruded. Exceptional fatigue endurance has led to its employment in automotive safety steering columns. Its impact strength and elastic recovery make it useful for snap-fit assemblies and springs, and its resistance to moisture and high abrasion and frictional resistance are advantageous in engineering uses generally. It also has good electrical properties over a wide temperature range.

Made mainly from trioxane (by condensation of formaldehyde) and ethylene oxide, acetal copolymers have properties resembling those of polyformaldehyde but in addition are notable for their stiffness at high temperatures, making them suitable for carburetor and automobile parts and for plumbing items. When glass-fibre-filled, they have exceptional resistance to creep (slow deformation over time when placed under stress).

A crystalline polymer, chlorinated polyether has a very low volume change from a melt to a solid state, permitting injection molding to very close tolerances. It has very high corrosion resistance to all reagents except oxidizing agents and is a valuable engineering plastic for valves, threaded socket connections, and precision machined gears, with stability up to at least 120° C (248° F). It can also be coated on metal parts by a number of coating techniques to give corrosion resistance.

Condensation polymers. The formation of polyamides has been briefly described earlier. By far the greatest production is of nylon-6,6, made by heating the amine and acid in an autoclave (an airtight vessel for carrying out chemical reactions under pressure) from which it is extruded on a chilled drum and cut into chips. It has a high melting point, strength, stiffness, toughness, and chemical resistance. After careful drying it can be extruded and injection molded.

Nylon-6,6 applications

Nylon plastics. These properties class it as an engineering plastic (a plastic used as a substitute for metal) and lead to a very wide range of applications such as gear wheels, oil seals, bearings, and parts for textile machinery. Extruded film is also used for boil-in-bag packaging. A large number of different molding powders is made by variation of molecular weight and copolymerization of various amide-forming constituents.

In addition, a number of different nylons are made by varying the molecular length of the diamines and the dibasic acids: nylon-6,10 using sebacic acid (ten carbon atoms); nylon-11, from an acid derived from castor oil; and nylon-12, made by a multistage process from butadiene. Terephthalic acid with methyl hexamethylenediamine gives a transparent polymer instead of the normal waxy product. The primary object of these variations is to decrease the moisture absorption.

Nylon-6 differs from these nylons in that it polymerizes directly, without elimination of water, from caprolactam. The polymer can be cast in place and is specially suited to irregular forms and thick sections because the low viscosity of the monomer prevents voids and polymerization is independent of part size, producing items ranging from 500-pound (227-kilogram) blocks to fuel tanks. The polymer can be extruded as easily sealable film and used for bake-in-bag bread and roast-in-bag meat. It is also used for packaging as a laminate with polyethylene.

Other variations are the use of amines with a ring structure, particularly the aromatic nylons, which give polymers with softening points above 300° C (572° F). A paper is made from chopped fibres from the phenylenediamine nylon that has electrical properties similar to mica.

Polyethylene terephthalate. Polyethylene terephthalate is a polyester, theoretically obtained by the reaction of glycol with dibasic terephthalic acid but actually made from glycol and dimethyl terephthalate by the process known as ester interchange. During the reaction methanol is produced instead of the water displaced during ordinary esterification. It is extruded as a transparent film in amorphous form and is immediately chilled. If it is stretched in two directions at right angles (biaxially stretched) at a temperature above 80° C (176° F), it is still mainly amorphous and can be used for shrink wrapping foodstuffs (a shrink wrapping shrinks after it is put around the object) and medical accessories by momentarily reheating it above the stretch temperature.

If, however, the film, biaxially stretched to three or four times its original dimensions, is held at a higher temperature, it crystallizes and develops very high tensile strength (25,000 pounds per square inch, or 1,760 kilograms per square centimetre) and dimensional stability.

Its main application is in food packaging, where it is used for meat packaging and for heat-and-serve packaging of convenience foods. For these applications it is frequently laminated with polyethylene. Its great strength and high dielectric strength make it ideal for insulation in electric motors, printed circuits, and cable insulation. Its dimensional stability and strength lead to its use for magnetic tape and increasingly for photographic film base. A special variety of the polymer has been developed that will mold to give the hardest plastic known.

Polycarbonates. Polycarbonates are long chain linear polyesters of carbonic acid and dihydric phenols, such as bisphenol A. This product is a faintly amber transparent plastic of good toughness, tensile strength, ductility, and dimensional stability maintained over a temperature range of 16° C to 150° C (60° F to 300° F).

It can be injection molded, blow molded, and extruded and is an ideal engineering plastic with good electrical insulating properties, finding applications in portable tool housings, safety helmets, computer parts, and vandal-proof windows and light globes. A number of special grades are made, including a special type to meet specific flammability requirements. Glass-reinforced polycarbonate can carry continuous loads up to 4,000 pounds per square inch (280 kilograms per square centimetre) at boiling-water temperature.

Uses of polycarbonates

Polyimides. Polyimides are made by a two-step reaction of dianhydride (tetrabasic acid with two molecules of water removed from each molecule of tetrabasic acid by chemical reaction) with an aromatic diamine giving first a polyamic acid that is converted by heat or catalysts to high molecular weight linear polyimides, available as film, laminating resins, adhesives, and molding compounds.

The polyamic acid is soluble; films can be cast from it and it can be used for laminates and adhesives. Polyimides are the best organic materials available for high-temperature insulators, capacitors, and bearings, and have special applications in space technology. Exposure at 275° C (527° F) for one year causes only a 1 percent elongation. Polyimides can also be produced as foamable powders for aerospace structural applications.

Other condensation polymers. If one amino group in *m*-phenylenediamine is converted to an amide group by formic acid, an intermediate is formed that on loss of water becomes benzimidazole. This reaction can be made using a wide range of reactants, producing resins (polybenzimidazoles) that can be used as high-temperature adhesives for laminates in the aerospace industry and compounded with tiny bubbles of silica (microballoons) and alumina fibres to provide insulation for re-entry vessels, fuel cells, and nuclear reactors. If the fillers are phenolic microballoons and carbon fibres, very high temperatures develop a carbonaceous char with excellent structural and ablative properties.

Aromatic polysulfones are transparent linear polymers made from the sodium salt of bisphenol A and *p, p'*-dichlorodiphenylsulfone. Light amber in colour and exceedingly resistant to oxidation, these plastics have extraordinary dimensional stability, a stress of 1,500 pounds per square inch (105 kilograms per square centimetre) for 100 hours at 150° C (300° F), producing barely visible deformation. These are first-class engineering plastics, used for underhood automotive parts, housings for power tools, and for dishwashers. Other phenols and sulfones are used for special high-heat-resistant materials.

Polyphenylene oxide is a linear polymer formed by oxidation of 2,6-xylenol by oxygen in the presence of a catalyst. It has very accurate dimensional tolerances at high temperatures and high chemical resistance. It is suitable for clock parts and surgical instruments and is in general a very good engineering plastic because of its stability, heat resistance, and low moisture take-up. Many variations in phenols are possible.

Though the starting materials for polyhydroxyethers, (phenoxy resins) are the same as for epoxy resins, phenoxy resins are of high molecular weight. An excess of epichlorohydrin is avoided to ensure thermal stability. Their properties can be varied by the use of different phenols, but the polymer from bisphenol A is commonly used. For coatings the hydroxyl group is usually esterified, and with drying oils thermosetting coatings can be made; phenoxy coatings are excellent primers because of high adhesion to metals and overall compatibility with topcoats of other plastics.

As molding powders they give tough, lightweight, clear moldings, easily printed and with good resistance to water and other liquids making them specially suitable for blown bottles. They have high abrasion and chemical resistance and can be hot gas welded. A special use is for jet-fuel tanks.

Steam at 950° C (1,740° F) converts *p*-xylene to an active intermediate that can be polymerized on a cooled surface to form poly *p*-xylylene. It is used only for very thin coatings as in miniaturized capacitors and for encapsulating such materials as lithium.

A prepolymer soluble in chlorinated solvents is made by

decomposing benzenesulfuryl dichloride in the presence of terphenyls at 300° C (570° F). Impregnation of reinforcing material with such a solution is followed by a series of heating processes with added benzenesulfuryl dichloride and catalysts, giving a cross-linked structure that is the best ablative resin, withstands temperatures up to 400° C (750° F), and remains unchanged for years at 200° C (390° F).

Coumarone–indene resins

Coumarone–indene resins, readily fusible, low-molecular-weight polymers, are not strictly plastics. They are made by polymerization of coumarone (benzofuran) and indene obtained from crude petroleum or from gas tar and coal tar resulting from the high-temperature carbonization of coal. The coumarone content is often less than 10 percent, and the polymer may contain styrene and methyl styrene, softening points varying with the composition up to 150° C (300° F). The catalysts are cationic (positively charged), and batch operation is traditional. Coatings of the resin are water and alkali resistant and are particularly useful vehicles in aluminum paint. They are also used in conjunction with chlorinated rubber and modified phenolics. They also greatly assist in the blending of fillers in rubber for floor coverings.

OTHER TYPES OF RESINS

Oil-soluble or modified resins. Alkyds used as coatings are polymers in which the spare hydroxyl, resulting from the esterification of a triol (usually glycerol) by a dibasic acid, is esterified by a long-chain (fatty) acid, which may contain double bonds. Since oils such as linseed are already glycerol esters, reaction is brought about by splitting off one or two acid chains—*i.e.*, reconstituting glycerol hydroxyl groups and re-esterifying these with more triol and a dibasic acid, usually phthalic acid, giving a complex polymer in which the acid chains polymerize among themselves and may retain double bonds which oxidize in air. As such, these are good drying oils.

Physical properties, including solubility, depend on composition, but their outstanding characteristic is compatibility with other resins.

Simple phenol– or cresol–aldehyde resins are insoluble in drying oils, but if resols are heated with an excess of an unsaturated organic acid such as rosin, they become soluble with removal of water and can be dispersed in ester gum (rosin glyceryl ester). By cooking with unsaturated oils such as tung and linseed, the rosin-modified resols disperse with some interesterification to produce air-drying varnishes.

The oil may also be incorporated by reacting phenol and formaldehyde in its presence or by reacting the phenol with the oil, with subsequent condensation with formaldehyde. Substituted phenols are sometimes used because they are more readily soluble in vegetable oils.

Plastics from other natural products. Casein plastic is made by reacting casein from milk with formaldehyde. It is produced in a variety of opaque colours, including pearl; it softens in hot water but takes a very high polish. Its use is greatly diminishing, but buttons and buckles are still made, partly because short runs, advantageous in fashion trades, are practicable.

Uses of lignin

Lignin is a mixture of complex thermoplastic phenolic polymers present in wood and straw. It can be used as a molding powder, but its chief use is for hard fibreboards, made by compressing high-lignin-content material at 10,000 pounds per square inch (700 kilograms per square centimetre) at 250° C (480° F), with or without previous treatment with phenol and formaldehyde. The lignin content is increased by disintegrating the cellulosic material and hydrolyzing part of it by treatment with steam and water at high pressure, the remaining fibre acting as reinforcement.

Synthetic resins for special applications. These applications include resin adhesives, paper, and textiles.

Apart from nitrocellulose and polyvinyl acetate resin, adhesives are thermosetting.

Resins for bonding

Phenol–formaldehyde and urea–formaldehyde resins are used chiefly for bonding wood, especially plywood. Phenol–formaldehyde joints are resistant to boiling water and by addition of resorcinol acquire greater strength, which, for metal to metal joints, achieves a maximum with the addition of polyvinyl formal. Epoxy resins also adhere well to metal and are popular for domestic as well as industrial use.

Urea–formaldehyde resins are cheap, show no glue line, and are used extensively in furniture. Polyurethane joints formed in place with diisocyanates are chiefly used to cement rubbers to metals or textiles. Polysulfones, polyimides, and polybenzimidazoles are used as high-temperature adhesives.

By the addition of small amounts of amino-resins to paper pulp, wet strength, bursting strength, rub resistance, and dry tensile strength are greatly improved. Urea–formaldehyde resin is also used to produce wallboard. Paper is rendered waterproof by coating with coumarone–indene resin and still more effectively by lamination with polythene film at very high speeds.

A number of polymers are used for external finishes, either in organic solvent solution or as emulsion, the most important being urea– and melamine–formaldehyde resins, and complicated derivatives thereof, acrylics and alkyds.

Amino resins are used for washfast, glazed, shrink-resistant chintz and for embossed effects. Acrylics stiffen without destroying flexibility and improve the abrasion resistance of both cotton and rayon, while polyethylene is applied in film form for permanent glazing. Butadiene–acrylonitrile improves the wear of cotton goods, and polyvinyl acetate is used extensively on linings that are often calendered to give a silk effect. The most interesting use of alkyds is that of a thermosetting castor-oil-modified glyptal that stiffens, is a fixing agent for pigments, and preserves mechanical finishes.

Plastics are also applied for crease resistance, especially of rayon and cotton. Urea–formaldehyde and melamine derivatives are chiefly used in solution, catalyzed by ammonium salts of strong acids; a variation is the inclusion of octadecyloxymethylpyridinium chloride, which confers water repellence, and of urea–formaldehyde resin with dichloropropanol, which produces an epoxy finish and is used in drip-dry materials.

III. Plastics manufacture and fabrication

NATURAL RAW MATERIALS

These include cellulose, coal, and petroleum products (see also COAL PROCESSING; PETROLEUM REFINING; RUBBER).

Cellulose. Originally the only source of cellulose for nitration, acetylation, and other processing was cotton linters, the short fibres left after ginning, which removes the long staple fibres. After cleaning and bleaching the linters are ready for acetylation and other processing.

Undegraded cellulose is a crystalline material with chains built up of about 3,500 repeating units of glucose anhydride, but this number falls to about 500 on acetylation, the fall being limited by process control as far as possible since molecular weight and physical properties are closely connected. Linters are now being replaced by cellulose derived from wood pulp by alkali treatment.

Coal and petroleum products. The products of the carbonization of coal are tar and light oils (or benzol), in proportions and quantities dependent upon the type of coal and the process employed. The amount of coal carbonized to make marsh gas (methane gas) is diminishing rapidly because of natural gas and the use of petroleum products, so that coke ovens are the largest source of by-products, with substantial additions from smokeless fuel plants.

Tar is distilled, and processing of the fractions yields phenol, cresols, and naphthalene, the first two used in formaldehyde resins and the latter for phthalic anhydride by oxidation.

Distillation of the light oils after alkali washing yields benzene, toluene, and xylene, and a residue of "heavy" naphtha containing coumarone and indene, which are separated by polymerization; coal tar is the only source of these polymers.

Coal by-products thus contain most of the raw materials for synthetic plastics, especially if acetylene from calcium carbide is included, but the vast quantities now required

for plastics can only be provided by recourse to the petrochemical industry.

Raw materials from oil

Ethylene and propylene are produced on an enormous scale by cracking naphtha in tubes heated by steam to 750–900° C (1,380–1,650° F) or by direct heat transfer. Ethylene is also formed by catalytic cracking at 700° C (1,290° F) of a special fraction that also gives a high content of aromatic hydrocarbons, benzene, toluene, and xylenes, as well as a residue from which napthalene is recovered.

Aromatics are also obtained by the platforming process in which feedstock is passed over platinum-coated silica–alumina, but naphthalene cannot be obtained directly from this product because most of it is present as an alkyl derivative and must be dealkylated by a special process that may also be applied to toluene to produce benzene.

Xylenes, which are also present, are separated by distillation and low-temperature crystallization, the *para*-xylene being required for oxidation to terephthalic acid; vapour-phase catalysis of the other xylenes produces more *para*-xylene.

Ethylene is now used to make vinyl chloride by way of ethylene dichloride; it combines with benzene to make styrene via ethylbenzene and yields glycol via ethylene oxide. Propylene is converted to glycerol through oxidation to acrolein; further treatment with hydrogen peroxide and chlorination of glycerol gives epichlorohydrin for epoxy resins. Ammoxidation, in which propylene reacts with air and ammonia, yields acrylonitrile. Acrolein is also an intermediate in the route to acrylic acid. The oxidation of isobutylene yields methyl methacrylates.

There are a number of ways of converting benzene to phenol—by sulfonation followed by fusion with sodium hydroxide, or chlorination and hydrolysis; and by the catalytic reaction of propylene with benzene, giving cumene, which is then oxidized with air to give cumene hydroperoxide, this product in turn being converted by strong acids to phenol and acetone. The mixture of phenol and acetone can be condensed directly to bisphenol A, widely used in epoxy resins and polycarbonates. Oxidation of benzene by air yields maleic acid, and treatment with methanol vapour gives cresols.

Methane is converted to acetylene, which reacts with acetic acid to form vinyl acetate; vinyl acetate can also be formed directly from ethylene and acetic acid in a reaction catalyzed by copper and palladium salts. Another petroleum product is cyclohexane, from which both adipic acid and hexamethylenediamine are made for polymerization to nylon-6,6.

CHEMICALS AND ADDITIVES

These include plasticizers, stabilizers, antioxidants, ultraviolet absorbers, organic peroxides, colorants, and flame retardants.

Plasticizers. Plasticizers, usually high-boiling, low-volatility esters, are incorporated with polymers by heating and mixing to increase flexibility by inserting themselves between the polymer chains and thus reducing the attractions between chains. They are essential in cellulose plastics and include triphenyl phosphate, a common ingredient to reduce flammability. The esters are usually diethyl and dimethoxy glycol phthalates but in polyvinyl chloride, which accounts for the largest proportion of plasticizer used, phthalate esters of much higher molecular weight are employed because the plasticizer proportion is higher and low volatility is imperative. It is also important that flexibility be maintained at low temperatures. The compound di-2-ethylhexyl phthalate (DOP) is much used, and for maximum low-temperature flexibility the adipate as well as the azelate are employed. These three plasticizers have the advantage of being safe in vinyl food-wrapping film. Of the phosphates, only octyl diphenyl phosphate may be used for this purpose.

Some materials are classed as extenders rather than plasticizers and cannot be used alone, nor are esters universally acceptable; nylon, for example, requires a glycol or an aromatic sulfonamide.

Another class of plasticizer, either a complete polyester made from a glycol, or an epoxy compound, usually of soya oil, is known as a "permanent" plasticizer because of its nonvolatility; the epoxy types are used with other plasticizers in vinyl compounds and also function as stabilizers.

Stabilizers and antioxidants. Polymers are subject, in varying degrees, to degradation by heat, light, and oxidation and can be protected by substances that only arrest but do not entirely prevent these processes.

Polyvinyl chloride in particular requires stabilization, especially with the development of rigid products, since the manipulation temperatures are high enough to initiate decomposition, but polyolefins and polystyrene are also improved by suitable additives. In polyvinyl chloride the main problem is emission of hydrochloric acid; this can be reduced by adding lead salts or soaps such as lead stearate. Though these are inexpensive, they cause opacity and are toxic; they are widely used for insulation compounds, however. Where a degree of transparency is required with freedom from stain, barium–cadmium salts, with or without zinc, are employed; they have the advantage that they can help to eliminate "plate out," the deposition of constituents on the metal of the processing plant.

With silicate and asbestos fillers, epoxy resins and phosphites are used, but for rigid transparent polyvinyl chloride, processed at high temperature, a different class of stabilizer, that of alkyl compounds of tin, is used.

Dibutyltin derivatives

After 1950 a series of dibutyltin derivatives, such as the dilaurate and maleate, were recognized as excellent though expensive stabilizers for high-clarity polyvinyl chloride, a property becoming more and more important with the increasing use of blown polyvinyl chloride bottles. These stabilizers are now being replaced by dioctyltin derivatives containing maleate groups or mercaptoacetates, which are far less toxic than the dibutyltin compounds.

For plastisols, barium, cadmium, and zinc compounds, together with organic additives, are preferred. Nontoxic stabilizers also include calcium–zinc compounds with phosphates and epoxies. Formulations are widely varied according to processing conditions, resin characteristics, and end uses.

Catalyst residues and metal traces from the processing plant promote degradation, which is diminished by antioxidants and ultraviolet absorbers. Organic phosphites are used to deal with metallic impurities and traces of heavy metals and also to decompose any trace of peroxide.

Antioxidants fall into five classes: phenol derivatives, amines, esters, organic phosphites, and miscellaneous—the choice depending on the polymer to be protected and the end use.

Low-density polyethylene is an important example of a polymer that must be protected by antioxidants. It is characterized by side chains, and each junction is vulnerable to oxidation, which is dealt with by phenols, alone or in conjunction with sulfur-containing esters or phosphites. Oxidation caused by light calls for ultraviolet absorbers, the most satisfactory of which is carbon black, the antioxidants in this case being phenols and naphthols containing sulfur.

Because of its structure, polypropylene is extremely susceptible to oxidation, a reaction that becomes important because the substance is handled at high temperatures. Antioxidants similar to those used for low-density polyethylene are effective, especially phenols and sulfur compounds, as distearyl thiodipropionate, used together.

Polystyrene is relatively stable to oxidation but sensitive to light degradation. High-impact polystyrene, however, because of its rubber content, requires the use of phenols. Most other polymers are less subject to oxidative degradation resulting from exposure to oxygen, but phosphite esters are often used, especially with polyurethanes, which are likely to discolour.

As already indicated, ultraviolet light combines with oxidation to promote degradation, and combinations of absorbers and antioxidants are common, the ultraviolet absorber being increasingly important as plastics are used in outdoor applications.

Classes of absorbers

The two classes of absorbers are screening agents and excited-state quenchers. Excited-state quenchers stabilize the polymer by removing energy from the molecule that has already absorbed ultraviolet light. Screening agents are the most widely used, absorbing most of the incident ultraviolet light and converting it to heat. The most effective of these are the benzophenones, especially their ethers and benzotriazoles. Ultraviolet absorbers also work by converting shortwaves into harmless longer light waves through fluorescence. Protection by fluorescent dyes is particularly applicable to films. Of the excited-state quenchers, nickel compounds are the most effective, especially organo-nickel complexes.

Organic peroxides. Organic peroxides and hydroperoxides are derivatives of hydrogen peroxide and split readily into the free radicals (groups of atoms that act as a unit in a chemical reaction) required in addition polymerization, over a temperature range determined by composition. Room temperature splitting can be brought about by activators, such as cobaltous ion or tertiary amine. Industry offers about 45 different peroxides.

Colorants. Colorants are inorganic or mineral pigments, organic pigments (either toners, insoluble organic salts, or metal complexes of dyes), or are lakes, dyes deposited on alumina.

Uniform dispersion and standardization of colour are usually accomplished by predispersion in a resin at high temperatures or, especially in the case of epoxy resins, polyesters, and urethane systems by dispersion in the plasticizer, monomer, or solvent.

Dyes must be used for transparent plastics, either dissolved in solvent or in such fine dispersions, as in the case of basic dyes, that no particles are visible.

Flame retardents. Where nonflammability is especially important, all polymers except rigid polyvinyl chloride and fluorocarbons require flame-retardant additives, even though some, such as nylons or polycarbonates, are self-extinguishing. The most important flame retardants are bromine, chlorine, antimony, and phosphorus compounds. Antimony compounds can be used only with opaque materials, however.

Additives vary with the polymer; various phosphates are used with plasticized polyvinyl chloride but with polyurethanes the phosphates must be brominated, and the same applies to polystyrene foam, where highly brominated phosphates or hydrocarbons are incorporated in the beads as supplied for molding. An alternative method is to build bromine-containing units into the polymer chains, as can be done in the case of polyesters, polyurethane foam, and epoxy resins.

Fillers and reinforcements. These fall into five main categories: mineral and synthetic inorganic powders, carbon blacks, cellulosics, metal powders, and microspheres.

Reinforcements include cotton and asbestos flocks; glass fibres, chopped or in the form of rovings, mats, or monofilaments; carbon fibres; and mineral whiskers.

A great variety of fillers especially suitable for different polymers and end uses is available—*e.g.*, wood flour, first used for phenolics, which is cheap and reduces mold shrinkage; asbestos flock (tiny fibres of asbestos), which greatly increases impact strength; various fillers that improve polyurethanes; and aluminum, magnesium, and titanium oxides, which give added stiffness to composites.

Fibre reinforcement

Cotton flock is essential in urea–formaldehyde resin moldings, and many synthetic fibres are used with different polymers; the great bulk of reinforcement, however, is glass fibre of less than 25 microns (ten millionths of an inch) in diameter and available as glass cloth; the cloth can be chopped for incorporation in moldings or used as mats or as filament for a thin surface mat or for winding.

Carbon fibres made by heating acrylonitrile fibres have enormous tensile strength, and plastics bonded with them can replace metals. Sapphire fibres and single-crystal whiskers also provide powerful reinforcement.

BASIC TECHNIQUES FOR PROCESSING PLASTICS

The chief plastics-processing techniques are extrusion, blow molding, film and sheet extrusion, calendering, thermoset molding, injection molding, rotational molding, thermoforming, laminating, casting, film casting, and foaming.

Extrusion. A major technique of the plastics industry, extrusion consists essentially of the melting and compression of plastic granules by the rotation of a screw conveyor in a long barrel to which heat or cold can be applied. The screw drives the plastic through a nozzle (see Figure 2). The technique is used on a large scale for the continuous production from thermoplastic materials of pipe, sections of various forms, sheeting and films, wire covering for insulation, plastic feed for blow-molding plants (see *Blow molding*), and preplasticizing systems for injection molding. Screws may vary considerably in design, but the most widely used is the constant pitch screw divided into feed, transition, and metering zones. The feed zone preheats the plastic granules in a deep screw channel of constant cross section and feeds them to the transition zone, which is of decreasing depth so as to compress and generally homogenize the melting plastic. This in turn discharges into the metering zone, which is of a constant depth less than that of the feed zone; it serves to complete the melting and mixing operation and to generate maximum pressure for pumping the molten plastic through the shaping nozzle. Additional homogenization may be obtained by including a mixing device; *e.g.*, a fluted cylindrical section in the screw that shears the molten plastic intensively as it passes through fine channels between the flutes and the barrel wall.

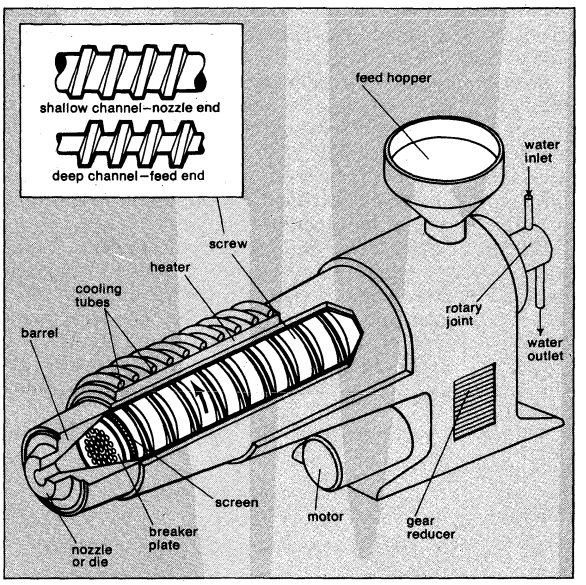


Figure 2: Extrusion-molding equipment. The screw has a shallow channel at the nozzle end and a deep channel at the feed end, as shown in inset.

The pressure of the plastic melt depends on the resistance of the nozzle but may be regulated by means of a throttle placed near the screw discharge. This may be a simple breaker plate with a screen pack, which also helps to homogenize the melt and to filter extraneous matter, or a streamlined adjustable valve that can be regulated while extrusion is in progress. Devices for automatic changing of screen packs without stopping extrusion are available.

Modern extruders may have a longer and more complex screw than the basic one described above. Such a screw is typically in two stages, each of which has a carefully selected combination of zones. Between the stages a decompression of the melt makes possible the incorporation of a vent hole in the barrel wall at this point for the removal of volatile materials and trapped air from the melt. Vacuum may be applied to this vent to increase the effect, so as to give a product of enhanced quality at high production rates.

The screws vary in design to give optimum results for different plastics. An important specification is the ratio of

length to diameter, generally about 25 to 1 although 35 to 1 or more is not uncommon for two-stage screws. Special alloys usually are used both for barrel and screw, the latter being bored for heating or cooling. Direct current motor drives are preferred for accuracy of control; the heating or cooling zones on the barrel, which may be built in, are controlled precisely, usually by electronic methods.

Normal nozzle diameters range up to eight inches (20 centimetres) but can be larger. For extruding pipe, a device called a mandrel, centred in the nozzle, determines the wall thickness.

Blow molding. Extrusion is used to push a molten tube (called a parison) into a bottle-shaped mold. Compressed air then forces the parison against the cold walls of the mold, hence blow molding. Molds are generally side fed, with wall thickness controlled by a tapered mandrel (core) or variable-orifice die. Continuous extrusion is possible by use of multiple blow molds.

Film and sheet extrusion. Thin polyethylene film is made in large quantities by extruding a tube through an annular nozzle behind which is a device through which air is introduced into the emerging tube. The air expands the tube to two or three times its original diameter. The plastic is then pressed between rollers over which the film passes as a double layer; subsequent slitting gives two flat films.

Thick-sheet extrusion

Thicker sheet is extruded through special nozzles (see Figure 3). Acrylonitrile–butadiene–styrene sheet one-

From M. Smith, *Manufacture of Plastics*; Chapman & Hall

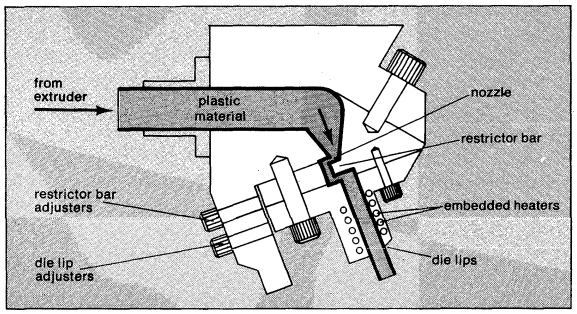


Figure 3: Apparatus for forming thick sheets by extruding plastic through a special nozzle.

quarter inch (six millimetres) thick is made in this way. Plastics that melt, such as nylon and polyethylene terephthalate, are deposited on a roll, cooled, and stretched before reeling. Controlled stretching at elevated temperatures in one or more directions can lead to improved clarity and strength (oriented films).

Calendering. Soft flexible polyvinyl chloride sheet, plain or embossed, is made in large quantities by feeding hot premix between rollers individually driven and heated to about 150–180° C (300–355° F), with an output ranging up to 10,000 pounds (4,500 kilograms) per hour. When desired, the sheet can be embossed, laminated to fabric, or laminated to another plastic sheet (to make tiles); this may be done immediately as the sheet emerges from the rollers.

Thermoset molding. Thermosetting resins are usually molded in hydraulic presses, using steel molds cored for heating or cooling. Pressures up to 3,000 pounds per square inch (210 kilograms per square centimetre) are required for cold powder or pellets, but these feeds are usually preheated by hot air or by infrared or radio frequency dielectric heating and pressures are lower. Preheating reduces curing time. Radio-frequency heating is particularly suitable for semi-automated molding.

Injection molding. Injection molding differs from compression molding in that the plastic material is rendered fluid in a separate chamber outside the mold, then is forced into the mold by external pressure. The mold is cooled and split, the two halves being locked by pressure during molding and opening automatically when the mold cavity has been filled by a shot of the liquefied plastic. The cooled article is then ejected from the mold and the process repeated.

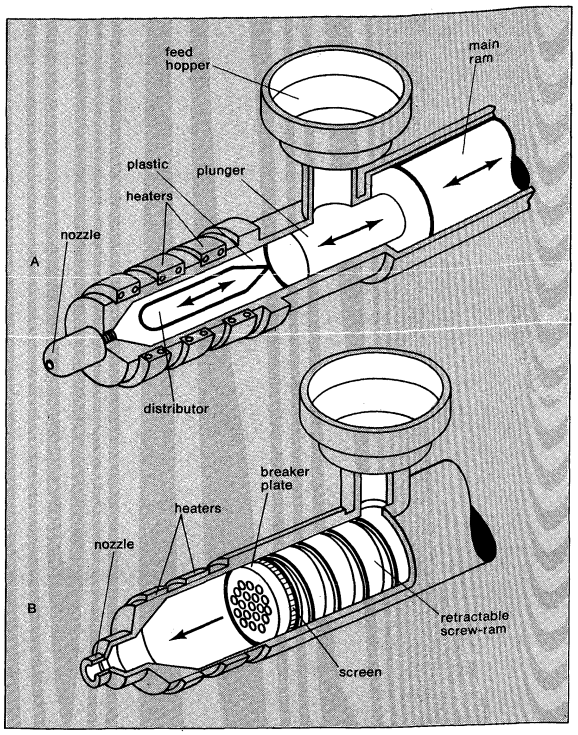


Figure 4: Thermoplastics injection machine with (A) hydraulically operated ram and (B) retractable screw-ram (see text).

In its simpliest form (see Figure 4A), the machine consists of a heated barrel with a hydraulically operated ram, an opening in the rear of the barrel being fitted with a hopper into which measured amounts of plastic pellets are fed. Falling in front of the plunger, the pellets are forced under high pressure into the heating section of the barrel, pushing out an equivalent volume of molten plastic into the mold. A torpedo-shaped distributor is fixed between the end of the ram and the nozzle feeding the mold.

The reciprocating screw feed

Of the various methods used to speed preplasticization, the most common involves a reciprocating screw feed that builds up a shot of molten plastic behind the nozzle, moving backward against the injection ram pressure until an electrical device stops rotation and the screw operates as a ram to force the plastic into the mold (Figure 4B). The two halves of the mold are clamped at a pressure that resists that exerted by the plastic entering the mold. This is usually achieved by a hydraulic ram or by a hydraulically operated toggle mechanism; the nozzle may be fitted with a device to prevent premature leakage of plastic at the nozzle and permit high pressure buildup before the shot.

Rotational molding. Plastics structures with heavy walls and complex shapes can be produced by rotational molding, in which a measured amount of plastic in powder or liquid form is placed in the cold mold, which is then rotated about both its vertical and its horizontal axis to distribute a layer of plastic evenly over the whole of the interior surface. While rotating, the mold is passed into a heated chamber where fusing or curing takes place; quick cooling solidifies the plastic, producing an accurate replica of the internal mold surface.

Heating is usually by hot air, up to 450° C (840° F); rotation speeds are infinitely variable up to about 12 revolutions per minute on the major axis and 40 revolutions per minute on the minor; the molds are inexpensive, usually made of cast aluminum. For extremely large cavities, sheet metal can be used, especially for prototypes. The ideal mold material has a high heat conductivity; beryllium copper is sometimes used. Polyethylene, polypropylene, and polyvinyl chloride are the plastics most widely used. Large commercial containers, special carrying boxes, and automotive headrests and armrests are made by rotational molding.

Thermoforming. A thermoplastic sheet, heated to its softening point, can be forced against the contours of a mold by air pressure, forming a desired shape after cooling. Many methods differing in detail are employed, but five principal techniques exist: simple vacuum forming, drape forming, and three modifications to permit deep drawing with maintenance of even thickness of walls and bottoms of the shapes. In simple vacuum forming, the plastic sheet is clamped over an airtight box, from which air is exhausted, rapidly drawing the heated plastic down on a "former" within the box, thus reproducing its shape. The sheet is heated above the box by electric heaters. In drape forming, the electric heater plate above the sheet is removed and a male former is driven up into the clamped soft sheet. Vacuum is employed to pull the sheet tight to the former.

Deep drawing

Both processes are modified to permit deep drawing. One technique involves plug-assisted drawing, in which a plug is forced down into the sheet before the application of vacuum. In drape forming, a controlled bubble is first blown upward in the sheet into which the former then enters. Another technique resembles plug-assisted forming except that the sheet is forced into the female former by compressed air.

Laminating. At first, only solutions of phenol–formaldehyde and urea–formaldehyde resins were used to impregnate paper, cotton fabric, absorbent forms of asbestos, and other fabrics. Later, melamine–formaldehyde resins were added. Material from a roll of fabric was passed through a bath and then a drying chamber; cut up sheets of the dried material were stacked in a press, then heated and cured at a pressure of over 1,000 pounds per square inch (70 kilograms per square centimetre); hence they were known as high-pressure laminates. Today, the development of polyester resins and epoxy resins that cure without giving off any by-products permits the use of low pressures, sufficient only to hold the reinforced layers in sheet form; these are known as low-pressure, or contact, laminates.

A variation of the sheeting method involves addition of the resin or its components to paper pulp, which is then sheeted according to standard paper methods. The most important of these are polyester resin laminates of glass-cloth and epoxy-resin laminates with carbon fibres. "Prepregs" and "postforming" laminates are impregnated uncured sheets that can be molded in metal molds.

Where tubes are required, the impregnated material is wrapped on suitable metal cores, revolved between heated rolls, removed, and baked to complete polymerization.

Casting. Polyesters, polyurethanes, silicones, and epoxy resins can be obtained in liquid form without solvents and can be formed by mixing in a catalyst and pouring the liquid into an inexpensive mold. Under suitable conditions, the mixture hardens into a solid shape. Electrical potting (sealing an electric component by placing it in a container and filling the container with the plastic compound) and industrial patterns are typical cast products. Specimens such as coins or butterflies can be embedded in attractive form in polymethyl methacrylate by immersing them in partially polymerized monomer and completing the polymerization very slowly.

Film casting. Film casting produces continuous rolls of thin film by spreading a polymer solution on a moving surface passing through a heated tunnel (see Figure 5). The solidified film is stripped as it returns to its casting point, and residual solvent is removed by passage through heated multiroller cabinets, in the course of which, if necessary, coatings can be applied. The moving surface may be a wheel (18 feet, or 5.5 metres, in diameter) with a polished chrome surface or a copper band, up to 200 feet (60 metres) long with a special polyvinyl alcohol surface. The process is applied almost entirely to cellulose triacetate and diacetate, the former for the production of photographic film base and the latter for packaging and diazo reproduction films.

Foamed plastics production. Foamed or expanded plastics are low-density materials made by forming gas bubbles in the plastic while in a liquid or molten state. A variety of processes and types of products are available,

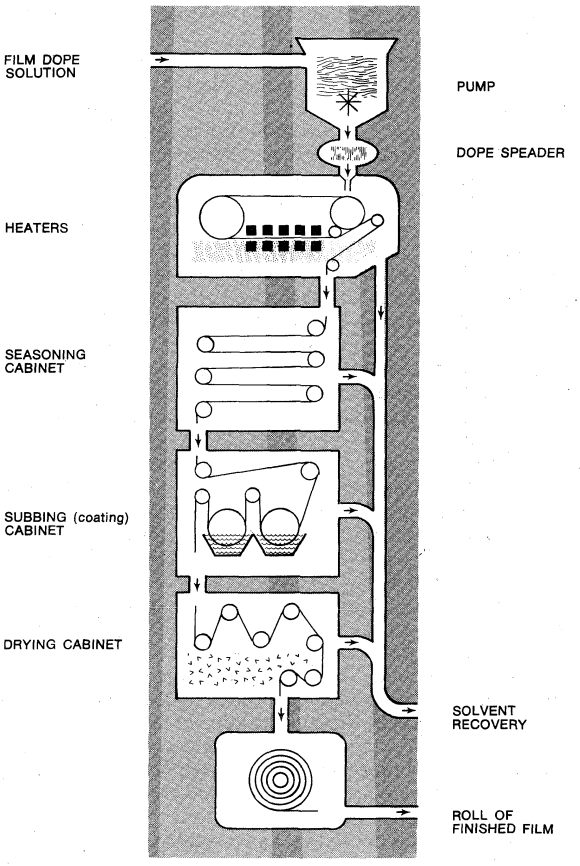


Figure 5: Basic operation of a film-casting machine.
Drawing by D. Meighan

depending upon the type of polymer and whether the structure is closed cell or open cell (*i.e.*, with the bubble cavities intercommunicating).

In some cases (*e.g.*, polyurethane foams), bubbles are produced in the low molecular weight mixture of reactants undergoing polymerization, partly arising from gases produced in the condensation process but usually supplemented by a blowing agent (*e.g.*, a volatile fluorocarbon liquid). Flexible polyurethane foams are of the open-cell type, like a sponge, and indeed are used for such purposes, as well as for heat insulation, upholstery, and other uses. Rigid or flexible thermoplastic materials (polyvinyl chloride, polyethylene, polystyrene) may be foamed in one of two general ways:

1. By dispersing a blowing agent in the material and heating it above its softening temperature. The blowing agent forms a large number of small bubbles in the material, and may be a volatile liquid such as pentane used in expanded polystyrene, or an unstable chemical such as azodicarbonamide, releasing much gas when heated.
2. By dissolving a gas such as nitrogen or carbon dioxide in the heated plastic at a high pressure. When the pressure is released, a large number of small bubbles are formed and grow in the soft plastic. This process may be carried out in autoclaves or continuously in extrusion machines, the plastic expanding as it leaves the extrusion nozzle.

Most of the foamed materials manufactured by the above two processes are produced in simple forms from which sheets or simple shapes are cut. In recent years so-called structural-foam processes have appeared, by which molded articles are produced from foamed plastics materials. Rigid plastic foams of relatively high density (*i.e.*, 0.5 grams per cubic centimetre) have a useful combination of lightness and rigidity, particularly if the molding process results in the formation of a relatively solid skin of plastic over a low-density expanded core.

FABRICATION AND FINISHING

Techniques of fabrication and finishing include mechanical machining, laser machining, roll-leaf hot stamping,

printing, welding, radiation processing, vacuum metallization, electroplating, painting and decorating, mold decoration, and inserts.

Mechanical machining. Most plastic articles do not require machining except for the removal of sprues (small projections) on injection molded parts. Parts may be machined from stock forms, however, such as tetrafluoroethylene rod or screw threads cut on pipe or laminates may be sawn and drilled. Machining properties of plastic are entirely different from those of metal. Plastic is more flexible, necessitating maximum support to prevent bending. Buildup of heat in the workpiece is rapid, causing it to soften. Furthermore, the plastic may deform slightly after machining, making it difficult to do highly accurate work. To diminish heat, tools are honed and polished. Lubrication is supplied, and the speed of cutting is adjusted for different plastics. Drills have wide polished flutes, and oversized taps are used to counter deformation after tapping.

Machining properties of plastics

Plastic sheet, especially acrylonitrile–butadiene–styrene, can be cold-formed under high pressures between metal dies. Good lubrication and polished drawings are required —these drawings control the drawing in of the sheet during forming. Attractive articles may be formed from previously decorated flat sheet. Plastics can also be forged like metals if the blank is heated nearly to the crystalline melting temperature. Forged parts can have ten times the impact strength of the blank.

Laser machining. A laser beam removes material by vaporizing it. Holes can be drilled in less than a thousandth of a second; thus perforations can be made on a moving plastic web, using a controlled pulse rate from the laser. The holes range in size from two to 50 mils (one thousand mils equal one inch or 25 millimetres) in diameter and are slightly tapered.

Aerosol nozzles in nylon and acetal are satisfactorily bored with lasers. Film can be slit with a continuous beam, without forming any dust; the beam can also be used for heat sealing.

Roll-leaf hot stamping. This process derives from the art of tooling gold leaf on leather. It is applied mainly to decoration or labelling of thermoplastic articles and consists of the application of a complex layer of which a typical composition includes a carrier film, usually polyglycol terephthalate, 12–25 microns (0.0005–0.001 inch) thick; a release coating on the carrier, based on a waxy material and about one micron thick; a lacquer coating that protects the decorative coating that is to follow (the lacquer is metallized with aluminum or gold to a thickness of a few microns, having been previously printed, or is coated with a pigment without a metallic layer); and, a final heat- and pressure-sensitive adhesive layer to give maximum adhesion to the plastic to be printed. The adhesive-coated side of the film is pressed by a heated die to the object to be decorated, and the adhesive bonded to the plastic under pressure. After a short cooling period to allow the adhesive to solidify, the carrier film is stripped off, that part of the resin coating carrying the metal coating or pigment that corresponds to the die form adhering to the object.

Decorating thermoplastic articles

Polyester film as a carrier has the advantage of withstanding high temperatures, which is important because the die may be at a temperature as high as 130° C (270° F), though the usual temperature is about 110° C (230° F). Metal dies are usually steel, though softer metals can be employed for short runs. There is, however, a growing use of silicone rubber, the latest development being a metal–rubber sandwich, the rubber layer being about 1/16 inch (1.6 millimetres) thick, faced with 1/32 inch (0.8 millimetre) of copper, which can be photoengraved. Such a die gives flexibility sufficient to cope with surface irregularities on the plastic faces.

Printing on plastics. There are two principal factors involved in printing on plastics: the choice of ink and the type of printing process. Ink choice is influenced by the printing method. The printing method in turn is determined by the shape and the plastic used, usually flexible material such as polyethylene and polypropylene, thin polyvinyl chloride, nylon, or high-density polyethylene. For these materials flexography, which involves printing with a rubber printing plate, gravure, using engraved metal rollers, and lithography are widely employed, chiefly for printing on the plastic web. Dry offset printing or stamping is employed for molded bottles, jars, tubes, and tubs. For bold decoration, the silk-screen process, a form of stencil printing, is employed. Another process under rapid development is electrostatic printing, in which ink powder is expelled from an image screen by a difference in electrostatic potential. Ink formulation is specific to the material and process. Alcohol-soluble polyamides are used for polyolefins; oxidizing alkyd resins and oxidizing thermosetting systems are used for polyvinyl chloride, polystyrene, nylon and also polyolefins. When printed sheets such as cut film are stacked as they come off an offset press, they are interleaved with paper panels (until the ink is dry) to prevent ink smudging.

Heat, high-frequency, and ultrasonic welding. Polyvinyl chloride sheet, in its various forms, can be welded by a hot jet of nitrogen or carbon dioxide and a rod of lower melting polyvinyl chloride. Overlap joints in tank linings of polyvinyl chloride, polyvinylidene chloride, and polyisobutylene can be welded with a type of soldering iron, followed by a small pressure wheel. Polymethyl methacrylate with a hot jet welds through contact with the monomer released by heat decomposition.

Polyvinyl chloride sheet is also particularly suitable for electronic heat sealing, as is any plastic that will absorb radio-frequency energy. A heavy duty steel press with brass dies is employed. A radio frequency generator is required, with power output ranging from three to 25 kilowatts. Electronic techniques are used to control positioning, sealing, and the length of time the press stays closed (dwell time). Very strong joints are made, and the process has been much used for applique decoration.

Electronic heat sealing

Ultrasonic (above the range of human hearing) vibrations, at about 20,000 hertz (cycles per second) are transmitted into the plastic assembly, generating heat at the joint surfaces where melting takes place. The harder the plastic, the more effective are ultrasonic welding techniques. Ultrasonic welding can be used for spot welding; staking, in which a plastic stud is inserted into a hole in metal; and for molding in inserts by means of a hole slightly smaller than the insert, heat being generated at the point of contact between metal and plastic in a second or less.

Radiation processing. Two plastics industrial processes employ high-energy electrons, the first being the curing of certain types of resin, which become thermosets (causing them to take on a rigid form) at room temperature and the second, the cross-linking of thermoplastics, causing the molecules to form a network structure, thus making the plastic infusible. Radiation sources are either electron accelerators or radioactive cobalt. In the case of accelerators, the material is passed in front of a window of the accelerator, where it is briefly exposed to radiation; the cure, or cross-linking, occurs rapidly without any catalyst. Thermoplastics, such as polyethylene, which have been cross-linked in this way, if stretched while hot, can be frozen in the new dimension; subsequent heating causes rapid and strong shrinking. Thus, for example, a four-core electrical cable can be capped with a four-tubed glove of treated polyethylene and shrunk down throughout by heat to make tight joints at all points. The process can be applied to polyethylene and other wire insulations by exposure to an electron beam, thus providing increased thermal stability and improved stress-cracking properties. Similar exposure of film gives increased clarity when shrunk in place, with better resistance to chemical attack but little reduction in permeability to water vapour.

Vacuum metallizing. Vacuum metallizing, the earliest method of metal coating of plastics, is accomplished by evaporating a metal, usually aluminum, onto plastic surfaces within a high vacuum chamber, up to seven feet (2.1 metres) in diameter and containing hangers so arranged that the workpieces can be rotated. Rotation is necessary because the metal particles travel in straight lines. Alongside the work are tungsten heater filaments on which pieces of aluminum are suspended. The tungsten

filaments are electronically heated to melt the aluminum and the temperature then raised to evaporate it while the workpieces are rotated. For vacuum metallizing, very high vacuum is required; to achieve this, most plastics are pretreated to remove moisture and any volatile substances. The time required for this process can be reduced by lacquering and baking the article, then applying a coat of lacquer after metallizing to protect the metal surface and provide colour if necessary.

Continuous coating

Continuous coating from rolls of plastic of many types is accomplished by use of special crucible sources and winding mechanisms. Metallizing is used largely for decoration; metallized polyester film is used for electrical capacitors and, finely slit, for metallized yarn.

Electroplating. The use of plastics such as acrylonitrile–butadiene–styrene, polypropylene, and polysulfone for automotive parts has been enhanced by electroplating with metal. The process depends on etching, which is easiest with acrylonitrile–butadiene–styrene because the dispersed rubber particles can be removed by oxidizing. Other plastics depend on strong acid attack. Immersion in a tin salt solution, followed by palladium salt, results in the deposition of catalytic palladium that adheres to the etched holes. Further immersion in copper or nickel salt solutions causes a layer of metal to be deposited on the plastic through reaction with the palladium, thus providing a conductive surface on the plastic for electroplating with copper, nickel, and chromium. The copper deposition may be thick enough to serve for printed electronic circuits. A number of intermediate steps are required and there are many proprietary variations. It is important that the moldings be free from strain and the visible surfaces as perfect as possible. Areas up to 350 square inches (2,260 square centimetres) can be plated.

Finishing. *Painting and decorating.* Paint can be used to obtain multicoloured or pearly finishes, to reduce gloss, and to reduce the typical plastic appearance of an object. It can also give light-coloured finishes to dark resin moldings and objects made from reclaimed scrap. Paint is usually sprayed but can be roller coated, especially for raised letters.

Mold decoration. Printed foils can be located, either mechanically or by static electricity, in a mold; thermosets are decorated with printed melamine-impregnated paper, as for dinnerware, the foil being inserted after partial curing by interruption of the molding process. For thermoplastics, film, preferably of the plastic being molded, is inserted printed-side inward.

Inserts and fasteners. Inserts, placed in the female half of a thermoset mold, can be molded in. For subsequent insertion, self-tapping and thread-forming inserts can be used. Locking devices such as serrated rings are used to secure bushings (metal linings for a hole). The cheapest inserts are driven into place, using projections that fit in grooves (splines) or expansion devices instead of threads.

IV. Industrial plastics

STRUCTURAL FOAMS

Sandwich forms

Rigid foams. Rigid polyurethane foams in sandwich forms have wide applications as building components because of the stiffness imparted by the thick foam centre for a given weight. They are also the best insulants known today and so have wide application in refrigeration and in buildings, where they are applied in fitted slab form or are foamed into cavities at the building site. They can also be applied by spraying about one-quarter inch (six millimetres) thickness with each pass of the spray gun. The ability to spray a foaming mixture through a single nozzle is a great advantage in application. An alternative application method is to mix the foaming ingredients under pressure with fluorocarbons and run the froth into the cavity, where foaming takes place through vapourization of the fluorocarbons by the reaction heat.

A very important use of rigid foam is for furniture parts to reproduce wood structures; these can be injection molded. Since the polyurethane remains in the mold a rather long time to allow foaming and setting, molds can be mounted on a rotatable table, coming around in turn for a shot of foamable material. Rigid polyurethane foam can be screwed and nailed with a retention about equal to white pine lumber.

A major advance in the manufacture of sandwich structures is a new method of injection molding invented in Britain, in which a large machine is used to produce moldings up to four feet (1.2 metres) square. A double valve provides differential feeds to the mold, the first supplying a skin of selected plastic, the second a foamable core material that forces the first polymer to the surface of the mold; the process resembles the blowing up of a balloon. Moldings of great strength and any desired surface are obtained.

Flexible foams. Flexible foams, usually polyether urethane, are made in slab form up to eight feet (2.4 metres) in width and as much as five feet (1.5 metres) high; these are then cut to required shapes or sizes or are molded. The molded foams may be hot molded. This involves filling heated aluminum castings and gives a product having high resistance to compression, as for automotive seats; or they may be cold molded, a process used particularly for semiflexible foams with high load-bearing properties and densities from five to ten pounds (2.3–4.5 kilograms) per cubic foot. Used almost exclusively by the automobile industry for crash pads, armrests, and dashboard covers, the process involves machine mixing the ingredients and pouring them into aluminum molds lined with vinyl or acrylonitrile–butadiene–styrene skins, which become the cover material for the part. These skins are being replaced by integral skins, the formation of which depends upon high fluorocarbon content of the foam and special diisocyanates, foaming occurring from the centre of the mold outward, vapor pressure consolidating the material on the mold surface. The same process can be applied to rigid foams.

Polyester flexible foams with a special silicone surface active agent are used for textile interliners, with densities about one to seven pounds (450 grams to 3.2 kilograms) per cubic foot; these can be made down to 30 mils thick and used for lamination.

Polystyrene foams

Polystyrene foams are made in a wide range of densities, from expandable beads, either by extrusion through slot-shaped openings to 40 times the original volume to form boards directly or by foaming in steam chests to form large billets. Using small beads in stainless steel molds, cups can be molded with thin sections.

Thin sheet for embossed packaging can also be made by the tube extrusion technique. Though packaging is a major use for forms made in closed molds, the largest use is for building panels; they can be plastered directly. Acrylonitrile–butadiene–styrene can be expanded from pellets and is particularly suitable for wood-grain effects and for the production of heavy sections.

Expanded vinyls can be made from plastisols for flooring or textile linings by calendering with a blowing agent and laminating to a fabric base, and by injection molding for insulation and such articles as shoe soles. An improved material is now obtained from cross-linked polyvinyl chloride and competes with polyester in glass reinforced plastic. Syntactic polybenzimidazole owes its expanded properties to the incorporation of extremely small hollow spheres (microballoons) of glass or polyfluoride resins. It is stable up to 300° C (570° F); its heat resistance and light weight adapt it particularly for use in space vehicles.

SHEETS AND FILMS

These include vinyls, fluoroplastics, and cellulose acetate.

Vinyls. Plasticized polyvinyl chloride, made by a calendering process, can be sewn, heat-sealed, or electronically sealed; it is used for apparel, shower curtains, protective clothing, and the like. Made in any colour, transparent, translucent, or opaque, polyvinyl chloride can be coated with pressure sensitive adhesives and printed with decorative patterns. Thicker sheet is coloured and embossed for women's handbags, luggage, seat covers, and upholstery.

Thin film is also used for packaging, especially of meat and fruits; if biaxially stretched, it forms a shrink film that retracts up to 60 per cent. Another important use is as a laminate for printed paper.

Flooring tiles, largely made of polyvinyl chloride, are built up by lamination and decorated either by printing or by rolling in colour chips; the commonest tile is vinyl asbestos, pressed into sheets on calenders and then embossed and cut into tiles.

Rigid polyvinyl chloride sheet has high dimensional stability and flame retardancy and is often used in corrugated form for building construction. Partitions, rain gutters, downspouts, industrial tank linings, vacuum-formed signs, and lighting panels are other uses.

Styrene film is widely used for semirigid food containers, especially packaging, being heat formable if held on all four sides to prevent shrinkage.

Acrylonitrile–butadiene–styrene possesses a wide range of properties, notably scuff resistance, plus high impact strength at lower temperatures, making it suitable for high-quality luggage, refrigerator linings, and food and detergent containers because of its chemical resistance and freedom from stress cracking. Because it is easily formed by heat, thick acrylonitrile–butadiene–styrene sheet is used for sports-car bodies and automobile doors.

Uses of nylon films

Nylon films are mostly from nylon-6,6 and are ideal for food packaging because of strength, impermeability to oils and greases, and high melting point. As such films are steam sterilizable, they find many uses in hospitals. They are frequently laminated. Acrylic films have high resistance to ultraviolet light and external exposure; their prime use is in surfacing laminates.

Fluoroplastics. Fluorinated ethylene–propylene and chlorotrifluoroethylene are the principal fluoroplastics. Fluorinated ethylene–propylene resists extreme temperatures from that of liquid hydrogen to 275° C (525° F) and is extremely inert in the presence of chemicals. Since it can be bonded to metal with special adhesives, fluorinated ethylene–propylene is used in printed circuitry, to insulate flat cable used in the computer industry, in space vehicles, and for heat-sealed industrial and biological packaging.

Both fluorinated ethylene–propylene and chlorotrifluoroethylene are noncombustible and have very low moisture vapour transmission rates; the fluorohalocarbons also have excellent thermoforming characteristics. They are used in packaging of critical hardware, in aircraft insulation, encapsulation, and in other electrical objects.

Polyurethane film and sheet are available in thicknesses from ½ mil to 100 mils (2.5 millimetres), the outstanding characteristic being toughness and resistance to tearing. Tensile strength and elongation are both very high, leading to uses for awkward packaging and for inflatable sheet used to stabilize ship cargoes.

Cellulose acetate. Cast film up to ten mils thick is used for rigid packaging, especially for "blister" packaging, designed to protect small articles by inserting them in preformed shapes.

Acetate film can be printed or drawn upon and hence is used widely for visual aids and in engineering and architectural drafting.

Cellulose triacetate is more moisture resistant than ordinary diacetate and, until recently, plasticized with high-boiling plasticizers, has been pre-eminent as a photographic film base, being cast, for example, to make rolls of motion picture film base 3,000 feet (910 metres) long and over 40 inches (one metre) wide. After emulsion coating, these are slit into standard motion picture film widths for edge perforation and addition of the sound track. Cellulose triacetate is similarly used for amateur film bases at 3.5 mil thickness and for X-ray films at 7.8 mil. In the field of colour-separation films, however, it is being replaced by polyethylene terephthalate, which has greater dimensional stability. Cellulose acetate film has a wide range of other uses, the result of its great tensile strength (25,000 pounds per square inch, or 1,760 kilograms per square centimetre), its high dielectric strength (withstands high voltages), and resistance to heat. Thus it is extensively used for electrical insulation of motors and capacitors; for packaging, which accounts for three-quarters of its utilization; and for magnetic recording tapes. Polycarbonate sheet is heat-, odour-, and stain-resistant and practically unbreakable. Polyimide film has no melting point and so can be used at high temperatures; it maintains its flexibility at 4° above absolute zero. Its toughness, flame resistance, and radiation resistance make a valuable combination for electrical uses, especially in space technology.

Uses of polyimide film

LAMINATES, REINFORCED PLASTICS, AND COMPOSITES

High-pressure laminates with phenolic resin binders are available in many grades, from strong paper laminates with varying quantities of resin to fabric-reinforcements of cotton, asbestos, nylon fibre, and glass-fibre fabric. All are rigid, can be built up to required thicknesses, can be readily machined into required dimensions, and in many cases can be punched. They are chiefly characterized by excellent electrical properties and resistance to humidity, combined with high mechanical strength, all these properties depending on choice of resin content and reinforcement employed. Cotton-fabric-reinforced laminates having very high impact strength are used in a wide variety of mechanical applications, especially gears.

Composites, which incorporate a third layer in addition to the resin and reinforcement, have special applications. Very large quantities of copper-clad laminates are used for printed circuits, and there is an increasing use of plasticized polyvinyl chloride laminated to steel or, in some cases, to aluminum. The polyvinyl chloride protects the metal and also carries pigmentation of desired colours.

Decorative laminates are widely used on a base of phenolic paperboard by laminating the surface with a printed sheet of wood grain or specific designs and finished with a urea– or melamine–formaldehyde transparent protective surface layer. Melamine, silicone resins, and epoxides are used instead of phenolic resins in applications requiring electrical properties under wet conditions or flexural (bending) strength at high temperatures. Where reinforced shapes other than sheets are required, postforming laminates or prepregs are used. Postforming laminates are sheet assemblies in which the phenol–formaldehyde retains a degree of thermoplasticity, so that it can be molded into shapes such as kitchen trays and covers for electrical apparatus, the molded depth depending on the extensibility of the reinforcing fabric. Prepregs are glass-fibre mats impregnated with resin, fillers, colour, and catalyst, the resin being polyester, epoxides, phenolics, or silicones, partially cured, the final curing process being effected in the mold. Prepregs are particularly adapted to mass-production techniques.

Low-pressure (contact) laminates

Contact laminates are formed by placing reinforcing materials with polyester or epoxide resins, in viscous liquid form, on a mold. The major plastics material used is the polyester type previously described; the reinforcing material is glass fibre coated with a silicone or silyl compound to promote adhesion to the plastic. The fibres are used either as rovings (slightly twisted strands or rolls of textile fibres), mats, yarns, or woven fabrics, or, for reinforcing molding, as powders in the form of short strands cut in lengths up to two inches (five centimetres).

The desirable quality of glass fibres is their tensile strength, amounting to 250,000 pounds per square inch in commercial form.

Because of this great strength, glass-reinforced plastic can be used for such structures as small ships, footbridges, etc. A glossy finish can be given to the surface of these laminates by finishing them with a thin mat of staple filaments.

Because of its strength and lightness, glass-reinforced plastic has been much used for transport applications, and there is a small production of all glass-reinforced-plastic car bodies, made by a hot-press molding process. Automobile manufacturers are beginning to turn attention to compression-molded polyester glass composites for large area moldings, including body panels.

A different method of producing glass-reinforced plastic

is filament winding, in which glass-fibre filament is passed through a resin bath and wound on a former to produce pipe and tanks of great strength. Pipes of large diameter can be made in this way.

Many other fibres are in use or under development; polyvinyl alcohol fibres, for example, adhere better than glass to plastic, giving higher impact strength and lighter weight; but the greatest interest lies in materials that offer greater stiffness and strength. Boron fibres are about six times as stiff as glass, but no stronger, whereas carbon fibres made from acrylonitrile filament combine increased stiffness with very high tensile strength ranging from 300,000 to 400,000 pounds per square inch. Composites of carbon fibre and resin can have the same tensile strength as steel but weigh only a quarter as much. Wear rate and friction can be much reduced, and resistance to high temperatures can be greatly increased.

MISCELLANEOUS INDUSTRIAL PLASTICS

Extruded forms, made on an extruder with a special die, though commonly produced as rods or tubes, may also be manufactured in geometrical shapes to suit given tasks (curtain runners, for example), as interlocking strips for panels and window frames, and as a special interlocking strip that is wound into a spiral to form polyvinyl chloride pipes of large diameter finished by continuously welding the spiral joint.

A special application of extruded film, generally polypropylene, is the use of tapes, which are woven into simulated jute products for sacks and carpet backing. Special treatment of the tapes leads to manufacture of carpet yarns.

Engineering plastics, materials that can substitute for metals, require tensile strengths well above 5,000 pounds per square inch and flexural strengths (resistance to bending or flexing) at least 15,000 pounds per square inch. Plastics with these properties as a rule have softening points well above 100° C (212° F), so that their stability is unaffected by any ambient temperatures to which they are likely to be exposed; also, they can be sterilized. The earliest use of plastics in engineering was as gears cut from laminated phenolics and used for heavy duty with water lubrication. In current use are thermoplastics such as acrylonitrile–butadiene–styrene, polycarbonates, and nylon, admirably adapted for light mechanisms such as clocks and cameras as well as heavy parts such as fuel tanks and many smaller automotive parts. The use of chopped-glass-fibre filling, which can increase tensile and flexural strengths by about 7,000 pounds per square inch, has increased the scope for such uses.

BIBLIOGRAPHY. M. KAUFMAN, *Giant Molecules* (1968), a superbly produced introduction to the technology of plastics, fibres, and rubbers; E.G. COUZENS and V.E. YARSLEY, *Plastics in the Modern World,* rev. ed. (1968), the best known work on plastics; M.L. MILLER, *The Structure of Polymers* (1966), a monumental reference book; F.W. BILLMEYER, *Textbook of Polymer Science* (1962), a comprehensive classic, suitable for first courses in polymer chemistry at the graduate level; G. ODIAN, *Principles of Polymerization* (1970), a well balanced treatment of all the types of reactions used to synthesize polymers; H. LEE, D. STOFFEY, and K. NEVILLE, *New Linear Polymers* (1967), a description of 10 new commercially available high-temperature polymers, developed since the mid-60s; W. MAYO SMITH (ed.), *Manufacture of Plastics,* vol. 1 (1964), information on all stages of the manufacture of commercial plastics, from monomer production to polymer finishing and shipping; J.A. BRYDSON, *Plastics Materials* (1969), one of the most readable, useful, and comprehensive books on plastics yet written; P.D. RITCHIE (ed.), *Physics of Plastics* (1965), a collection of 8 monographs by members of the Plastics Institute; *Modern Plastics Encyclopedia,* probably the most popular and well used reference book on plastics.

(E.G.C.)

Plata, Río de la

The Río de la Plata (literally River of Silver; in English often called the River Plate) is a tapering intrusion of the Atlantic Ocean occurring on the east coast of South America between Uruguay to the north and Argentina to the south. While some geographers regard it as a gulf, others even as a sea, and still others as a river, the majority regard it as the estuary of the Paraná and Uruguay rivers. As such, the Río de la Plata receives waters draining from the basin of these two rivers, which covers northern Argentina, the greater part of Uruguay, Paraguay, eastern Bolivia, and much of southern Brazil and represents an area of 1,600,000 square miles (4,144,000 square kilometres). Montevideo, the capital of Uruguay, is located on the northern shore, and Buenos Aires, the capital of Argentina, is on the southwestern shore.

The head of the Río de la Plata is where the delta of the Paraná and the mouth of the Uruguay meet; the eastern, or Atlantic, extremity lies about 180 miles seaward and would be represented by a line on a map drawn from Cabo San Antonio, Argentina, to Punta del Este, Uruguay. The breadth of the estuary increases from the head seaward: it is 31 miles from Punta Lara on the southern (Argentine) shore to the port of Colonia del Sacramento on the northern (Uruguayan) shore, and 136 miles from shore to shore at the Atlantic extremity of the estuary. To those who regard the Río de la Plata as a river, it is the widest in the world, with a total area of 13,500 square miles (35,000 square kilometres). (See also city articles on BUENOS AIRES; MONTEVIDEO; for an associated physical feature see PARANA RIVER).

The natural environment. *Physiography.* The two contributory rivers bring down about 2,000,000,000 cubic feet of silt every year. The muddiness of the water in the Río de la Plata itself is increased by the tides and winds that prevent the settling of the deposit on the bed. The deposit thus forms great shoals, banks, or bars of clay, sand, and organic matter: the Playa Honda Shoal is just off the Paraná Delta; the Ortiz and Chico shoals are further downstream; and the Rouen, the Inglés, the Alemán, and the Arquímedes shoals are still farther out. The depth of the water—varying from six feet above the shoals to 65 feet in the intervening channels—is reduced along the south coast by an offshore shoal. Silt and shoals

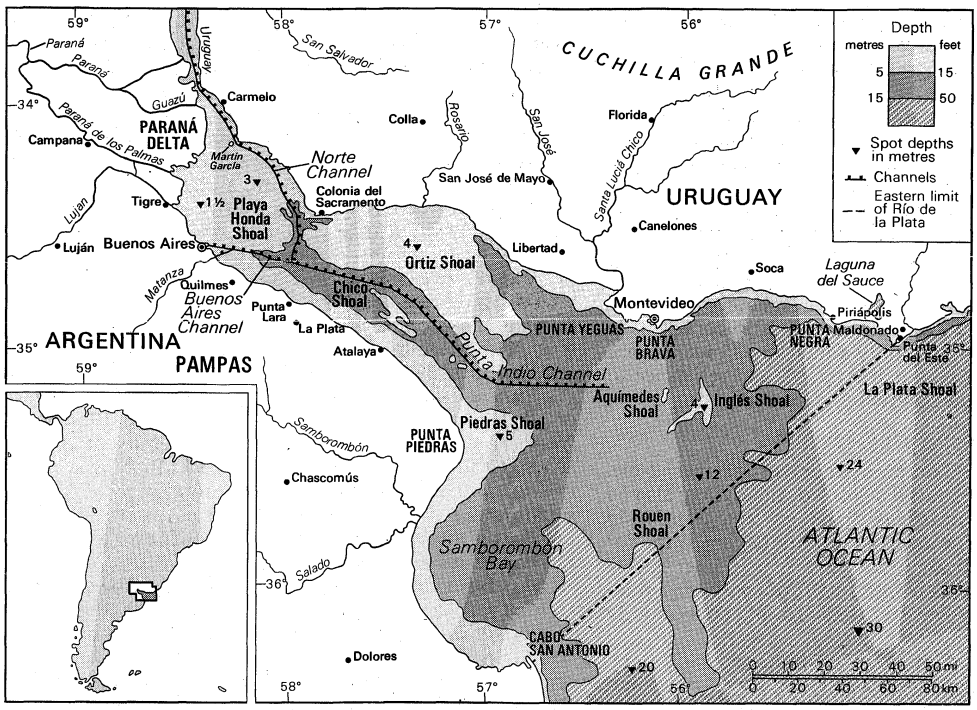
The Argentine coast of the estuary is low-lying; its banks are of marine debris and coarse sand, and the coast is subject to flooding in places. The Uruguayan coast stands quite high and consists largely of rocks and dunes. Off the Uruguayan coast there are several small islands, such as Hornos, San Gabriel, López, Lobos, Farallón, and—opposite the mouths of the Uruguay and Paraná Guazú rivers—Martín García.

Hydrology. The volume of water discharged by the Río de la Plata into the Atlantic is estimated at 776,925 cubic feet per second. Although the water of the tributary rivers is so widely distributed over the length and breadth of the estuary that variations in their volume do not affect the level of the water, the level is considerably affected by variations of the tides and, especially, of the winds reaching it. The ocean tides, weak though they may be, flow 120 miles up the Paraná and the Uruguay rivers from their mouths on the estuary. The average tidal range is six inches at Montevideo and two and a half feet at Buenos Aires. The pampero (the southwest wind) and the southeast wind both exert a great influence on the Río de la Plata: the pampero, when it is most powerful, drives the water on to the Uruguayan coast, so that the water level drops on the Argentine side; the southeast wind has the contrary effect, flooding parts of the Argentine shore, and causing the level to drop on the Uruguayan coast.

Climate. The Río de la Plata has a temperate climate and copious rainfall at all seasons, amounting to 39 inches a year. The mean annual temperature is 55° F (13° C), and monthly averages are always over 50° F (10° C). Humidity, at 70 percent, is notably high. The prevailing winds are the north, the pampero, and the southeast. Frosts are frequent in the winter months.

Fish. The estuary is very rich in fish; pejerrey (kingfish), corbina (white sea bass), patí (large scaleless river fish), surubí (the largest Argentinian river fish), dorado (a large golden fish resembling a salmon), and pacú (a large river fish almost as high as it is long) are among the edible species.

Human ecology. For the people living along its shores, the Río de la Plata has been useful as a waterway; for their essential livelihood they have looked inland to the



Río de la Plata.

pampas (grassy plains). Nevertheless, the estuary as a thoroughfare for trade is of major importance not only to the people of the coasts but also to the remotest inhabitants of the whole drainage basin.

Exploration. After the European discovery of the Pacific Ocean at the Isthmus of Panamá (1513), King Ferdinand II of Aragon commissioned the Spanish navigator Juan Díaz de Solís to look for a sea passage from the Atlantic to the Pacific. In 1516 Solís found his ship in freshwater at the mouth of the Río de la Plata estuary, which he named Mar Dulce (Freshwater Sea). Solís, however, was killed by Indians on the mainland. His Mar Dulce was subsequently renamed Río de Solís in his honour.

Early nomenclature

The Portuguese navigator Magellan reached the estuary in 1520 and explored it cautiously before continuing his voyage to the south. The Italian navigator Sebastian Cabot made a detailed exploration in the years 1526 to 1529, finding the Uruguay and the Paraná rivers and sailing up the Paraná River and a long way up the Paraguay River as well. It was his interest in the silver objects possessed by the local Indians that led in 1533 to the estuary's being named the Río de la Plata. The Spanish colonization of the shores was begun by Pedro de Mendoza, who founded Buenos Aires in 1536.

Navigation. The chief ports of the estuary are Montevideo, Colonia del Sacramento, and Maldonado on the left bank and Buenos Aires and La Plata on the right. While Montevideo has direct access to deep water, Buenos Aires requires the constant dredging of the Punta Indio Channel, which extends for 130 miles from the Ortiz Shoal to the Argentine mainland; 22,000,000 tons of deposit are taken from this channel every year in order to maintain a depth of 30 feet. Navigation upstream from Buenos Aires, likewise, requires the dredging of the Buenos Aires Channel (Martín García Channel), which, in its full length, runs between the Ortiz and Playa Honda shoals and forms a transverse link between the Punta Indio and Norte channels. The Norte Channel, north of the Ortiz Shoal, continues to the northwest, running between Martín García Island and the Uruguayan coast to the mouths of the Uruguay River and of the Paraná Guazú in the Paraná Delta. Martín García Island, which belongs to Argentina, stands less than two miles off the coast of Uruguay and commands the approach from the sea to the rivers and so to the whole basin of the Río de la Plata. The Paraná de las Palmas, the second mainstream of the Paraná Delta, has a bar at its mouth.

Channels

Economic resources. The shores of the Río de la Plata are the most densely populated areas of Argentina and of Uruguay alike. The great ports, Buenos Aires and Montevideo, where the density is highest, are primarily concerned with exporting meat from the hinterland; the refrigerators and shipyards required for this trade are located in the coastal zone, as are flour mills, factories for vegetable oils, textile industries, metallurgical plants, and petroleum refineries.

The port of Buenos Aires handles 80 percent of Argentina's imports by value and 70 percent of its exports. Of Paraguay's exports, 50 percent are destined for the ports of the Río le la Plata; nearly all the remainder goes down the estuary to overseas countries. Montevideo handles 75 percent of Uruguay's foreign trade. Buenos Aires is linked with Colonia del Sacramento and with Montevideo by daily boat services.

The water of the estuary is thus exploited for urban and industrial purposes as well as for navigation. Fishing is of little importance commercially, except that seals are caught around the Uruguayan islands. Some of the Uruguayan beaches—at Pocitos, Carrasco, Piriápolis, and Punta del Este—have been developed for vacationers.

BIBLIOGRAPHY. ORGANIZATION OF AMERICAN STATES, *Developmental Plans for the River Plate Basin* (1967), brief description of the river system and analyses of economic resources, transportation, and commerce between countries of the river basin; and *The River Plate Basin: Studies for Development* (1969), general information about the river system together with an inventory of climatic and hydrographic facts; ARGENTINE REPUBLIC, SECRETARIA DE MARINA, *Derrotero Argentino*, pt. 1, *Río de la Plata* (1965), a detailed study.

(D.O.)

Plateaus and Basins

Plateaus are high, flat-surfaced areas of the Earth's surface that typically are bordered on one or more sides by escarpments. Basins are low, flat-surfaced, elliptical or circular areas of the Earth's surface surrounded by mountains or highlands. Both terms are so loosely used that they cannot be defined rigidly. The term plateau, for example, is commonly applied to broad areas of hilly uplands that are plateau-like only in a most general sense. Nearly all of Africa is considered a plateau because it consists of a high, gently rolling region separated from a narrow, low coastal belt by an escarpment.

The term basin also has several different meanings. It is applied to drainage areas of large rivers, sediment-filled

structural downwarps of the Earth's crust (*e.g.*, the Paris Basin), and to major depressions of the ocean floor.

Plateaus and basins occur on every continent. About 30 percent of the Earth's land surface consists of plateaus (excluding ice plateaus), if the term is applied in its most general sense to include most of the continent of Africa. Topographic basins, however, constitute only about 3 percent of the land.

Kinds of plateaus

Plateaus develop in a variety of ways and occur in many different geologic settings. They can be classed as (1) low plateaus in stable ancient continental areas (shields) developed by differential erosion of flat-lying cover rocks (*e.g.*, the Brazilian Highlands) or by uplift of ancient crystalline rocks previously planed off by erosion (as in southern Africa); (2) low plateaus on flat-lying resistant rocks within sedimentary basins (plateaus in the Paris structural basin); (3) plateaus on flat-lying or gently folded strata on the flanks of major mountain systems (Appalachian Plateau); (4) plateaus developed within major mountain systems by uplift of previously developed erosion surfaces (plateaus in the eastern highlands of Australia); (5) intermontane plateaus high in mountain systems that occur over flat-lying or moderately deformed sedimentary rocks between flanking mountain ranges (Tibetan Plateau); (6) lava plateaus developed by the piling up of basaltic lava (Ethiopian Plateau); (7) volcanic plateaus high in mountain systems formed by the accumulation of vast quantities of volcanic materials such as lava and tuff (Altiplano and Puna of the Andes mountains); and (8) large continental ice plateaus (Greenland Ice Plateau).

Basins form by downwarping or downfaulting (vertical movement along fractures in the Earth's crust) and occur within mountain ranges, in areas of flat-lying sedimentary rocks, and in areas of ancient crystalline rocks.

This article describes the distribution of plateaus and basins by continent and presents specific examples of these landforms. For an overview of the Earth's surface features, see EARTH, PHYSIOGRAPHY OF; and for information on oceanic analogues, see OCEAN BASINS. See also the articles on the individual continents (*e.g.*, ASIA).

Occurrence and distribution. *North America.* Plateaus and basins are widespread in the intermontane region of the western cordillera (mountain system) of North America. In Alaska the area between the coastal mountains on the south and the Brooks Range on the north is called the Yukon Plateau. It consists of a vast, plateau-like rolling upland but also includes mountain ranges, river valleys, and irregular basins. In Canada the region between the coastal ranges and the Rocky Mountains consists of tablelands and low mountains and is referred to as the British Columbia Interior Plateau. It is generally 160 kilometres (100 miles) wide and has general summit elevations of about 1,100 metres (3,500 feet). In the United States the broad intermontane area between the coast ranges and the Rocky Mountains has a diverse topography including two typical plateau regions: (1) the Columbia Plateau, a Tertiary (2,500,000 to 65,000,000 years ago) lava plateau averaging 300 metres (1,000 feet) in elevation, and (2) the Colorado Plateau, composed mostly of flat-lying Mesozoic (65,000,000 to 225,000,000 years ago) sedimentary rocks lying at elevations of 1,200 metres (4,000 feet) or more. The remainder of the intermontane United States consists of mountain ranges separated by basins, forming the Basin and Range Province. The mountain ranges of North America extend southward into western Mexico, where most of the Sierra Madre Occidental is a volcanic plateau from 2,000 to 3,000 metres (6,000 to 10,000 feet) high flanked by regions of basin and range structure. East of the volcanic plateau in central Mexico is the "Central Mesa" (Central Plateau), which is more properly called a basin because it is a generally flat elliptical area surrounded by mountains or plateaus.

The Columbia and Colorado plateaus

In the eastern United States, a low, gently rolling region cut on crystalline rocks east of the Appalachian Mountains is commonly called the Piedmont Plateau, although it is a plateau only in the sense that a gentle slope separates it from the coastal plain to the east. West of the Appalachian Mountains, a region of plateaus 1,000 to 1,500 metres (3,000 to 5,000 feet) high underlain by resistant Paleozoic (225,000,000 to 570,000,000 years ago) sedimentary rocks is referred to as the Appalachian Plateau. Low plateaus on stratified (layered) rocks occur west of the Appalachian Plateau in Kentucky and Tennessee. Another high, hilly region (the Ozark Plateau) occurs in Arkansas and southern Missouri.

South America. Plateaus occur in five principal regions of South America: the Guiana Highlands, the Brazilian Highlands, the Paraná structural basin, the Patagonia region, and the Andes mountains. Topographic basins are not characteristic of South America.

The Guiana Highlands in the northern part of the South American continent and the Brazilian Highlands in the central part are broad interior regions of hilly uplands, low mountains, and tabular plateaus at a general elevation of 1,000 to 1,200 metres (3,000 to 4,000 feet). These plateaus are cut mainly on little deformed Paleozoic and Mesozoic sedimentary rocks and are within a stable continental block that has undergone little deformation since Precambrian (older than 570,000,000 years) time.

The low plateaus of the Paraná structural basin of southern Brazil, northern Argentina, Paraguay, and Uruguay are on relatively undeformed flood basalts much like those of the Columbia Plateau of North America.

Patagonia, along the east side of the Andes mountains in Argentina, is a relatively high, plateau-like region underlain by nearly horizontal sedimentary formations.

Extending for about 2,400 kilometres (1,500 miles) along the Andes from central Peru across western Bolivia to Argentina is a high intermontane volcanic plateau, called the Altiplano in Bolivia and the Puna in Argentina. It is generally 3,000 to 4,500 metres (10,000 to 15,000 feet) above sea level and is interrupted in places by high, intermittently active volcanoes. In Bolivia a segment of the plateau 800 kilometres (500 miles) long and 160 kilometres (100 miles) wide has been downdropped to form a basin of interior drainage containing Lake Titicaca, the world's highest large body of navigable water.

The Altiplano

Europe. In Europe plateaus occur in Spain, in a broad region north of the Alps, in Scandinavia, and in the western part of the Soviet Union. Basins occur within and adjacent to the Alps and in Spain.

The central part of Spain is occupied by the Meseta Central, a generally elevated plateau-like area formed by the irregular uplift of a flat-surfaced earth block composed of folded Paleozoic rocks. Part of the Meseta Central consists of broad basins drained by major rivers.

Some of the plateaus of central Europe north of the Alps, such as the Massif Central of southern France, the Vosges and Black Forest of France and Germany, and Bohemia in Czechoslovakia and Germany, are like the Meseta Central of Spain; they consist of uplifted blocks of folded Paleozoic rocks. The Massif Central of France, with large areas over 1,000 metres (3,000 feet) in elevation, is perhaps the most plateau-like.

Other plateaus in central Europe are developed in areas of flat-lying sedimentary rocks, where resistant formations have been uplifted bordering mountain ranges (*e.g.*, the Podolian Plateau, east of the Carpathian Mountains) or where sedimentary layers have been differentially eroded (*e.g.*, the low plateaus within the Paris structural basin).

Much of the southern part of the Scandinavian peninsula is a plateau-like area lying 1,000 to 1,200 metres (3,000 to 4,000 feet) above sea level. It has a glaciated surface developed on Precambrian crystalline and Paleozoic metamorphic rocks (those altered by high temperature and pressure). Also, most of the southern part of Finland can be considered a plateau, although it lies only 75 to 150 metres (250 to 500 feet) above sea level.

In the western part of the Soviet Union, plateaus occur within areas of relatively flat-lying rocks bordering mountain ranges (*e.g.*, the Ufa Plateau, west of the Ural Mountains) or within platform areas (long-stable structural plains).

Asia. The most spectacular plateaus and basins in the world occur in or adjacent to the great mountain chains

of Asia. These ranges are composed largely of complexly folded and faulted Paleozoic and Mesozoic sedimentary rocks that have had a varied geologic history.

The Anatolian Plateau

The Anatolian Plateau of central Turkey lies between two major ranges in the westernmost part of the mountain belt of Asia. It averages 1,000 metres (3,000 feet) in elevation and contains numerous depressions or basins, some of which hold salt lakes. The central part of Iran is also a high plateau containing a major basin of internal drainage. Farther east and lying north of the Himalayas in southern China is the world's most extensive high plateau, the Tibetan Plateau, covering an elliptical area of about 1,400,000 square kilometres (530,000 square miles) at a general elevation of 4,000 metres (14,000 feet). To the east and southeast, the highly dissected Yunnan, Kweichow, and Shan plateaus of China and Burma are in part a continuation of the Tibetan Plateau. North of the Tibetan Plateau in western China lies a vast region of mountains and basins. The basins have internal drainage and range in elevation from 300 metres (1,000 feet [Dzungaria]) to 2,700 metres (9,000 feet [Tsaidam]). Northeast of these basins is the Mongolian Plateau, varying from 1,000 to 1,500 metres (3,000 to 5,000 feet) in altitude. The highly diverse mountain region of the southern Soviet Union, extending almost continuously across Asia, contains many plateaus and a few basins. Some of the plateaus are developed in areas of relatively flat-lying rocks on the flanks of fold mountains; others are uplifted and dissected former erosion surfaces within fold mountains.

Extensive plateaus also occur in lowland areas of Asia. The Arabian Peninsula consists largely of an eastward-sloping plateau developed on uplifted Precambrian crystalline rocks and flat-lying Phanerozoic (younger than Precambrian) sedimentary rocks. The northwestern part of the peninsula of India is a lava plateau, the so-called Deccan Plateau, formed by the extrusion of large volumes of basalt during Late Mesozoic and Early Tertiary time. The Turgay area of the Soviet Union, south of the Siberian plain, and the Ustyurt area, within the Aralo-Caspian Basin in the southeast, contain low plateaus developed on generally flat-lying Cenozoic (of the last 65,000,000 years) strata. In addition, the central part of the Soviet Union consists of a broad plateau (the central Siberian plateau), formed on uplifted Precambrian crystalline rocks and flat-lying Paleozoic sedimentary rocks, that occurs at a general elevation of 500 metres (1,600 feet).

Africa. The entire African continent is considered to be a plateau because it contains huge areas of level or slightly undulating land that occur at moderate elevation. The continent is a rigid earth block composed of Precambrian metamorphic and crystalline rocks overlain by nearly flat-lying Phanerozoic sedimentary rocks.

The southern part of Africa contains an inner region of plateaus at an altitude of 1,000 to 1,200 metres (3,000 to 4,000 feet) above sea level. These plateaus are separated from coastal areas by the Great Escarpment, which can be traced, with some breaks, in an arc around southern Africa. The western and northern parts of Africa contain vast areas of low plateaus and broad topographic basins at elevations of 150 to 300 metres (500 to 1,000 feet). Some of the low plateaus are developed on flat-lying Mesozoic or Tertiary sedimentary rocks, whereas others are on uplifted Precambrian rocks. The Congo Basin, Chad Basin, and Qattara Depression are basins resulting from broad structural downwarps.

At the northern extremity of the African continent, high plateaus underlain by relatively flat-lying rocks separate two branches of the Atlas Mountains, the only Tertiary fold-mountain system on the continent.

East Africa is dominated by plateaus divided by a great rift-valley system that extends north to south for about 6,000 kilometres (4,000 miles). The rift valleys (*q.v.*) are grabens (downdropped blocks) formed along the crests of broad structural upwarps. In Ethiopia the plateaus are developed mostly on Tertiary flood basalts that occur in the area of the upwarp. Farther south, the plateaus are ancient surfaces developed mainly on the Precambrian and Paleozoic rocks bordering the downdropped rift valleys.

Australia. Australia is divided into three major landform regions: the western plateau, the central-eastern lowland, and the eastern highlands.

The western plateau, which occupies three-fifths of the continent, is much like Central Africa, as it is a shield area of Precambrian rocks 225 to 450 metres (750 to 1,500 feet) high and sloping gradually to narrow coastal plains. Its surface is monotonously flat and has developed through long periods of geologic time.

The central–eastern lowland is a wide belt of plains and basins occupying about one-third of the continent. It contains little land above 150 metres (500 feet), and it was completely covered by seas in Mesozoic time. The Great Artesian Basin in the central part of the region is one of internal drainage, ending below sea level at Lake Eyre.

The eastern highlands, which are mostly less than 1,500 metres (5,000 feet) high, consist of low mountains and plateaus. Their plateau-like character is a result of the Late Cenozoic uplift of a relatively flat erosion surface and subsequent erosion that has not advanced enough to destroy the fairly uniform summit elevations. In places, Tertiary basalt flows cap the surface, enhancing its plateau-like character.

Specific examples. *Lava plateaus.* The Ethiopian and Somalian plateaus of northeast Africa are typical lava plateaus built by the accumulation of vast quantities of basalt. Other plateaus of this type are the Columbia Plateau of the western United States, plateaus in the Paraná structural basin of South America, the plateau or tableland of Iceland, and the Deccan Plateau of India.

Ethiopian and Somalian plateaus

The Ethiopian and Somalian plateaus cover an irregular area of over 500,000 square kilometres (200,000 square miles) lying at an average elevation of about 2,000 metres (7,000 feet), higher than any other comparable region of Africa. The Ethiopian Plateau on the west and the Somalian Plateau on the east are separated by a north–south trending belt of rift valleys 60 kilometres (40 miles) wide, and their highest elevations are generally 1,000 metres (3,000 feet) above the valley floors. The limits of the plateaus are sharp in some areas and indistinct in others. Their upper surfaces are irregular and hilly, but summit elevations are generally concordant, and flat, lava-capped tablelands are characteristic. An exception, however, is the northern part of the Ethiopian Plateau, where mountain ranges rise to elevations of 3,700 to 4,600 metres (12,000 to 15,000 feet).

The oldest rocks of the plateau are Precambrian metamorphic and granitic rocks, and resting on them in places and on flat-lying Mesozoic strata elsewhere is a thick pile of basaltic rocks of Early Tertiary age that forms the upper surface of the plateaus. These lavas have a thickness exceeding 3,000 metres (10,000 feet) in places, and their volume is 170,000 cubic kilometres (40,000 cubic miles). The general uplift of the Ethiopian and Somalian plateaus took place in several phases extending to Quaternary time (the last 2,500,000 years). Some of the earliest lavas were extruded at or below sea level, but the region was considerably uplifted before the rift valleys started to form in Late Tertiary time.

According to modern theories, the eruption of the basalts, the uplift of the plateaus, and the formation of the rift valleys in Ethiopia are all genetically related. All these features are a consequence of events that are causing East Africa to separate and drift away from the remainder of the continent. During the initial phases of this separation, fractures develop in the Earth's crust, and these serve as conduits for the release of basaltic lava derived from the mantle (zone beneath the crust) far below. During and subsequent to these eruptions, other large quantities of the mantle-derived magma (molten silicate material) become entrapped below the crust, forming a pillow-shaped body that causes the overlying crust to dome or swell. The rift valleys are formed by later downdropping of crustal blocks (grabens) in the previously swollen arch.

Intermontane plateaus. Intermontane plateaus, broad high areas of diverse topography that are more or less

surrounded by mountains, occur within the boundaries of most of the major mountain ranges of the world. The Tibetan Plateau provides a typical example.

The Tibetan Plateau

The Tibetan Plateau is the largest high upland area in the world. It covers an irregular east–west trending elliptical area 2,400 kilometres (1,500 miles) long and 700 kilometres (430 miles) wide (about 1,400,000 square kilometres [530,000 square miles]) and lies at a general elevation of about 4,000 metres (14,000 feet). Extending parallel to the southern border of the plateau is a 72-kilometre- (45-mile-) wide tectonic trench (a deep trough produced by deformational forces in the Earth's crust) that controls the upper courses of the Indus and Brahmaputra rivers. South of this trench is the main part of the Himalayan mountain system, which has an average elevation of 6,000 metres (20,000 feet) and contains the highest peak in the world—Mt. Everest (8,848 metres [29,028 feet]). To the north of the plateau are the Kunlun Mountains, with peaks over 6,000 metres above sea level. To the west the plateau is enclosed by a group of mountains, and to the east it extends into the highland or plateau of Yunnan.

The surface of the Tibetan Plateau consists of broad plains and many easterly trending ranges rising from a few hundred to a few thousand feet above the general level of the plateau. Most of the drainage is internal: the streams lose themselves in innumerable salt lakes or salty swamps. The plateau is one of the most desolate areas of the world. It is cold, windy, and treeless; precipitation is generally less than ten centimetres (four inches) a year.

The Tibetan Plateau is underlain primarily by Mesozoic marine sedimentary rocks that were folded and faulted. The Himalayas to the south were formed by uplift, folding, and overriding during several orogenic (mountain-building) stages that started in Late Mesozoic time and continued throughout the Cenozoic. The most violent movements were in the Middle Tertiary, when granitic masses were intruded. The mountains have undergone great uplift, probably over 5,000 metres (16,000 feet) during the Cenozoic.

The events that formed the Himalayan mountain system are also responsible for the uplift of the Tibetan Plateau. The Himalayan mountain system is the result of the Cenozoic collision of two drifting Earth plates; the vast continental mass of Asia on the north is one plate, and the small continental mass of the Indian peninsula on the south is the other. The collision of these two plates caused the Earth's crust to buckle in a highly complex manner, forming the Himalayas. Because it was behind the Himalayas, the Tibetan Plateau was not so severely deformed, but the movement was great enough to cause the piling up of a great thickness of crustal and upper mantle material.

Western North America

Basin and range topography. Basin and range topography is best developed in the western United States and Mexico. It consists of closely spaced, parallel mountain ranges separated by elongate basins, some of which have internal drainage.

In western North America, basin and range topography occurs in an irregular area extending chiefly from northern Nevada and western Utah southward and southeastward for about 2,400 kilometres (1,500 miles) into northern Mexico. In the United States the region of basin and range topography is nearly 720 kilometres (450 miles) across. In Mexico the region of basin and range topography is divided into two parts by the Sierra Madre Occidental; the whole belt, including the medial mountains, is about 550 kilometres (350 miles) across. The most systematic development of basin and range topography is in Nevada and western Utah, where long north–south trending mountains about 16 kilometres (ten miles) wide that commonly crest at elevations of 2,500 to 3,000 metres (8,000 to 10,000 feet) separate valleys of comparable widths lying at elevations of 1,200 to 1,500 metres (4,000 to 5,000 feet). In the United States south of Nevada and Utah and in northern Mexico, the mountains and valleys are lower, and their distribution pattern is less systematic. There many of the mountains are isolated masses protruding from accumulated alluvium.

The mountains owe their origin to block faulting in which vertical uplift has occurred along extensive faults on one or both sides of the mountain block. Two different types of block faulting have been suggested: in one type, a giant crustal block is tilted, forming a mountain on the uplifted part and a valley on the downtilted part; in the other type, crustal blocks are alternately upthrown and downthrown—the upthrown blocks (horsts) forming the mountains, the downthrown blocks (grabens) forming the valleys. To form this structure by either method requires a considerable amount of extension of the Earth's crust. In Nevada and western Utah, this extension has been estimated to be about 70 kilometres (45 miles) in an east–west direction, or approximately 10 percent of the width of the basin and range region.

Ice plateaus. Most of Greenland and Antarctica are covered by thick sheets of ice that have relatively flat upper surfaces and high elevation. These ice regions are referred to as ice plateaus. The ice that forms them is over 1,500 metres (5,000 feet) thick in much of Greenland and Antarctica, and in both places their bases, where they are thickest, are below sea level. The ice is largely a relic of the Pleistocene ice age (10,000 to 2,500,000 years ago), when continental glaciers covered much of the Arctic, sub-Arctic, and even some temperate regions of the world (see further ICE SHEETS AND GLACIERS).

BIBLIOGRAPHY. Detailed information on plateaus and basins is available under descriptions of individual mountain ranges, countries, and continents. The following books contain information on specific continents and countries: W.D. THORNBURY, *Regional Geomorphology of the United States* (1965); M.G. RUTTEN, *The Geology of Western Europe* (1969); S.P. SUSLOV, *Physical Geography of Asiatic Russia* (1961; orig. pub. in Russian, 2nd ed., 1954). Details of the geology of the Ethiopian and Somalian plateaus are given by P.A. MOHR in "The Ethiopian Rift System," *Bull. Geophy. Obs. Addis Ababa*, no. 11 (1967); of the Himalaya mountains and the Tibetan Plateau by AUGUSTO GANSSER in "The Indian Ocean and the Himalayas: A Geologic Interpretation," *Eclog. Geol. Helv.*, 59:831–848 (1966); and of Basin and Range topography by J.H. STEWART in "Basin and Range Structure: A System of Horsts and Grabens Produced by Deep-Seated Extension," *Bull. Geol. Soc. Am.*, 82:1019–1044 (1971). Background information on plate tectonics related to the origin of plateaus and basins is in R.S. DIETZ and J.C. HOLDEN, "The Breakup of Pangaea," *Scient. Am.*, 223:30–41 (1970); and J.H. ILLIES, "Graben Tectonics As Related to Crust-Mantle Interaction," *Graben Problems: Proceedings of the International Rift Symposium*, pp. 4–27 (1970).

(J.H.St.)

Platinum Metals, Products and Production of

Platinum and palladium are far more important commercially than the other members of the platinum-group metals, which also include iridium, osmium, rhodium, and ruthenium. The platinum metals are notable for their wide range of applications, many of which are based on their high melting points, resistance to corrosion, and catalytic properties.

History. The first platinum to reach Europe came from placers in the Chocó district of what is now Colombia, South America. Gold also was recovered from the same placers, which are still being worked. The new material was brought to England by Sir Charles Wood in 1741 and aroused widespread scientific interest as its complexities were gradually unravelled. A method for separating quite pure platinum from the ore was devised, and the metal was found to be ductile.

The high melting points of these metals were a limitation to their use, but researchers in England, France, Germany, and Russia devised methods for consolidating and working platinum into useful forms.

Ores and mining. Platinum ores usually contain more platinum than the other platinum metals, except for the relatively rare osmiridium.

Where deposits containing platinum minerals have been subjected to weathering, the high density of platinum and its corrosion resistance have led to its concentration in riverbeds. In a number of locations, gold also is present in the same placers, but the original sources differ. The

Chocó placers mentioned earlier were overshadowed by platinum from the Urals of Russia, which dominated the world supply from 1825 until about 1914.

Platinum was long known to exist as an arsenide, sperrylite, in Canadian copper–nickel–iron–sulfide deposits, and some recovery was made beginning in 1900. Not until the adoption of electrolytic methods for refining copper and nickel in Canada and the carbonyl process in England, however, did substantial production result. Platinum and palladium dominate in these deposits, but all of the platinum metals are present and are now recovered.

Before 1930, palladium was relatively rare, being derived from various platinum ores and from the electrolytic refining of copper, a situation that changed in 1930 when production by the International Nickel Company of Canada, Ltd., sharply increased and the metal became available in quantities adequate for industrial use. Production of palladium as well as platinum from copper nickel deposits in the U.S.S.R. began to increase in the 1960s.

Recovery from South African pyroxenite deposits. In working alluvial deposits, the high density of the platinum minerals is relied upon for recovery, as is the case with gold. In treating platinum ore from the pyroxenite deposits in the Merensky horizon in South Africa, now the largest sources of platinum, the ore is crushed, then finely wet-ground in ball mills. Then the pulp flows over corduroyed tables that recover free metallic particles and coarse platinum-bearing minerals. The recovered material is further concentrated and shipped to a refinery in England. The tailing from the table separation is thickened to remove much of the water and is then sent through a flotation cycle that recovers a concentrate of nickel, copper, and iron plus the remaining platinum minerals. The recovered material is dewatered, filtered, dried, and pelletized. The pellets are fed to blast furnaces, then converted into a copper–nickel matte containing substantial amounts of platinum metals, which can be treated to separate the copper and nickel sulfides and ultimately produce electrolytic copper and nickel. The platinum metals are recovered from the anode slimes from the nickel electrolysis. Alternatively, the Wise–Vines process can be employed, in which a small amount of silicon is added to the molten matte, followed by slow solidification. A nickel-rich bottom, which carries most of the platinum metals, separates and can be worked up promptly in a platinum metals refinery.

The Wise–Vines process

Recovery from Canadian copper–nickel ores. At the operations of the International Nickel Company of Canada, Ltd., the ore is selectively floated to yield fractions containing copper and nickel. The nickel fraction, after blast furnace and converter treatment, yields a nickel-rich matte, is cooled, crushed, ground, and passed over a magnetic separator to recover nickel-rich metallics, which contain most of the platinum metals. The remaining material is given further treatment and again selectively floated to yield a low-copper and a high-copper sulfide. The latter and the earlier copper concentrates are refined by the usual electrolytic methods, as is the nickel sulfide concentrate. Gold and silver are recovered from the copper electrolysis; any remaining platinum metals are recovered from the nickel electrolysis.

Some nickel matte is converted to oxide and shipped to Clydach, Glamorgan, where it is treated by the Mond carbonyl process, in which the nickel is converted into gaseous nickel carbonyl. The platinum metals remain behind, and this material, along with platinum metal concentrates from Canadian operations, is treated at Acton, England, to isolate the individual platinum metals.

In refining, it is essential that all the gold be precipitated and that no trace redissolves, as any gold will form the explosive fulminate when ammonia is added in the palladium recovery operation. Even a small amount of fulminate can cause a serious explosion when the palladium precipitate dries. It also should be noted that the rhodium compound $(NH_4)_2NaRh(NO_2)_6$ will explode if heated. Osmium is easily oxidized, forming volatile osmium tetroxide, OsO_4, which will seriously damage the eyes.

Applications of pure metals and alloys. The mechanical properties of the six platinum metals differ greatly. Platinum and palladium are rather soft and very ductile; these metals and most of their alloys can be worked hot or cold. Rhodium is initially worked hot, but cold working can be done later with rather frequent annealing. Iridium can be worked hot, as can ruthenium, but with difficulty. Neither metal can be cold-worked appreciably.

Pure platinum and pure palladium are too soft for many applications and so are hardened by alloying, generally with one of the other metals of the platinum group, usually iridium or ruthenium. Rhodium is a less potent hardener for platinum but confers special advantages for high-temperature and certain catalytic uses. Platinum–rhodium, platinum thermocouples are standard for the measurement of high temperatures. (A thermocouple is a junction of dissimilar metals, in this case pure platinum and a platinum–rhodium alloy. A voltage is generated when this junction is heated.) All of the platinum metals can be electrodeposited, palladium being the easiest to process. The deposited metal is much harder than the wrought metal.

Hardening by alloying

Platinum. Platinum alloyed with 3.5 percent of rhodium is the standard material for crucibles in the laboratory, where they are subject to a variety of corrosives, including molten salts. A large-scale use for platinum and platinum with 10 percent and 20 percent rhodium for lining tanks has developed in the production of high-quality glass. Even more important is the use of the platinum–10 percent rhodium alloy for the bushings through which molten glass flows in producing glass fibre, a new industry made possible by platinum.

Rayon fibre is produced by forcing a cellulose xanthate solution through hair-fine holes in a spinneret made of platinum–rhodium or a gold–platinum–palladium alloy.

Insoluble platinum-iridium anodes are essential to the production of perchlorates and persulfates, while platinum-coated titanium is used as an insoluble anode in some electroplating operations and in cells for producing chlorate as well as in cathodic corrosion protection systems. Titanium coated with ruthenium dioxide is used for chlorine production.

Platinum–10 percent rhodium in the form of gauze, operating at bright-red heat, serves as the catalyst in the reaction between ammonia and air to produce nitrogen oxide and ultimately nitric acid, a standard process throughout the world. By feeding in some methane along with the ammonia mixture, hydrocyanic acid can be produced.

Platinum has long been employed as the catalyst in many other processes, but the most important new application is in the petroleum industry to produce improved gasoline and other hydrocarbon products. In this and many other applications, most of the platinum in the spent catalyst can be recovered. Plant expansion accounts for most of the demand for the metal.

Use by petroleum industry

The use of platinum for fine jewelry began about 1906. It continues to be used in jewelry, but industrial demand has provided the impetus for opening new mines.

Platinum alloys have long been used as spark-resistant electrical contact materials, but for some years the largest electrical use was for lead-ins for incandescent lamps. Because of its resistence to spark erosion, platinum was used in spark plugs for combat aircraft in World War II.

Palladium. Pure palladium was found to be very desirable for contacts in the relays of telephone and other wire communications systems, because it provided long life and a high level of reliability. This use continues to be the major one.

Palladium in relay contacts

Hydrogen diffuses rapidly through palladium and some of its alloys at moderately elevated temperatures. Industrial application is made of this property by utilizing tubes with hair-thin walls of a palladium–25 percent silver alloy to separate high-purity hydrogen from mixed gases.

Palladium hardened with ruthenium is used in certain types of jewelry, and a number of palladium-content brazing alloys are in use in the aircraft industry.

The range of corrosives that can be handled with titanium can be widened by alloying it with small amounts of palladium; alloys containing about 0.2 percent palladium are of commercial importance.

Other platinum metals. Rhodium is the whitest of the platinum metals, a property responsible for its extensive

use as an electroplate in applications ranging from the finishing of jewelry to the production of 60-inch searchlight reflectors. Rhodium has useful catalytic properties that are receiving attention. It is also suitable for electrical contacts and is particularly useful as an electroplated surface on rubbing electrical contacts. The use of rhodium as an alloy with platinum for high-temperature use is important, and the platinum, platinum–10 percent rhodium thermocouple is standard for measurement of temperatures up to about 1,500° C. For higher temperatures, thermocouples made of platinum with 7 percent and 30 percent rhodium are available. Platinum–rhodium alloys also are used for the heating windings in laboratory furnaces. Pure rhodium as well as pure iridium crucibles are used in growing crystals for laser and related applications.

Iridium, the most corrosion-resistant of the platinum group metals, is a very effective hardener for platinum. Its extremely high melting point renders it useful in special situations where oxygen can be limited, as in growing special crystals. In addition to the small percentage of iridium found in most platinum deposits, natural alloys of iridium with osmium are found in a few locations around the world. In addition, a few thousand ounces of osmiridium are recovered yearly from the gold ores of South Africa. Iridium has outstanding resistance to lead compounds at high temperatures, one of the reasons for its use in ultra-long-life spark plugs.

Sources of iridium

Ruthenium is a good hardener for platinum, the 12 percent alloy being used for contacts in voltage regulators, thermostats, and magnetos. It is used to harden platinum and particularly palladium for use in jewelry. Ruthenium content alloys are also used for tipping the nibs of fountain pens. The electrodeposition of ruthenium is possible, and its hardness recommends it in instruments where a low-pressure rubbing contact is required.

Osmium is the rarest and hardest of the group and has the highest melting point, but its ready oxidation is a limitation. It has been employed in the synthesis of cortisone.

Purity. Traces of impurities in refined platinum metals can be estimated by the spectrography of the arc emissions. In the case of platinum, the measurement of the thermal electromotive force (EMF) against a sample of extremely pure wire often is employed, as is a measurement of the temperature coefficient of electrical resistivity between 0° and 100° C.

Within the temperature range of −183° C (the boiling point of oxygen) and 630° C, the platinum resistance thermometer defines the temperature on the International Practical scale. The range from 630° C to 1,064° C is covered with the platinum, platinum–10 percent rhodium thermocouple. In view of the importance of both the temperature coefficient and the thermal EMF against a sample of extremely pure platinum, a detailed study has been made of the effect of single impurities on these properties.

Production. The major sources of platinum metals are South Africa, the U.S.S.R., and Canada. Mining in South Africa is directed primarily to the production of the platinum metals; in the U.S.S.R., platinum metals as well as copper and nickel are involved; in Canada, nickel and copper are the primary goals.

BIBLIOGRAPHY. The history of the production of platinum is well covered in D. MCDONALD, *A History of Platinum from the Earliest Times to the Eighteen-eighties* (1960). The nickel ores that are the principal sources of platinum metals in the U.S. are discussed in J.R. BOLDT and P. QUENEAU, *The Winning of Nickel: Its Geology, Mining and Extracting Metallurgy* (1967). A large world source of platinum metal in South Africa is well described in "Platinum Mining at Rustenburg," *Platin. Metals Rev.*, 13:15–22 (1969). An excellent and current summary with a discussion of areas of future demand is given in L.B. HUNT and F.M. LEVER, "Availability of the Platinum Metals," *ibid.*, 13:126–138 (1969). The properties and uses of important platinum metal alloys and compounds are described in "Platinum Group Metals," *Kirk-Othmer Encyclopedia of Chemical Technology*, 2nd ed., 15:832–860 (1968); a pioneering summary of properties is presented in R.F. VINES and E.M. WISE (eds.), *Platinum Metals and Their Alloys* (1941). The properties of palladium and its alloys are presented in E.M. WISE, *Palladium* (1968). A bibliography dealing with the catalytic properties of metals other than platinum and palladium is A.E. REA and M. BEBBINGTON (comps.), *Annotated Bibliography on Ruthenium, Rhodium and Iridium as Catalysts* (1959). A good summary of the chemical analysis of platinum metals is in the *Treatise on Analytical Chemistry*, vol. 8, pt. 2, pp. 379–522 (1963). See also D.W. RHYS and P. TAIMSALU, "Effect of Alloying Additions on the Thermoelectric Properties of Platinum," *Engelhard Inds. Tech. Bull.*, 10:41–47 (1969).

(E.M.W.)

Plato

Plato was, with Socrates and Aristotle, one of the three philosophers of ancient Greece who between them laid the philosophical foundations of Western culture. Building on the life and thought of Socrates, whom his dialogues memorialized, Plato developed a profound and wide-ranging system of philosophy that was strongly ethical, and at times even mystical, in spirit, while remaining basically rationalistic. Through his foundational metaphysics of eternal Ideas or Forms, each answering to some changeless abstract concept (or universal), Plato sought to establish a realm of absolutes, which, by incorporating ethical as well as other concepts, could allay the moral confusion of his day.

THE LIFE OF PLATO

Plato was born, the son of Ariston and Perictione, in Athens, or perhaps in Aegina, in *c.* 428 BC, the year after the death of the great ruling statesman Pericles. His family, on both sides, was one of the most distinguished in Athens. Ariston is said to have traced his descent through Codrus, the last king of Athens, to the god Poseidon; on the mother's side, the family was related to the early Greek lawmaker Solon and traced its ancestry back to Dropides, archon (one of the nine principal magistrates of Athens) in 644 BC. His mother, Perictione, apparently married as her second husband her uncle Pyrilampes, a prominent supporter of Pericles; and Plato was probably brought up chiefly in his house. Critias and Charmides, leaders among the extremists of the oligarchic terror of 404, were, respectively, cousin and brother of Perictione; both were old friends of Socrates, and through them Plato must have known the philosopher from boyhood.

Distinguished family

His own early ambitions were political. A conservative faction urged him to enter public life under its auspices —at the age of 24—but he wisely held back until the faction's policy should be declared. He was soon repelled by its members' violent acts, particularly their attempt to implicate Socrates in the illegal execution of a victim. He hoped for better things from the restored democracy and

By courtesy of the Staatliche Museen zu Berlin



Plato, Roman herm probably copied from a Greek original, 4th century BC. In the Staatliche Museen, Berlin.

even clung to this hope for a considerable time after it condemned Socrates to death. Eventually, however, he became convinced that there was no place for a man of conscience in active politics. On the execution of Socrates in 399 BC, Plato and other Socratic men took temporary refuge at Megara with Euclid, founder of the Megarian school of philosophy. The next few years are said to have been spent in extensive travels in Greece, in Egypt, and in Italy. Plato's own statement is only that he visited Italy and Sicily at the age of 40 and was disgusted by the gross sensuality of life there but found a kindred spirit in Dion, brother-in-law of Dionysius I, the tyrant of Syracuse.

The Academy and Sicily. About 387, or soon after, Plato founded the Academy as an institute for the systematic pursuit of philosophical and scientific research. He presided over it for the rest of his life, making it the recognized authority also in mathematics and in jurisprudence. From allusions of Aristotle it would seem that Plato lectured without manuscript. "Problems" were propounded for solution by the joint researches of students. On the political side there are traces of tension between the Academy and the rival school of Isocrates.

Political action in Syracuse

The one outstanding event in Plato's later life was his intervention in Syracusan politics. On the death of Dionysius I in 367, Dion conceived the idea of bringing Plato to Syracuse as tutor to his successor, Dionysius II, whose education had been neglected. Plato himself was not optimistic about the results; but because both Dion and Archytas of Tarentum, a philosopher-statesman, thought the prospect promising, he felt bound in honour to risk the adventure. The plan was, by training Dionysius II in severe science, to fit him for the position of a constitutional king who might hold Carthaginian encroachment on Sicily at bay. The scheme was crushed by Dionysius' natural jealousy of the stronger Dion, whom he drove into virtual banishment. Plato later paid a second and longer visit to Syracuse in 361–360, still in the hope of effecting an accommodation; but he failed, not without some personal danger. Dion then captured Syracuse by a *coup de main* in 357. After the murder of Dion in *c.* 354, the philosopher drew up his important *Epistles* vii and viii, reviewing and justifying the policy of Dion and himself and making proposals, unsuccessfully, for a conciliation of the Sicilian parties. Plato is said to have died (in 348/347 BC) "at a marriage feast," or, while "writing," and to have been buried in the Academy.

As for Plato's personality, apart from the reflections of it in the *Epistles*, the most valuable light thrown upon it is that of Aristotle, who described him as a man "whom it is blasphemy in the base even to praise."

Influence of the Academy

To his readers through the ages, Plato has been important primarily as one of the greatest of philosophical writers; but to himself, the foundation and organization of the Academy must have appeared to be his chief work. In *Epistles* vii, he contrasts the impact of written works with that of the contact of living minds as a vehicle of "philosophy," and he utters on his own account the same comparatively unfavourable verdict on written works that he ascribes to Socrates in the *Phaedrus*. He apparently intended his dialogues in the main to interest an educated outside world in the more serious and arduous labours of his "school."

All of the most important mathematical work of the 4th century was done by friends or pupils of Plato. The first students of conic sections, and possibly Theaetetus, the creator of solid geometry, were members of the Academy. Eudoxus of Cnidus—author of the doctrine of proportion expounded in Euclid's *Elements*, inventor of the method of finding the areas and volumes of curvilinear figures by exhaustion, and propounder of the astronomical scheme of concentric spheres adopted and altered by Aristotle—removed his school from Cyzicus to Athens for the purpose of cooperating with Plato; and during one of Plato's absences he seems to have acted as the head of the Academy. Archytas, the inventor of mechanical science, was a friend and correspondent of Plato. The Academy can thus be viewed as the link between the mathematics of the 5th-century Pythagoreans and that of the geometers and arithmeticians of Alexandria.

Nor were other sciences neglected. Speusippus, Plato's nephew and successor, was a voluminous writer on natural history; and Aristotle's biological works have been shown to belong largely to the early period in his career immediately after Plato's death, before the breach between the younger philosopher and the Academy. The comic poets found matter for mirth in the attention of the school to botanical classification. The Academy was particularly active in jurisprudence and practical legislation. As the biographer Plutarch testifies,

> Plato sent Aristonymus to the Arcadians, Phormion to Elis, Menedemus to Pyrrha. Eudoxus and Aristotle wrote laws for Cnidus and Stagirus. Alexander asked Xenocrates for advice about kingship; the man who was sent to Alexander by the Asiatic Greeks and did most to incite him to his war on the barbarians was Delios of Ephesus, an associate of Plato.

The creation of the Academy as a permanent society for the prosecution of both exact and humane sciences was, some assert, the first establishment of a university.

Formative influences. The most important formative influence to which the young Plato's mind was exposed was that of Socrates. It does not appear, however, that Plato, whose first ambitions were political, belonged as a "disciple" to the innermost circle of Socrates' intimates. In *Epistles* vii he is careful to speak of Socrates not as a "master" but as an older "friend," for whose character he has a profound respect; and he has recorded his own absence (through indisposition) from the death scene of the *Phaedo*. It would seem that his own vocation to philosophy dawned on him only afterward, as he reflected on the treatment of Socrates by the democratic leaders.

Political setting of Plato's life

Plato was also connected with Pyrilampes, a Periclean politician, and Critias, known as a democrat until, after the collapse of the Periclean government, his moral balance was upset and he became one of the most unscrupulous of the ensuing Thirty Tyrants. Plato's early upbringing in a family of Periclean politics having connection with Solon may explain why Plato's own estimate of democracy in the *Statesman* and *Laws* is much less unfavourable than that which he earlier ascribed to Socrates.

Plato's early experiences covered the disastrous years of the Deceleian War, the shattering of the Athenian empire, and the fierce civil strife of oligarchs and democrats in the year of anarchy, 404–403. He was too young to have known anything by experience of the imperial democracy of Pericles and Cleon or of the tide of the Sophistic movement. It is certainly not from memory that he depicted Protagoras, the earliest avowed professional Sophist, or Alcibiades, a turncoat Athenian general, as they were in their great days.

GENERAL FEATURES OF THE DIALOGUES

The canon and text of Plato was apparently fixed at about the turn of the Christian era. By reckoning the *Epistles* as one item, the list contained 36 works, arranged in nine tetralogies. No genuine work of Plato has been lost, and there is a general agreement of modern scholars that a number of small items—*Alcibiades I*, *Alcibiades II*, *Theages*, *Erastae*, *Clitopho*, *Hipparchus*, and *Minos*—are spurious. Most, though not all, contemporary scholars also regard the *Epinomis* as an appendix to the *Laws*, added by the mathematician Philippus of Opus. The *Hippias Major* and the *Menexenus* are regarded as doubtful by some, though Aristotle seems to have regarded them as Platonic. Plato's will is pretty certainly authentic.

Order of composition. Plato's literary career extended over the greater part of a long life. *The Apology* must have been written early, while the memory of Socrates' trial was still fresh. The *Laws*, on the other hand, was the work of an old man, and the state of its text bears out the tradition that Plato never lived to give it its final revision. Since there is no evidence that Plato began his career with a fully developed "system" that had only to be exposed, and since there is every logical reason to believe that his mind developed, the order in which the various dialogues were written takes on importance. Only through it can the development of Plato's thought be adequately charted. Plato himself has given a few clues to the order:

he linked the *Sophist* and the *Statesman* with the *Theaetetus* externally as continuations of the conversation reported in that dialogue. Similarly, he seems to have linked the *Timaeus* with *The Republic*. And Aristotle noted that the *Laws* was written after *The Republic*.

Stylometric studies of the dialogues

Modern scholars, by the use of stylistic criteria, have established that the *Sophist, Statesman, Philebus, Timaeus* (with its fragmentary sequel *Critias*), and *Laws* form a distinct linguistic group, belonging to the later years of Plato's life. The whole group must be later than the *Sophist*, which professes to be a sequel to the *Theaetetus*, which, in turn, can be dated with some accuracy, since it commemorates the death of the eminent mathematician after whom it is named (probably in 369 BC). The dialogue may thus be ascribed to *c.* 368, the eve of Plato's departure for Syracuse.

The earlier group is generally believed to have ended with the *Theaetetus* and the closely related *Parmenides*. Besides this conclusion, perhaps all that can be said with certainty is that the great outstanding dialogues, *Symposium, Phaedo*, and *The Republic* (and perhaps also *Protagoras*), in which Plato's dramatic power was at its highest, mark the culmination of this first period of literary activity. The comparative decline of dramatic power, accompanied by a compensating maturity of critical acumen, is the most striking contrast between the dialogues of the second period and those of the first period.

Persons of the dialogues. The great difficulty that initially besets the modern student is that created by the dramatic form of Plato's writings. Since Plato never introduced himself into his own dialogues, he is not formally committed to anything taught in them. The speakers who are formally bound by the utterances of the dialogues are their protagonists—Socrates, the Rationalist Parmenides of Elea (who held that all is a static One), and the Pythagorean Timaeus—and all of these are real historical persons. It is, therefore, reasonable to question whether or not Plato was putting his own doctrines into the mouths of these characters, whether his purpose was dogmatic and didactic or mainly dramatic, and whether he should be held responsible for what is said.

Whether Plato gives his own views

George Grote, an English historian, held that Plato allowed himself to develop freely in a dialogue any view that interested him for the moment, without pledging himself to its truth. Thus, Plato can make Socrates tolerate hedonistic utilitarianism in the *Protagoras* and denounce it in the *Gorgias*. Others argued that some of Plato's characters, notably Socrates and Timaeus, are "mouthpieces" through whom he inculcates tenets of his own, without concern for dramatic or historical propriety. Thus it often has been held that the "theory of Ideas," the doctrine of "recollection," and that of the tripartite soul were originated by Plato after the death of Socrates and consciously fathered on the older philosopher. In fact, neither extremity seems a reasonable view.

Thought of the earlier and later dialogues. There is undeniably a real difference between the thought of the dialogues that are later than the *Theaetetus* and those that are earlier. But there are no serious discrepancies of doctrine between individual dialogues of the same period. Plato may have announced his own personal convictions on certain doctrines in a second group of dialogues by a striking dramatic device: in the *Sophist* and *Statesman* the leading part is taken by an Eleatic (one belonging to a school of monistic Rationalists) and in the *Laws* by an Athenian, who are the only anonymous, indeed almost certainly the only imaginary, personages of any moment in the whole of Plato's writings. It seems likely, therefore, that these two characters have been left anonymous so that the writer may be free to use them as mouthpieces for his own teaching. Plato thus takes on himself the responsibility for the logic and epistemology of the *Sophist* and of the *Statesman* and for the ethics and educational and political theory of the *Statesman* and of the *Laws* in a specially marked way, thus evoking the question of how far the earlier utterances of Socrates are to be taken as his own.

Doctrine of Forms. Scholars have definitely established that there is a specific philosophical doctrine running through the earlier dialogues that has as its main features a theory of Forms (or of Ideas), the theory that knowledge is "recollection," and the theory of the "tripartite soul." In the *Phaedo*, for example, Socrates is made to describe the theory of Ideas as something quite familiar that he has for years constantly canvassed with his intimates (nearly all of whom were certainly living when the *Phaedo* was circulated). In the dialogues of the second period these tenets appear only in the mouth of Timaeus, a 5th-century Pythagorean older than Socrates; and the most important of them all, the theory of Forms, is actually made the object of what looks like a refutation in the *Parmenides*. The question thus arises as to whether Plato himself had two distinct philosophies, an earlier and a later, or whether the main object of the first group of dialogues was to preserve the memory of Socrates, with the philosophy expounded being, in the main, that of Socrates—coloured, no doubt, but not consciously distorted, in its passage through the mind of Plato. Thus, on the second view, Plato had no distinctive Platonic philosophy until a late period in his life. This interpretation remained viable for many years in spite of objections from many scholars. In its favour were several considerations long viewed as weighty.

Whether Plato has two philosophies

Socrates and Plato. It is significant that the only dialogue later than the *Theaetetus*, in which Socrates takes a leading part, is the *Philebus*, the one work of the second group that deals exclusively with the ethical problems on which the thought of Socrates had concentrated. This is usually explained by supposing that Plato was unwilling to make Socrates the exponent of doctrines that he knew to be his own property. On the other hand, it is hard to understand his misgivings if he had already been employing him in that very capacity for years. It is notable, too, that Aristotle, who apparently knew nothing of an earlier and a later version of Platonism, attributed to Plato a doctrine that is quite unlike anything to be found in the first great group of dialogues. It was also the view of Neoplatonic scholars that the theory of Ideas of the great earlier dialogues really originated with Socrates; and the fact that they did not find it necessary to argue the point may show that this had been the standing tradition of the Academy.

Today, however, a modified view is regarded as much more likely. Though the theory of Forms, in this view, could not have arisen with Socrates, Plato nevertheless attributed it to him because he saw it as the theoretical basis and elaboration of what Socrates did teach.

THE EARLIER DIALOGUES

Three main areas of thought

In *The Republic*, the greatest of all the dialogues that precede the *Theaetetus*, there are three main strands of argument deftly combined into an artistic whole, the ethical and political, the aesthetic and mystical, and the metaphysical. Other major dialogues belonging to this period give special prominence to some one of these three lines of thought; the *Phaedo* to the metaphysical theme; the *Protagoras* and the *Gorgias* to the ethical and political; the *Symposium* and the *Phaedrus* to the aesthetic and mystical.

Dialogues of search. The shorter dialogues, dealing with more special problems, generally of an ethical character, mostly conform to a common type: a problem in moral science, often that of the right definition of a virtue, is propounded, a number of tentative solutions are considered, and all are found to be vitiated by difficulties that cannot be dispelled so that the reader is left, at the end of the conversation, aware of his discreditable ignorance of the very things that it is most imperative for a man to know. He has formally learned nothing but has been made alive to the worthlessness of what he had hitherto been content to take as knowledge and to the need of seeking further enlightenment.

The effect of these dialogues of search is thus to put the reader in tune with the spirit of Socrates, who had said that the one respect in which he was wiser than other men was in his keen appreciation of his own ignorance of the most important matters. He learns the meaning of Socrates' ruling principle that the supreme business of life is to "tend" the soul and his conviction that "goodness of

soul" means *knowledge* of good and evil. The three dialogues directly concerned with the trial of Socrates have manifestly a further purpose. They are intended to explain to a puzzled public, as a debt of honour to his memory, why Socrates thought it a matter of conscience neither to withdraw from danger before trial, nor to make a conciliatory defense, nor, after conviction, to avail himself of the opportunity of flight.

Themes of the shorter dialogues

In the *Hippias Major* Plato propounds the question "What is the 'fine' (or 'beautiful')?" In the *Hippias Minor* he deals with the paradox that "wrong-doing is involuntary." In the *Ion* he discredits the poets, who create not "by science" but by a nonrational inspiration. The *Menexenus,* which professes to repeat a funeral oration learned from Aspasia, a famous *hetaira,* or member of a class of highly educated concubines, is apparently meant as a satire on the patriotic distortion of history. The *Charmides, Laches,* and *Lysis* are typical dialogues of search. The question of the *Charmides* is what is meant by *sōphrosynē,* the virtue that is shown both in graceful command of the passions and in dutiful behaviour to parents and official "superiors," in balance and sanity amid the turns of fortune. It seems that this virtue can be identified with "knowledge of self"—the self-knowledge that Socrates had valued so highly—when man is confronted with an ambiguity. The *Laches* is concerned with valour, the soldier's virtue; and the *Lysis* examines in the same tentative way, friendship, the relation in which self-forgetting devotion most conspicuously displays itself.

The question of whether names are significant by nature or by convention is considered in the *Cratylus—i.e.,* whether there is some special appropriateness of the sounds of names to the objects called by them, or whether it merely reflects the "usage of the community." Thus Plato shows that language is an instrument of thought; that the test of its rightness is not mere social usage, but its genuine capacity to express true thought accurately. The dialogue *Euthydemus* satirizes the "eristics"—those who try to entangle a person in fallacies because of the ambiguity of language. Its more serious purpose, however, is to contrast this futile contradiction mongering with the "protreptic," or hortatory, efforts of Socrates, who urges that happiness is not guaranteed by the *possession* of things but by the right *use* of them—and particularly of the gifts of mind, body, and fortune.

Ethical and political dialogues. The *Gorgias,* the *Protagoras,* and several lesser dialogues give special prominence to ethical and political themes.

Gorgias. The *Gorgias* begins ostensibly as an inquiry into the nature and worth of rhetoric, the art of advocacy professed by Gorgias, and develops into a plea of sustained eloquence and logical power for absolute right—as against expediency—as the sovereign rule of life, both private and public. It ends with an imaginative picture, on Orphic lines, of the eternal destinies of the righteous and of the unrighteous soul. Literature contains no more impressive presentation of the claim of conscience to unqualified obedience and the impossibility of divorcing the politically right from the morally right.

Gorgias holds that "rhetoric" is the queen of all "arts." If the statesman skilled in rhetoric is clever enough, he will, though a layman, carry the day even against the expert specialist. Socrates, on the other hand, declares that "rhetoric" is not an "art" but a mere "empiric knack" of humouring the prejudices of an audience. There are two arts conducive to health of soul, those of the legislator and of the judge. The Sophist counterfeits the first, as the rhetorician does the second, by taking the pleasant instead of the good as his standard. The rhetorician is thus not the wise physician of the body politic but its toady. This severe judgment is disputed by Polus, an ardent admirer of Gorgias, on the ground that the successful rhetorician is virtually the autocrat of the community, and to be such is the summit of human happiness because he can do whatever he likes.

Autocratic power of rhetoric and its refutation

Socrates rejects this view. To suffer a wrong is an evil, but to inflict one is much worse. If rhetoric is of real service to men, it should be most of all serviceable to an offender, who would employ it to move the authorities to inflict the penalties for which the state of his soul calls. This is denied by Callicles of Acharnae, an otherwise unknown politician, who proceeds recklessly to develop the doctrine of the "will to power." It may be a convention of the herd that unscrupulous aggression is discreditable and wrong, but "nature's convention" is that the strong are justified in using their strength as they please, while the weak "go to the wall." To Socrates, however, the creators of the imperialistic Athenian democracy were no true statesmen; they were the domestic servants of the democracy for whose tastes they catered; they were not its physicians. That would be a condition like that of the fabled sinners who are punished in Hades by being set to spend eternity in filling leaking pitchers. The truly happy life is not, as Callicles supposes, to have vehement passions and be able to gratify them to the full; it is to have that measure in which the gratification of desire is strictly regulated by regard for justice and *sōphrosynē* (see above).

Meno. The *Meno* is nominally concerned with the question of what virtue is and whether it can be taught; but it is further interesting for two reasons: it states clearly the doctrine that knowledge is "recollection"; it also introduces as a character the democratic politician Anytus, the main author of the prosecution of Socrates.

Whether virtue can be taught

Whether virtue can be taught depends on what virtue is. Any definition, however, is futile because it is idle to inquire into what you already know, useless to inquire into what you do not know, for you could not recognize the unknown, even if you found it. But this difficulty would vanish if the soul were immortal and had long ago learned all truth, so that it needs now only to be reminded by sense experiences of truths that it once knew and has forgotten. Thus a youth who has never studied geometry can be brought to recognize mathematical truths. He produces the right answer "out of himself." The point is thus the presence of a priori elements in mathematical truth. But if virtue is knowledge, there must be professional teachers of it. But Anytus insists that the Sophists, who claim to be such professionals, are mischievous impostors; and even the "best men" have been unable to teach it to their own sons. If a man should arise who could actually teach statesmanship, he would be one who really *knew* what the Good is.

Protagoras. The *Protagoras,* a finely dramatic dialogue, gives the most complete presentation to be found in Plato of the main principles of the Socratic morality. In this dialogue, Socrates meets the eminent Sophist Protagoras, who explains that his profession is the "teaching of goodness"—*i.e.,* the art of making a success of one's own life and of one's city. Socrates urges, however, that public opinion and the practice of the eminent few, who have never taught "goodness" to their sons, alike suggest that the conduct of life is not teachable. But the problem arises as to whether the various commonly recognized virtues are really different or all one. Protagoras is ultimately ready to identify all of them but courage with wisdom or sound judgment. Socrates then attempts to show that, even in courage, goodness is meaningful inasmuch as, by facing pain and danger, one escapes worse pain or danger. Thus all virtues can be reduced to the prudent computation of pleasures and of pains, and virtue, according to Socrates, must then be one, namely, knowledge; and wrongdoing is thus involuntary. The true explanation of Socrates' doubts, however, is that, though true virtue, being knowledge, is teachable, what Protagoras is trying to teach is not true virtue. Success depends on personal tact, and tact cannot be learned from an instructor.

Euthyphro, The Apology, Crito. The main purpose of these three works, which deal with the bearing of Socrates before, during, and after his trial, is to obviate possible misunderstandings of his position and motives; their theme is the true meaning and importance of the care or tending of the soul. The problem of the *Euthyphro* is what is religion. The respondent Euthyphron is a kind of Orphic sectary. The central, though tacit, point is that the service of God, which in religion means cooperation with him, lies not in the fulfillment of purely arbi-

trary commands but in the production of a noble work, presumably that of the "tending of the soul." Consideration of *The Apology* and *Crito* in detail belongs to the study of Socrates (*q.v.*) rather than to that of Plato.

Metaphysical foundation of Plato's doctrine: "Phaedo." In the works so far considered, the foundation of a Socratic moral and political doctrine is laid, which holds that the great concern of man is the development of a rational moral personality and that this development is the key to man's felicity. Success in this task, however, depends on rational insight into the true scale of good.

Need for insight into absolute Good

The reason why men forfeit felicity is that they mistake apparent good for real, the conditionally for the absolutely Good. If a man ever knew with assurance what absolute Good is, he would in practice never pursue anything else; it is in this sense that "all virtue is knowledge." The philosophical moralist, who has achieved an assured insight into absolute Good, is thus the only true statesman, for he alone can tend to the national character. These convictions clearly imply a far-reaching metaphysic as their foundation and justification. The principles of this metaphysic are expounded more explicitly in the dialogues now to be considered, in which an explicit theory of knowledge and of scientific method are also discernible.

The object of the *Phaedo* is to justify faith in immortality by showing that it follows from a fundamental metaphysical doctrine (the theory of Ideas, or the doctrine of Forms), which seems to afford a rational clue to the structure of the universe. The writer has a passionate faith in "personal immortality." What is being maintained, in fact, is the divinity of the soul, of which its survival of death is a consequence. The argument is, briefly, that a true philosopher's whole life has been spent in trying to liberate the soul from dependence on her body. In life, the body is always interfering with the soul's activity. Its appetites and passion interrupt her pursuit of wisdom and goodness.

Arguments for immortality

There are three reasons for thinking that the soul survives death.

1. There is a belief that the soul has a succession of many lives. The processes of nature in general are cyclical; and it is reasonable to suppose that this cyclicity applies to the case of dying and coming to life. If this were not so, if the process of dying were not reversible, life would ultimately vanish from the universe. Secondly, the doctrine that what men call "learning" is really "recollection" suggests that men must have become acquainted with this knowledge before they were confined to their bodies.

2. The soul is relatively immutable, like the fixed ideal standards or eternal objects that she contemplates in her scientific thinking; and she, herself, has the likeness of that which she contemplates. Thus the divine element in man, the soul, can resist destruction. It is true that the soul could be an epiphenomenon, the tune (*harmonia*) given out by the body, which will vanish when the instrument is broken; but Socrates replies that what can be more or less in tune is clearly not itself a tune.

3. Socrates' thought was permeated by a fundamental unproved postulate, the theory of Ideas, which holds that there really is a single determinate and immutable something answering to every significant general, or abstract, term and apprehended only by pure thought. The sensible things of which a man predicates general terms temporarily partake in or communicate with the Idea or Form. When he says that a thing becomes beautiful, what he means is that the Form "beauty" begins to be present to that thing. This is the true account of the cause of "coming into and passing out of being." As applied to immortality, Socrates argues first that there are certain sensible things of which it is an essential character to partake of a given Form. Such things will never admit an incompatible Form. Similarly, it is an essential character of a soul to be alive, to partake of the Form life. It refuses to partake of the Form death. The proof of immortality is thus hypothetical; it is shown to be involved as a consequence of the doctrine of Forms. (For further discussion of the theory of Forms, see METAPHYSICS.)

Aesthetic and mystical dialogues. Both the *Symposium* and the *Phaedrus* present the Forms in a special light, as objects of mystical contemplation and as stimuli of mystical emotion.

Symposium. The immediate object of the *Symposium*, which records several banquet eulogies of *eros* (desirous love), is to find the highest manifestation of the love that controls the world in the mystic aspiration after union with the eternal and supercosmic beauty. It depicts Socrates as having reached the goal of union, and puts the figure of Alcibiades, who has sold his spiritual birthright for the pleasures of the world, in sharp opposition to him.

Ascent to the supreme Form of Good

The main argument may be summarized thus: *Eros* is a reaching out of the soul to a hoped-for good. The object is eternal beauty. In its crudest form, love for a beautiful person is really a passion to achieve immortality through offspring by that person. A more spiritual form is the aspiration to combine with a kindred soul to give birth to sound institutions and rules of life. Still more spiritual is the endeavour to enrich philosophy and science through noble dialogue. The insistent seeker may then suddenly descry a supreme beauty that is the cause and source of all the beauties so far discerned. The philosopher's path thus culminates in a supreme beatific vision of the "Form of Good," the supreme Form that stands at the head of all others.

Phaedrus. Though the immediate subject of the *Phaedrus* is to show how a really scientific rhetoric might be built on the double foundation of logical method and scientific study of human passions, Plato contrives to unite with this topic a discussion of the psychology of love, which leads him to speak of the Forms as the objects of transcendent emotion and, indeed, of mystical contemplation. The soul, in its antenatal disembodied state, could enjoy the direct contemplation of unbodied reality; *i.e.*, of the Forms. Sense experience can suggest the Form of beauty in an unusually startling way: through falling in love. The unreason and madness of the lover mean that the wings of his soul are beginning to grow again; it is the first step in the soul's return to its high estate.

The consummate dialogue: "The Republic." As earlier stated, it is in *The Republic* that Plato interweaves the ethicopolitical, the aesthetic and mystical, and the metaphysical strands of his thought. In *The Republic*, the immediate problem is strictly ethical. What is justice? Can it be shown that justice is always a boon, injustice a curse, apart from all consideration of consequences? Is there a rational and self-authenticating principle at the root of moral distinctions? Plato holds that there is. Each man has a specific contribution that he can make to a rational society. Morality, justice, is to discharge that vocation. For to live otherwise is to be spiritually diseased.

This leads to a sketch of the institutions of a society in which justice holds sway and, particularly, of its moral and religious education. To show, next, that the principles that regulate public life are also recognizably those of the great virtues of private life, Plato provides, through the doctrine of the tripartite soul, a psychology of voluntary action. Moral excellence is thus grounded in absolute loyalty to a sound moral tradition enforced by education —a goal to be assured through statesmen who are philosophers as well as kings.

But the way for the vision of the Good must be prepared by an intellectual discipline in hard thinking that leads through study of the exact sciences to that of their metaphysical principles. The central books of *The Republic* thus present an outline of metaphysics and a philosophy of the sciences. As the various stages of degeneration through which national and personal character pass when the true moral-ideal decays are reviewed, one is confirmed in his conviction that, in respect of happiness, the life of regard for right is immeasurably superior, and this conclusion is finally clinched by reaffirmation of the immortality of the soul, which makes the choice of infinite moment.

The virtues: in individual and state

In the ethical scheme of *The Republic*, "three lives" are distinguished: those of the philosopher, of the votary of enjoyment, and of the man of action. The end of the first is wisdom, of the second, the gratification of appetite,

and of the third, practical distinction. These reflect the three elements or active principles within a man: (1) considered rational judgment of good; (2) a multitude of clamant and conflicting appetites for particular gratifications; and (3) a factor of spirited higher ideal emotion, manifested as resentment against infringements both by others and by our own appetites.

The same three distinctions reappear in the structure of society; there are: (1) the statesmen; (2) the general civilian population, who provide for material needs; and (3) the executive force (army and police). These three orders are thus, respectively: the judgmental, appetitive, and spirited elements. The four cardinal virtues are thus: wisdom, the excellence of the thinking part; courage, that of the spirited part (loyalty to the rule of life laid down by judgment); temperance, that of the appetitive part (acquiescence of the nonrational elements to the plan of life prescribed by judgment); and justice—the state in which each of the elements is vigorously executing its own function and, in loyal contentment, confining itself within its limits. Such a society is a true aristocracy, or rule of the best. Successive deviations from the ideal are timocracy (the benign military state), oligarchy (the state dominated by merchant princes, plutocracy), democracy (the state run by the inexpert) and, worst of all, tyranny (the state subjected to an irresponsible or criminal will).

Philosophic rulers: their virtues and training

The tripartite psychological schema here reflected proves inadequate when applied to the moral life of the philosopher-king, of whom a higher level of moral goodness is demanded. His courage, for example, is declared to be no mere loyalty to right opinions but a high serenity arising from knowledge of the insignificance of man's brief individual life in a vast universe. For him, in fact, "all virtue is one thing, knowledge." In the ideal state itself, however, virtue is not a complete unity. A special demand may be made, for example, of the fighting force, which is the valour of the whole society. This is because their moral convictions arise not from personal insight but from opinions implanted by education and taken on trust. But the rulers must know with a personal knowledge, by their insight into a system of absolute values embodied in the very structure of the universe. The virtues are thus transubstantiated and can no longer be distinguished from one another; they fuse in knowledge of the Good.

In *The Republic* the Forms appear in the double character of objects of all genuine science and formal causes of events and processes. It is expressly denied that there can be knowledge, in the proper sense, of the temporal and mutable. In Plato's scheme for the intellectual training of the philosophic rulers, the exact sciences—arithmetic, plane and solid geometry, astronomy, and harmonics—are first studied for ten years to familiarize the mind with relations that can only be apprehended by thought. Five years are then further given to the still more severe study of dialectic, a study which proceeds "by means of Forms, through Forms, to Forms"—what in the 1970s would be called a critical metaphysic of the sciences, which examines the *hypotheses* seeking to discover some still more ultimate, really self-evident principle from which they follow as consequences.

The Form of Good

This principle is the Form of Good, which, like the sun in relation to visible things, is the source at once of the reality of all things and of the light by which they are apprehended, as well as of their value. As in the *Symposium*, the Good is the supreme beauty that dawns suddenly upon the pilgrim of love as he draws near to his goal. It is true that the Good holds the place taken in later philosophies by the God who is thought of as the "Light of the world." But it is not God. Though *The Republic* is permeated by religious faith, theism as a principle of metaphysical explanation only makes its appearance in Plato's latest dialogues.

Dialogues of critical reconstruction. The two works that probably anticipate the dialogues of Plato's old age, the *Parmenides* and *Theaetetus*, display a remarkable difference of tone, clearly the result of a period of fruitful critical reconstruction.

The theory expounded in the *Phaedo* and *The Republic* does not allow enough reality to the sensible world. They tell the reader that a sensible thing is a complex or meeting place of a plurality of Forms: what else it may be they do not say. Clearly, however, the relation between a thing and a Form (*e.g.*, beauty), which has been called participation, needs further elucidation. In these dialogues truths of fact, of the natural world, have not yet come by their rights.

Plato clearly had an external motive for the re-examination of his system as well. The *Parmenides*, the *Theaetetus*, and the *Sophist* all reveal a special interest in the Eleatic philosophy—of which Parmenides was the chief representative—and the first and third show an anxiety on Plato's part to maintain that, in spite of important divergences, he, and not the professed Eleatics, is the true spiritual heir of Parmenides. The doctrine of his friend Eucleides of Megara, like that of Parmenides, was that sensible appearances are illusions with no reality at all. Continued reflection on this problem thus led straight to the discussion of the meaning of the copula "is" and the significance of the denial "is not," which is the subject of the *Sophist*.

Parmenides. Formally, the *Parmenides* leads to an impasse. In its first half the youthful Socrates expounds the doctrine of the participation of things in Forms to the Eleatic philosophers Parmenides and Zeno as the solution of the problem of the one and the many. Parmenides raises what appear to be insoluble objections and hints that the helplessness of Socrates under his criticism arises from insufficient training in logic.

In the second half, Parmenides gives an example of the logical training that he recommends. He takes for examination his own thesis, "The one is," and constructs an elaborate set of antinomies, or contradictions, thus yielding a patently ironical conclusion.

Objections to the doctrine of Forms

The Eleatic objections to the doctrine of the participation of Form in thing are, first, that it does not really reconcile unity with plurality, since it leads to a perpetual regress. It says that the many things that have a common predicate, or characteristic, participate in or imitate a single Form. But the Form itself also admits of a common predicate, and therefore a second Form must exist, participated in alike by the sensible things and the first Form, and so on, endlessly. Secondly, a graver difficulty is that the relations between Forms must belong to the realm of Forms, and those between sensible things to the realm of things. Thus men, belonging to the second, can know nothing of the true realities, the Forms. The purpose of the objections is thus to suggest that the manifold of sense has not even a derivative reality; it is mere illusion. The inference is that Plato is reproducing Eleatic and Megarian criticisms of the doctrine ascribed by himself to Socrates. The purpose of the antinomies may be deliberate parody, to show that the methods of the Megarian logicians are even more damaging to themselves than to the doctrine of participation.

Theaetetus. The *Theaetetus* is a straightforward discussion of the question of how knowledge should be defined. Plato shows (1) that knowledge cannot be identified with sensation nor with any formless simple apprehension; and (2) that pure relativism is as impossible in epistemology as in metaphysics. The increasing value that Plato is coming to put upon natural knowledge is marked by the use of the word *doxa* (uncertain opinion) in the new sense of "judgment." The *Theaetetus* strikingly treats knowledge at length without making any reference to the Forms or to the mythology of recollection. It remains to this day the best of introductions to the problem of knowledge. The main argument is as follows:

Efforts at defining knowledge

1. It seems plausible to say that knowledge is sensation. "What seems to me is so to me; what seems to you is so to you" (Protagoras). This relativistic doctrine arrives at a point at which the only reality, both within and without, is motion—a view clearly derived from certain philosophers of change presumably influenced by Heracleitus. When the motions of both organ and environment impinge on one another, they give rise to the twin product, felt sensation–sensible quality.

There is thus no common perceived world and therefore no standard of truth or reality other than that of the individual percipient.

Even the relativists, however, do not hold that each man is the only and inerrant measure of his *future* sensations. A man's own opinion as to whether a certain course will be expedient or good for him is often far from being the soundest. What the soul perceives through bodily organs must be distinguished from what she apprehends by herself without organs—such as number, sameness, likeness, being, and good. Since these comprise the formal element in all knowledge, knowledge must be found not in sensations but in the judgment of the mind upon them.

2. Is knowledge, then, true judgment? The answer is "hardly," for the statement implies a knowledge of what is meant by error, for which Plato cannot give a satisfactory account.

3. Finally, is knowledge "true judgment accompanied by discourse, true judgment for which a man can give grounds"? This doctrine has difficulties of its own; and, in any case, the proposed definition amounts to saying that knowledge is true judgment about an object plus *knowledge* of its *differentia* and so is circular.

Because Plato's argument nowhere appeals to his favorite doctrine of Forms and because the dialogue ends so inconclusively, it is likely that Plato wanted to show that the problem of knowledge is insoluble without the Forms.

THE LATER DIALOGUES

Sophist and Statesman. Formally the important dialogues, the *Sophist* and the *Statesman*, are closely connected, both being ostensibly concerned with a problem of definition. The real purpose of the *Sophist*, however, is logical or metaphysical; it aims at explaining the true nature of negative predication, or denials that something is so, and thus at disposing of the Eleatic thesis that the sensible realm, containing, as it does, a negative moment, must be mere unreal illusion. The object of the *Statesman*, on the other hand, is to consider the respective merits of two contrasted forms of government, personal rule and constitutionalism, and to recommend the second, particularly in the form of limited monarchy. The *Sophist* thus lays the foundations of all subsequent logic, the *Statesman* those of all constitutionalism. A more temporary purpose in both dialogues is to illustrate the value of careful classification as a basis for scientific definition.

In the *Sophist*, the Sophistic criticism of any definition couched in terms of falseness is presented. The Sophists argued that there can be no such thing as a false statement or a false impression, for the false means "what is not," and "what is not" is nothing at all and can neither be uttered nor thought. Plato argues, on the other hand, that "what is not, in some sense also is," and "what is, in some sense is not"; there are, that is, significant negative propositions. Plato refutes Parmenidean monism by drawing the distinction between absolute and relative nonbeing. A significant denial, *A is not B*, does not mean that *A* is nothing, but that *A* is other than *B*; every one of the great categorical features of reality—being, identity, difference, motion, and rest—is other than every other feature, and the true business of dialectic is to treat the Forms themselves as an interrelated system, with relations of compatibility and incompatibility among themselves. Negation is a moment in the system of intelligible reality, and therefore its presence in the sensible realm does not stamp that realm as illusion.

Absolute versus relative nonbeing

In the *Statesman* the conclusion is reached that government by a benevolent dictator is not suitable to the conditions of human life, in which the direction is not that of a god. The surrogate for direction by a god is the impersonal supremacy of inviolable law. Where there is such law, monarchy is the best and democracy the least satisfactory form; but where there is no law, this situation is inverted. A sovereign democracy is preferable to an irresponsible autocrat.

Philebus. The *Philebus* contains Plato's ripest moral psychology. Its subject is strictly ethical—the question of whether the Good is pleasurable feeling (hedonism) or thought (the exercise of intelligence). Under the guidance of Socrates a mediating conclusion is reached: the best life contains both elements, but intelligence predominates.

Philosophically most important is a classification adopted to determine the formal character of the two claimants to recognition as the good. All components of the actual belong to one of four classes: (1) the infinite or unbounded, (2) the limit, (3) the mixture (of infinite and limit), (4) the cause of the mixture. Inasmuch as, in the *Philebus*, all of the good things of life belong to the third class, that is, are produced by imposing a definite limit upon an indeterminate continuum, this dialogue is the immediate source of the famous doctrine of the mean. The establishment of such a ratio is a genesis into being, and the cause in such a process is always intelligence, human or divine.

Timaeus. The *Timaeus* is an exposition of cosmology, physics, and biology. Timaeus first draws the distinction between eternal being and temporal becoming and insists that it is only of the former that one can have exact and final knowledge. The visible, mutable world had a beginning; it is the work of God, who had its Forms before him as an eternal model in terms of which he molded the world as an imitation. God first formed its soul out of three constituents: identity, difference, being. The world soul was placed in the circles of the heavenly bodies and the circles were animated with movements. Subsequently, the various subordinate gods and the immortal and rational element in the human soul were formed. The human body and the lower components of its soul were generated through the intermediacy of the "created gods" (*i.e.*, the stars).

The forming of the world and things

The *Timaeus* combines the geometry of the Pythagoreans with the biology of Empedocles, the founder of Sicilian medicine, by a mathematical construction of the elements, in which four of the regular solids—cube, tetrahedron, octahedron, and icosahedron—are assumed to be the shapes of the corpuscles of earth, fire, air, and water. (The fifth, the dodecahedron, comprises the model for the whole universe.)

Among the important features of the dialogue are its introduction of God as the intelligent, efficient cause of all order and structure in the world of becoming and the emphatic recognition of the essentially tentative character of natural science. It is also noteworthy that, though Plato presents a corpuscular physics, his metaphysical substrate is not matter but *chōra* (space). The presence of space as a factor requires the recognition, over and above God or mind, of an element that he called *anankē* (necessity), a name for the fact that there is always in the actual an irreducible remainder of brute datum that cannot be rationalized completely and that one has to be content to accept. Its function is everywhere to be instrumental to the intelligent and beneficent purpose of God. It is the expression, however, of a rational faith in Providence and in the supremacy of the moral order. The details of the cosmology, physiology, and psychophysics are of great importance for the history of science but metaphysically of secondary interest.

Laws and Epinomis. The *Laws*, Plato's longest and most intensely practical work, contains his ripest utterances on ethics, education, and jurisprudence, as well as his one entirely nonmythical exposition of theology. The immediate object is to provide a model of constitution making and legislation to assist in the actual founding of cities.

Plato's attempt to do work of this kind at Syracuse had given the Academy a recognized standing as a school of scientific politics and jurisprudence.

The practical level of the *Laws*

The problem is thus not the construction of an ideal state as in *The Republic*, but the framing of a constitution and code that might be successfully adopted by a society of average Greeks. Hence the demands made on average human nature, though exacting, are not pitched too high; thus the communism of *The Republic* is dropped.

Purely speculative philosophy and science are excluded from its purview; and the metaphysical interest is introduced only so far as to provide a basis for a moral theol-

ogy. In compensation, the *Laws* is exceptionally rich in political and juristic wisdom and appears, indirectly, to have left its mark on the great system of Roman jurisprudence.

The ethical ideal is still that familiar from earlier dialogues. All unnatural vices shall be completely suppressed and monogamous marriage with strict chastity shall be the rule. (In *The Republic*, however, at least the guardian class enters into temporary unions or "sacred marriages," with community of wives and children, to foster a concern for the common good.) In politics, Plato favours a mixed constitution, a balance between *eleutheria* (popular control) and *monarchia* (personal authority), and suggests a system to secure both genuine popular representation and the proper regard for personal qualifications. The basis of society is to be agriculture, not commerce. What amounts to a supertax of 100 percent is to be levied on incomes beyond the statutory limits. Education is regarded as the most important of all the functions of government. The distinction between the sexes is to be treated as irrelevant.

Careful attention is to be paid to the right utilization of the child's instinct for play and to the demand that the young shall be taught in institutions where expert instruction in all the various subjects is coordinated (the origin of the grammar school). Members of the supreme council of the state shall be thoroughly trained in the supreme science, which "sees the one in the many and the many in the one"; *i.e.*, in dialectics.

Plato instituted regulations that would ensure that trials for serious offenses would take place before a court of highly qualified magistrates and would proceed with due deliberation. Also, provision was made for appeals, and a foundation was laid for the discrimination between civil and criminal law.

There are passages showing that Plato believed in the Earth's motion. In the *Laws* it is said that the real orbit of each planet is a single closed curve; in the *Epinomis* the view that the circle of the stars communicates its motion to those of the planets is rejected. What Plato asserts is that each planet has only one proper revolution; the remaining revolutions are apparent, not real. The implication is that the latter are reflections of a motion of the Earth itself. The diurnal movement, however, would not be a rotation on an axis, but a diurnal revolution around a centre, apparently a Pythagorean "central fire."

Plato's natural theology

In the *Laws* Plato, for a practical purpose, creates a natural theology. There are three false beliefs, he holds, that are fatal to moral character: atheism, denial of the moral government of the world, and the belief that divine judgment can be bought off by offerings. Plato claims that he can disprove them all. His refutation of atheism turns on the identification of the soul with the "movement which can move itself." Thus all motion throughout the universe is ultimately initiated by souls. It is then inferred from the regular character of the great cosmic motions and their systematic unity that the souls that originate them form a hierarchy with a best soul, God, at their head. Since some motions are disorderly, there must be one soul that is not the best, and there may be more. (There is no suggestion, however, that there is a worst soul, a devil.) This is Plato's way of excluding pantheism, as incompatible with the reality of evil. The other two heresies can be similarly disposed of. Plato thus becomes the originator of the view that there are certain theological truths that can be strictly demonstrated by reason; *i.e.*, of philosophical theology. Plato goes on to enact that the denial of any of his three propositions shall be a grave crime. He is thus the inventor of the proposal to make an official creed for the state and to treat dissent from it as criminal. Plato's last word, then, on the problem of how the sensible comes to partake of form is that it does so through the agency of divine goodness and wisdom, as God molds the sensible upon the pattern of the intelligible Forms.

ARISTOTLE'S ACCOUNT OF PLATO'S THOUGHT

Since Plato refused to write his own metaphysic, knowledge of its final shape has to be derived from statements by Aristotle, according to whom Plato's doctrine of Forms was, in its general character, not different from Pythagoreanism, the Forms being actually called numbers. The two points on which Aristotle regards Plato as disagreeing with the Pythagoreans are: (1) that, whereas the Pythagoreans said that numbers have as their constituents the unlimited and the limit, Plato taught that the Forms have as constituents the one and the great-and-small; and (2) that, whereas the Pythagoreans had said that things are numbers, Plato intercalated between his Forms (or numbers) and sensible things an intermediate class of mathematicals.

Number-Forms: the one and the great-and-small

Light is thrown on the statement that the Forms are numbers if it is recalled that in the *Timaeus* the behaviour of bodies is explained by the geometrical structure of their corpuscles, and the corpuscles themselves are analyzed into complexes of elementary triangles, and a triangle may be regarded as a triplet of numbers. Moreover, according to Aristotle, the mathematicals that are intermediate between the Forms and sensible things, like the geometer's figure, differ from Forms because they are many (whereas the Form is one) and differ from sensible things in being unchanging. The relations of numbers are thus the key to the whole mystery of nature, as is actually said in the *Epinomis*.

There is also a difficulty about the notion of a number as a collection of units, which must have been forced on Plato's attention by his manifest interest in irrational numbers (not expressible as a fraction), which cannot possibly be reached by any process of forming collections of units. Thus numbers are determinations of a continuum (a great and small) by a limit. The one, however, can no longer be regarded as a blend of unlimited and limit but must be, itself, the factor of limit. Thus Plato's way of generating numbers is intended to permit the inclusion of rationals and irrationals in the same series.

The influence of Plato can be traced throughout Western culture from Aristotle through St. Augustine, from Pascal to A.N. Whitehead, an eminent 20th-century metaphysician who declared: "The safest general characterization of the European philosophical tradition is that it consists of a series of footnotes to Plato" (see PLATONISM AND NEOPLATONISM).

MAJOR WORKS

Apologia Sokratous (*c.* 399 BC).

THE DIALOGUES: *Hippias Elasson* (Major—doubtful); *Laches; Charmides; Ion; Protagoras; Euthyphron; Kriton; Gorgias; Menon; Lysis; Menexenos* (doubtful; *c.* 386 BC); *Euthydemos; Kratylos; Symposion* (*c.* 384 BC); *Phaidon; Politeia* (*The Republic*); *Parmenides* (*c.* 370 BC); *Theaitetos* (*c.* 368 BC); *Phaidros; Sophistes* (360 BC or later); *Politikos; Philebos; Timaios; Kritias; Nomoi* (*Laws*).

EPISTLES: There are 13 epistles, *Thirteen Epistles of Plato* (1925), attributed by some to Plato, but their authenticity is open to considerable doubt. The most important is *Epistles vii* concerning Plato's visit to Sicily (367 BC) and his relations with Dion and the tyrant Dionysius.

The Apology and the Dialogues are translated in the "Loeb Series." Several of the Dialogues are translated in the "Penguin Classics Series."

BIBLIOGRAPHY. Plato's role and context in the growth of Greek philosophy are concisely described by F.M. CORNFORD, *Before and After Socrates* (1932); and W.K.C. GUTHRIE, *The Greek Philosophers from Thales to Aristotle* (1950). His relations with the Sophists and Socrates are sketched in G.C. FIELD, *Plato and His Contemporaries* (1930); W.K.C. GUTHRIE devotes *A History of Greek Philosophy*, vol. 3, *The Fifth-Century Enlightenment* (1969), entirely to a detailed analysis of the Sophists and Socrates, as background for understanding Plato (vol. 4, forthcoming, concentrates on Plato). E.A. HAVELOCK, *Preface to Plato* (1963), traces the linguistic developments that furnished the terminological context of the dialogues.

Two still influential introductions are A.E. TAYLOR, *Plato: The Man and His Work*, 7th ed. (1960); and G.M.A. GRUBE, *Plato's Thought* (1935, reprinted 1958). More recent syntheses are G.C. FIELD, *The Philosophy of Plato* (1949); and R.S.H. BLUCK, *Plato's Life and Thought* (1949). PAUL FRIEDLANDER, *Platon*, 2 vol. (1928–30; Eng. trans., 3 vol., 1958–69), represents continental scholarship on Plato as myth-maker, cosmologist, and city planner. See also ALEXANDRE KOYRE, *Introduction à la lecture de Platon* (1945; Eng. trans., *Discovering Plato*, 1945); and J. STENZEL, *Studien zur*

Entwicklung der platonischen Dialektik von Socrates zu Aristotles (1917; Eng. trans., *Plato's Method of Dialectic*, 1940), a sound textual account; K.M. SAYRE, *Plato's Analytic Method* (1969), suggests an analogy between Plato and recent analysis. Stages of methodic development are distinguished in RICHARD ROBINSON, *Plato's Earlier Dialectic*, 2nd ed. (1953); the mature ontology and epistemology of the *Theaetetus* and the *Sophist* are presented in W.G. RUNCIMAN, *Plato's Later Epistemology* (1962). The strong mathematical component in Plato is examined formally by A. WEDBERG, *Plato's Philosophy of Mathematics* (1955); and in concrete applications by R.S. BRUMBAUGH, *Plato's Mathematical Imagination* (1954, reprinted 1968). Philosophy as a cure of souls is the theme of R.E. CUSHMAN, *Therapeia: Plato's Conception of Philosophy* (1958).

I.M. CROMBIE, *An Examination of Plato's Doctrines*, 2 vol. (1962–63), makes an in-depth analysis, furnishing the ethical-social-aesthetic-religious context for the more abstract issues of knowing and reality. W.F.R. HARDIE, *A Study in Plato* (1936), focusses on his evolving epistemology and metaphysics; W.D. ROSS, *Plato's Theory of Ideas* (1951), emphasizes the gradual transcendence of Forms; GILBERT RYLE, *Plato's Progress* (1966), criticizes the dialectic; and J.E. RAVEN, *Plato's Thought in the Making* (1965), gives balanced attention to the great middle dialogues. R.E. ALLEN (ed.), *Studies in Plato's Metaphysics* (1965); and R. BAMBROUGH (ed.), *New Essays on Plato and Aristotle* (1965), present collections of metaphysical and epistemological studies.

JOHN GOULD, *The Development of Plato's Ethics* (1955), furnishes a good conspectus; the religious aspects are studied by F. SOLMSEN, *Plato's Theology* (1942). A.W. GOULDNER, *Enter Plato* (1965), covers Plato's social analysis and legislation, with its tension between detachment and engagement. Plato's mixed influence on the history of freedom and authoritarianism is debated by KARL R. POPPER, *The Open Society and Its Enemies*, 5th ed. rev., 2 vol. (1966); and R.B. LEVINSON, *In Defense of Plato* (1953). Papers on the controversy are in R. BAMBROUGH (comp.), *Plato, Popper and Politics* (1967); and T.L. THORSON (ed.), *Plato: Totalitarian or Democrat?* (1963).

The Plato translations made by F.M. CORNFORD, R. HACKFORTH, and J. SKEMP also include commentaries. Other commentaries include: A. SESONSKE and N. FLEMING (eds.), *Plato's Meno: Text and Criticism* (1965), with essays on Socratic paradox, definition, and education; R.E. ALLEN, *Plato's "Euthyphro" and Earlier Theory of Forms* (1970), a thorough introduction to problems of the Forms; N.R. MURPHY, *The Interpretation of Plato's Republic* (1951), bringing out its many conceptual problems and pitfalls; A.E. TAYLOR, *A Commentary on Plato's Timaeus* (1928), which (together with Cornford's later commentary) examines his cosmology and myths; G.R. MORROW, *Plato's Cretan City* (1960), analyzes the social situation, internal organization, and general principles of Plato's *Laws;* a companion commentary on Plato's role in Sicilian politics appears in G.R. MORROW's translation of Plato's *Epistles* (1962).

See also the two studies by HAROLD F. CHERNISS on the influence of Plato upon the Academy: *The Riddle of the Early Academy* (1945, reprinted 1962); and *Aristotle's Criticism of Plato and the Academy* (1944, reprinted 1962), on Plato as critically interpreted by Aristotle. Plato's impact upon subsequent Greek philosophy is felt throughout *The Cambridge History of Later Greek and Early Medieval Philosophy*, ed. by A.H. ARMSTRONG (1967), especially in PHILIP MERLAN'S synthesizing view of "Greek Philosophy from Plato to Plotinus," pp. 14–132.

(Ed.)

Platonism and Neoplatonism

The term Platonism can be applied to any philosophy that derives its ultimate inspiration from the *Dialogues* of Plato. But these can be read in many different ways, often very selectively, and it may be that all that the various kinds of Platonism can be said to have in common is an intense concern for the quality of human life—always ethical, often religious, and sometimes political, based on a belief in unchanging and eternal realities, independent of the changing things of the world perceived by the senses. Platonism sees these realities as giving life value and meaning. It is this belief in absolute values rooted in an eternal world that distinguishes Platonism from the philosophies of Plato's immediate predecessors and successors and from later philosophies inspired by them—from the immanentist naturalism of most of the pre-Socratics (who interpreted the world monistically in terms of nature as such), from the relativism of the Sophists, and from the correction of Platonism in a this-worldly direction carried out by Plato's greatest pupil, Aristotle.

"Platonism" as construed in this article, however, does not include the work of Plato himself but covers, instead, that of his successors in the entire movement that issued from his work. As so defined, Platonism therefore begins with Aristotle.

GREEK PLATONISM FROM ARISTOTLE THROUGH MIDDLE PLATONISM: ITS NATURE AND HISTORY

Plato as construed by Aristotle

Aristotle's statements about the philosophy of his master seem at first sight to be very different from anything to be found in the *Dialogues*, though there are a few points of contact in some of the later ones. It was generally assumed in antiquity that Aristotle's main source was an unpublished lecture or course of lectures on "The Good" (or "On Goodness") delivered by Plato in his later years; and it seems likely that the doctrines criticized by Aristotle and developed by Plato's immediate successors in the leadership of his Academy, Speusippus (head from 347–339 BC) and Xenocrates (head from 339–314 BC), do in fact derive from the later oral teaching of their master.

The main feature of Plato's later thought seems to have been a certain mathematicization. The Ideas or Forms—the eternal realities that give existence, meaning, and value to everything else—were thought of somehow in numerical terms; between them and physical things there was an intermediate realm of mathematical objects, which are changeless like the Forms, but—unlike them—not unique; there can be many of each kind (*e.g.*, there is only one Idea of Circle, but there are many geometrical circles, which can intersect, etc.). This intermediate realm of "mathematicals" was connected in some way with Soul—perhaps by Plato, certainly by his immediate successors. The ultimate principles of the Forms, and so of all reality, were also described in mathematical terms as the One and the Indefinite Dyad. The One seems to have been thought of as a principle of limit, form, and order, and so of goodness; the Indefinite Dyad, though necessary if any plurality of realities were to exist, was the principle of vague, unbounded multiplicity, and so of evil. These doctrines, developed with individual variations by Plato's two successors, played an important part in later Platonism. Aristotle himself had little use for this sort of mathematical metaphysics and rejected Plato's doctrine of transcendent eternal Forms altogether. Nonetheless, something of Platonism survived in Aristotle's system in his belief (1) that the reality of anything lay in a changeless (though wholly immanent) form or essence comprehensible and definable by reason and (2) that the highest realities were eternal, immaterial, changeless self-thinking intellects that caused the ordered movement of the universe by a desire for their transcendent perfection that kept the heavenly spheres rotating and to which man's intellect at its highest was akin. This Aristotelian doctrine of Intellect (nous) was easily recombined with Platonism in later antiquity.

The Academy through the Hellenistic Age

Little is known about the contributions, if any, to the development of Platonism that were made by the three successors of Xenocrates in the headship of the Academy: Polemon, Crates, and Crantor. It is possible that the influence of Crantor (*c.* 330–270 BC) on later Platonism—at least as a transmitter of the ideas of the Old Academy—may have been considerable, as he is said to have been the first to write commentaries on Plato; and in particular he is said to have written one on the *Timaeus*, in which he maintained that its description of the creation of the universe in time was not to be taken literally and that Plato really believed that the physical world had always existed.

After Crantor the Academy passed through a Skeptical phase and was preoccupied for about two centuries with the serious questioning of man's knowledge claims. This began with Arcesilaus (*c.* 315–*c.* 240 BC), who is described as the founder of the Middle Academy. (The distinction between the Middle and the New Academies, though ancient, is artificial, since there was a real continuity of thought throughout the Skeptical period in the school's history.) The Skeptical Academics were probably inspired

both by a genuine desire to recover the critical, questioning, and Agnostic attitude of the Socrates of Plato's early dialogues—the Socrates who knew only that he did not know—and by philosophical exasperation with the crude dogmatism of contemporary Hellenistic philosophers, especially the Stoics. It is likely that Arcesilaus was influenced to some extent by Pyrrhon (*c.* 360–270 BC), founder of the tradition to which the name Skeptic was applied in antiquity. The Skeptical Academics developed a formidable critique of Stoic dogmatism. They denied that certainty on any subject was possible and worked out a sophisticated theory of probability as a guide to practical decision making. Their critical dialectic and probability theory were brought to their highest point of refinement by Carneades (214/213–129/128 BC), called the founder of the New Academy, who made a considerable impression at Rome when he went there on a diplomatic mission from Athens in 155 BC. The definitive and final return of the Academy from Skepticism to dogmatic and positive philosophical teaching was brought about by Antiochus of Ascalon, who was head of the school in 79–78 BC, when Cicero heard some lectures delivered by him.

Transition to Neoplatonism

The next important phase of Platonism, Middle Platonism or Pre-Neoplatonism, was significant through the influence that it exerted in more than one direction. In the direction of Semitic culture (further described in a later section), it formed the Greek philosophical background of the efforts of the Jewish philosopher Philo of Alexandria to create a philosophical system on the basis of the Old Testament heritage. Though the origins of Middle Platonism are obscure, the outlines of its main direction of development became clear in the 1st century AD. It seems to have been linked from the beginning with the closely related revival of Pythagoreanism (a philosophy holding that reality is number, and sometimes showing, after the revival, a tendency to superstitious occultism). The somewhat Platonized Stoicism of Poseidonius (*c.* 135–50 BC), whose dualism of matter and reason enhanced the roles of emotion and will, may have influenced its beginnings; and Stoic influence, especially in the ethical field, remained important in its later developments. There was sometimes a considerable influence of Aristotle, though there was also a strong anti-Aristotelian tradition. The general characteristics of this revised Platonic philosophy (and the closely related Neo-Pythagoreanism) were the recognition of a hierarchy of divine principles with stress on the transcendence of the supreme principle, which already was called occasionally "the One"; the placing of the Platonic Forms in the divine mind; a strongly otherworldly attitude demanding a "flight from the body," an ascent of the mind to the divine and eternal; and a preoccupation with the problem of evil, attributed either to an evil world soul or to matter. The best known of the Middle Platonists is the biographer and essayist Plutarch of Chaeronea (*c.* AD 46–120). Other important figures in the 2nd century AD were Albinus, Atticus, and Numenius of Apamea. It was from the thought of these and other Middle Platonists that the foremost Neoplatonist, Plotinus, started in developing his own interpretation of Platonism, which was both profoundly original and firmly rooted in an established school tradition.

NEOPLATONISM: ITS NATURE AND HISTORY

Neoplatonism is the modern name given to the form of Platonism developed by Plotinus in the 3rd century AD, which came to dominate the Greek philosophical schools in the 4th century and remained predominant till the teaching of philosophy by pagans ended in the second half of the 6th century AD. It represents the final form of pagan Greek philosophy. It was not a mere syncretism (or combination of diverse beliefs) but a genuine, if one-sided, development of ideas to be found in Plato and earlier Platonism—though it incorporated important Aristotelian and Stoic elements as well. There is little real evidence for Oriental influence. A certain Gnostic (ancient theosophical) tone or colouring sometimes may be discerned in the thought of Plotinus. But he was consciously a passionate opponent of Gnosticism; and in any case there was often a large element of popular Platonism in Gnosticism itself—*i.e.*, in the Gnostic systems then current. Moreover, the theosophical works of the Chaldean theologians (late 2nd century AD)—the *Chaldean Oracles*, which were taken as inspired authorities by the later Neoplatonists—seem to have been a hotchpotch of popular Greek religious philosophy.

Core concepts of Neoplatonism

Neoplatonism began as a complex (and in some ways ambiguous) philosophy and developed in a long, vigorous and varied growth; it is therefore not easy to generalize about it. But the leading ideas in the thought of philosophers who can properly be described as Neoplatonists seem always to have included the following:

1. There is plurality of spheres of being, arranged in hierarchical descending order, the last and lowest comprising the universe, which exists in time and space and is perceptible to the senses.
2. Each sphere of being is derived from its superior, a derivation that is not a process in time or space.
3. Each derived being is established in its own reality by turning back toward its superior in a movement of contemplative desire, which is implicit in the original creative impulse of outgoing that it receives from its superior; thus the Neoplatonic universe is characterized by a double movement of outgoing and return.
4. Each sphere of being is an image or expression on a lower level of the sphere above it. The relation of archetype and image runs through the whole of the Neoplatonic scheme.
5. Degrees of being are also degrees of unity; in each subsequent sphere of being there is greater multiplicity, more separateness, and increasing limitation—till the atomic individualization of the spatiotemporal world is reached.
6. The supreme sphere of being, and through it all of what in any sense exists, derives from the ultimate principle, which is absolutely free from determinations and limitations and utterly transcends any conceivable reality, so that it may be said to be "beyond being." As it has no limitations, so it has no division, attributes, or qualifications; it cannot really be named but may be called "the One" to designate its complete simplicity. It may also be called "the Good" as the source of all perfections and the ultimate goal of return; for the impulse of outgoing and return that constitutes the hierarchy of derived reality comes from and leads back to the Good.
7. Since this supreme principle is absolutely simple and undetermined (or devoid of specific traits), man's knowledge of it must be radically different from any other kind of knowledge: it is not an object (a separate, determined, limited thing) and no predicates can be applied to it; hence it can be known only if it raises the mind to an immediate union with itself, which cannot be imagined or described.

Plotinus and his philosophy. As far as is known, the originator of this distinctive kind of Platonism was Plotinus (AD 205–270). He had been the pupil at Alexandria of a self-taught philosopher called Ammonius, who also taught the Christian, Origen, and the latter's pagan namesake, and whose influence on his pupils seems to have been deep and lasting. But Ammonius wrote nothing, and so nothing is really known about his thought. A number of distinguished scholars have recently made attempts to reconstruct it, but their speculations go far beyond the evidence. Plotinus must thus be regarded as the first great Neoplatonist, and his collected works, the *Enneads* (Greek *ennea*, "nine"—six sets of nine treatises each, arranged by his disciple Porphyry), are then the first and greatest collection of Neoplatonic writings. (For further details about the life and works of Plotinus, see PLOTINUS.)

Plotinus, like other philosophers of late antiquity, was a religious and moral teacher as well as a professional philosopher engaged in the critical interpretation of a long and complicated school tradition. He was an acute critic and arguer, with an exceptional degree of intellectual honesty for his, or any, period; but philosophy for him was not so much a matter of abstract speculation as a way of life in which, through an exacting intellectual and moral self-discipline and purification, those who are capable of the ascent can return to the source from which they came. His written works explained how, from the eternal

Metaphysics: system of emanation

creative act—at once spontaneous and necessary—of that transcendent source, the One or Good, proceeds the world of living reality constituted by repeated double movements of outgoing and return in contemplation; and this account, showing the way for the human self—which can experience and be active on every level of being—to return to the One, was itself an exhortation to follow that way.

Plotinus always insisted that the One or Good is beyond the reach of thought or language; what he said about this supreme principle was intended only to point the mind along the way to it, not to describe or define it. But though no adequate concept or definition of the Good is possible, it was nonetheless for Plotinus a positive reality of superabundant excellence. Plotinus often spoke of it in extremely negative language, but his object in doing so was to stress the inadequacy of all of man's ways of thinking and speaking to express this supreme reality or to clarify the implications of the claim that the Good is absolutely one and undetermined, the source of all defined and limited realities.

The original creative or expressive act of the One is the first great derived reality, Nous (which can be only rather inadequately translated as "Intellect" or "Spirit"); from this again comes Soul, which forms, orders, and maintains in being the material universe. It must be remembered that, to Plotinus, the whole process of generation is timeless; Nous is eternal, and time is the life of Soul as active in the physical world, and there never was a time when the material universe did not exist. The "levels of being," then, though distinct, are not separate but are all intimately present everywhere and in everyone. To ascend from Soul through Intellect to the One is not to travel in space but to awake to a new kind of awareness.

Intellect, Soul, world

Intellect for Plotinus is both thought and object of thought; it is a mind that is perfectly one with its object. As object, it is the world of forms or ideas, the totality of real being in the Platonic sense. These forms, being one with Intellect, are not merely objects but are living, thinking subjects, each not only itself but, in its contemplation, the whole. They are the archetypes and causes of the necessarily imperfect realities on lower levels, souls and the patterns or structures that make bodies what they are. Men at their highest are intellects, or souls perfectly conformed to Intellect; they become aware of their intellectual nature when, passing not only beyond sense perception but beyond the discursive reasoning characteristic of the life of Soul, they immediately grasp eternal realities.

Soul for Plotinus is very much what it was for Plato, the intermediary between the worlds of Intellect and Sense and the representative of the former in the latter. It is produced by Intellect, as Intellect is by the One, by a double movement of outgoing and return in contemplation, but the relationship between the two is more intimate and the frontier less clearly defined. For Plotinus, as for Plato, the characteristic of the life of the Soul is movement, which is the cause of all other movements. The life of the Soul in this movement is time, and on it all physical movement depends. Soul both forms and rules the material universe from above; and in its lower, immanent phase, which Plotinus often calls nature, it acts as an indwelling principle of life and growth and produces the lowest forms, those of bodies. Below these lies the darkness of matter, the final absence of being, the absolute limit at which the expansion of the universe—from the One through diminishing degrees of reality and increasing degrees of multiplicity—comes to an end. Because of its utter negativity, such matter is for Plotinus the principle of evil; and although he does not really believe it to be an independent principle forming, with the good, a dualism, his language about it often has a strongly dualistic flavour.

He was not, however, really dualistic in his attitude toward the material universe. He strongly maintained its goodness and beauty as the best possible work of Soul. It is a living organic whole, and its wholeness is the best possible (though very imperfect) reflection on the space-time level of the living unity in diversity of the world of forms in Intellect. It is held together in every part by a universal sympathy and harmony. In this harmony external evil and suffering take their place as necessary elements in the great pattern, the great dance of the universe. Evil and suffering can affect men's lower selves but cannot touch their true, higher selves and so cannot interfere with the real well-being of the philosopher.

Ethics and religion: purification

As souls within bodies, men can exist on any level of the soul's experience and activity. (The descent of souls into bodies is for Plotinus—who had some difficulty in reconciling Plato's various statements on this point—both a fall and a necessary compliance with universal law.) Man can ascend in spirit to the level of universal Soul, become that whole that he already is potentially, and, in Soul, attain to Intellect; or he can isolate himself on the lower level, shutting himself up in the experiences, desires, and concerns of his lower nature. Philosophical conversion—the beginning of the ascent to the One—consists precisely in turning away, by a tremendous intellectual and moral effort, from the life of the body, dominating and rising above its desires, and "waking to another way of seeing, which everyone has but few use." This, Plotinus insisted, is possible while one is still in an earthly body and without neglecting the duties of his embodied state. But the body and bodily life weigh a man down and hamper him in his ascent. Plotinus' language when speaking of the body and the senses in this context is strongly dualistic and otherworldly. Platonists in general think much more dualistically about their own bodies than about the material universe as a whole. The physical world is seen positively as a noble image of the intelligible; the individual, earthly, animal body, on the contrary, tends to be regarded negatively as a hindrance to the intellectual and spiritual life.

When a man's philosophical conversion is complete and he has become Intellect, he can rise to that mystical union in which the One manifests his continual presence, carried on the surging current of the impulse of return to the source (in its strongest and final flow), the pure love of Intellect for the Good from which it immediately springs. There is no consciousness of duality in that union; one is not aware of himself; but neither is he destroyed or dissolved into the One—because even in the union he is still Intellect, though Intellect "out of itself," transcending its normal nature and activity. This mystical union for Plotinus was the goal of all his effort and the source of the continuing power of his teaching. Philosophy for him was religion, the effort to actualize in oneself the great impulse of return to the Good, which constitutes reality on all its levels; and religion for him was philosophy. There was no room in his thought and practice for special revelation, grace, and repentance in the Christian sense, and little for external rites or sacraments. For him the combination of moral purification and intellectual enlightenment, which only Platonic philosophy as he understood it could give, was the only way to union with the Good.

Porphyry and Iamblichus

The later Neoplatonists. Porphyry (*c.* AD 234–*c.* 305), a devout disciple of Plotinus and a careful editor of his works, occupied a special position in the development of later Neoplatonism. In some ways his thought paralleled that of the later pagan Neoplatonists, but in others it quite opposed them. The most distinctive features of his thought seem to have been an extreme spiritualism, an insistence, even sharper than that of Plotinus, on the "flight from the body" and—more philosophically important—a monistic interpretation of the Plotinian system; he tended, that is, to think that the One, Intellect, and Soul are ultimately only one reality. Soul in its true nature is for him practically identical with Intellect (which was already true in some of the writings of Plotinus), and Intellect is the self-determination of the life springing from the One, which is unknowable, infinite, pure being; the three—Being, Life, and Intellect—known as the "horizontal" triad, are phases in the eternal self-determination of the ultimate (and only) reality. This triad, which had its origins in the thought of Plotinus, became one of the most important elements in the complex metaphysical structures of the later Neoplatonists. But Iamblichus and his successors reacted strongly against Porphyry's monistic tendencies.

Iamblichus (*c.* AD 250–*c.* 325) was the real originator of the type of Neoplatonism that came to dominate the Pla-

tonic schools in the 5th and 6th centuries AD. This kind of Neoplatonism sharpened and multiplied the distinctions between the levels of being. The basic position underlying its elaborations is one of extreme philosophical Realism: it is assumed that the structure of reality corresponds so exactly to the way in which the mind works that there is a separate real entity corresponding to every distinction that it can make. In the fully developed late Neoplatonic system, the first principle of reality, the ultimate One, was removed to an altogether ineffable transcendence, mitigated by two factors: (1) the presence of the expressions or manifestations of its unifying power, the "henads"—identified with the gods of paganism—at every level of reality, and (2) the possibility of return to absolute unification through the henad with which one is linked. Below the One a vast structure of triads or trinities reached down to the physical world: this was constructed by combining Plotinus' "vertical" succession of the levels of Being, Intellect, and Soul (much complicated by internal subdivision and the interposition of mediating hypostases at every stage) with the horizontal triadic structure, giving a timeless dynamic rhythm of outgoing and return, such as that already encountered in Porphyry.

Neoplatonism as a pagan theology

This later Neoplatonism aspired to be not only a complete and coherent metaphysical system but also a complete pagan theology. The maintenance and defense of the old religion in a world more and more intolerantly dominated by its triumphant rival, Christianity, was one of the main concerns of the Platonists after Plotinus. By the study of Aristotle and of the religious metaphysics of Plato (above all in the *Parmenides* and the *Timaeus*), in which they employed remarkable processes of forced exegesis, they believed it possible to arrive at a complete understanding of divine truth cryptically revealed by the gods themselves through the so-called theologians—the inspired authors of the Orphic poems and of a strange theosophical farrago called the *Chaldean Oracles*, published in the second half of the 2nd century AD, to which Porphyry first gave some guarded and qualified recognition but which was inspired scripture to Iamblichus and his successors. Their view of the human soul was a humbler one than that of Plotinus: it was for them a spiritual being of low rank, which had descended altogether into the material world; they could not therefore aspire, like Plotinus, to that return to and unification with the divine that remained for them the goal of human life through philosophy alone. Help from the gods was needed, and they believed that the gods in their love for men had provided it, giving to all things the power of return in prayer and implanting even in inanimate material things —herbs and stones and the like—sympathies and communications with the divine, which made possible the secret rites of theurgy, through which the divine gave the needed spiritual help by material means. Theurgy, though its procedures were generally those of late Greek magic, was thus not thought of merely as magic; and the degree of attention paid to external rites varied considerably from philosopher to philosopher: there seem to have been men even in the last generation of pagan Neoplatonists who had little use for or interest in such things and followed a mystical way much like that of Plotinus.

Neoplatonic schools

The different schools of late Neoplatonism seem to have differed less from each other than has sometimes been supposed. The school of Pergamum, founded by Aedesius, a pupil of Iamblichus, made perhaps the least contribution to the philosophical development of Neoplatonism, but it was not entirely given over to theurgy. Its greatest convert was the emperor Julian, called by Christians the "Apostate." By the end of the 4th century AD, the Platonic Academy at Athens had become Neoplatonist in the manner of Iamblichus. It was particularly fervent and open in its paganism and attracted Christian hostility. Though maintaining itself for a surprisingly long time against this hostility, it eventually yielded to it and was closed by Justinian in AD 529. In the interim, however, it had produced the greatest and most influential systematic expositor of later Neoplatonism, Proclus (AD 410–485). The head of the school at the time of its closing, Damascius, was also a notable philosopher. The school of Alexandria in the 5th and 6th centuries does not seem to have differed very much in its fundamental philosophical outlook from that of Athens. Philosophers like Simplicius, the greatest of the Neoplatonic commentators on Aristotle, passed easily from one school to the other. But the Alexandrians seem to have been more discreet in their attitude toward Christianity and consequently devoted much time to the comparatively safe and neutral study of Aristotle. An openly pagan (and very inferior) philosopher, Olympiodorus, was still teaching at Alexandria well into the second half of the 6th century. Finally, in the 7th century, after philosophical teaching had passed peacefully into Christian hands, the last known Alexandrian philosopher, the Christian Stephanus, was called to teach in the University of Constantinople under Heraclius. Stephanus formed one of the most important links in the chain of transmission of Platonism to the Byzantine Middle Ages.

PLATONISM IN THE WORLD OF THE REVEALED RELIGIONS

Early Jewish Platonism. Well before the beginning of the Christian Era, Jews with some Greek education had begun to make casual use of popular Greek philosophy in expounding their revealed religion: there are traces of this in the Wisdom literature of the Old Testament. In Paul's speech to the Areopagus in Acts 17, commonplaces of Stoic philosophy were employed for apologetic purposes. But, as far as is known, the first Jew who was really well read in Greek philosophy and used it extensively in the exposition and defence of his traditional religion was Philo of Alexandria (*c.* 15 BC–after AD 40), an older contemporary of St. Paul. Philo expressed his philosophical religion in the form of lengthy allegorical commentaries on the Jewish Scriptures, especially on Genesis, in which he showed to his own satisfaction that the ancient revelation given to Moses accorded with the teaching of the best Greek philosophers, which, in his view, was later and derivative. The Greek philosophy that he preferred and found to be most in accordance with revelation was an early form of Middle Platonism (see above *Transition to Neoplatonism*). Philo was neither approved of nor read by later orthodox Jews, but his influence on Greek-speaking and Greek-educated Christians from the 2nd century AD was great; and in important ways he determined the tone of their religious speculation.

Philo of Alexandria

Ancient and medieval Christian Platonism. Like Philo, the Christian Platonists gave primacy to revelation and regarded Platonic philosophy as the best available instrument for understanding and defending the teachings of Scripture and church tradition. But, also like Philo, they did not believe that truth could conflict with truth and were confident that all that was rationally certain in Platonic speculation would prove to be in perfect accordance with the Christian revelation. Their unhistorical approach and unscholarly methods of exegesis of texts, both pagan and Christian, facilitated this confidence. The general attitude of Christian Platonists was one of relatively moderate and humane otherworldliness (the cruder sorts of Christian otherworldliness and hatred of the body seem to derive from non-Platonic and non-Greek sources). They stressed the transcendence of God though, by insisting that it is a transcendence that is also the deepest immanence, they acknowledged his intimate presence within the world as well. They took a dualistic view of soul and body (though accepting bodily resurrection) and emphasized the primacy of the spiritual, while insisting on the goodness of God's material creation.

Patristic Platonism. From the middle of the 2nd century AD, Christians who had some training in Greek philosophy began to feel the need to express their faith in its terms, both for their own intellectual satisfaction and in order to convert educated pagans. The philosophy that suited them best was Platonism. Though Stoicism had exerted a considerable influence on Christian ethical thinking (which has persisted till our own times), Stoic corporealism—the belief that God and the soul are bodies, of a subtle and peculiar kind—repelled most Christians, and Stoic pantheism was incompatible with Christianity. The Platonism that the first Christian thinkers knew was of

course Middle Platonism, not yet Neoplatonism. Its relatively straightforward theism and high moral tone suited their purposes excellently; and the influence of this older form of Platonism persisted through the 4th century, and beyond, even after the works of Plotinus and Porphyry began to be read by Christians.

Justin, Origen, the Cappadocians

The first Christian to use Greek philosophy in the service of the Christian faith was Justin Martyr (martyred *c.* 162–168), whose passionate rejection of Greek polytheism, combined with an open and positive acceptance of the essentials of Platonic religious philosophy and an unshakable confidence in its harmony with Christian teaching, was to remain characteristic of the Christian Platonist tradition. This was carried on in the Greek-speaking world by Clement of Alexandria (*c.* 150–*c.* 215), a persuasive Christian humanist, and by the greatest of the Alexandrian Christian teachers, Origen (*c.* 184–254), who was consciously more hostile to and critical of Platonic philosophy than either Justin or Clement but was nonetheless more deeply affected by it and produced a synthesis of Christianity and late Middle Platonism of remarkable originality and power, which is the first great Christian philosophical theology. In spite of subsequent condemnations of some of his alleged views, his influence on Christian thought was strong and lasting. The Greek philosophical theology that developed during the Trinitarian controversies over the relationships among the persons of the Godhead, which were settled at the ecumenical councils of Nicaea (325) and Constantinople (381), owed a great deal to Origen, on both sides, orthodox and heretical. Its greatest representatives on the orthodox side were the three Christian Platonist theologians of Cappadocia, Basil of Caesarea (*c.* 330–379), Gregory Nazianzus (*c.* 330–390), and Basil's brother Gregory of Nyssa (died 394). Of these, Gregory of Nyssa was the most powerful and original thinker (and the closest to Origen). He was the first great theologian of mystical experience, at once very Platonic and profoundly Christian, and he exerted a very strong influence on later Greek Christian thought.

Pseudo-Dionysius and others

At some time between the period of the Cappadocian Fathers and the early years of the 6th century, a new turn was given to Christian Platonism by the remarkable writer who chose to publish his works under the name of St. Paul's convert at Athens, Dionysius the Areopagite. The kind of Platonism that the Pseudo-Dionysius employed for his theological purposes was the 5th-century Neoplatonism that is best represented by Proclus (see above). Almost everything about this mysterious author is vigorously disputed by scholars. But there can be no doubt about the influence that his system of the hierarchic universe exerted upon later Christian thought; his vision of man's ascent through it—carried up by divine love, to pass beyond all hierarchy and all knowledge into the darkness of the mystical union with God—had its impact both in the East, where one of the greatest of Greek Christian Platonist thinkers, Maximus the Confessor (*c.* 580–662), was deeply influenced by the Dionysian writings and commented extensively upon them, and in the West, where they became known and were translated into Latin in the 9th century. In the Latin West there was more than one kind of Christian Platonism. An impressive and extremely difficult philosophical theology, employing ideas approximating Porphyry's version of Neoplatonism to explain and defend the orthodox doctrine of the Trinity, was produced in the second half of the 4th century by the rhetorician and grammarian Marius Victorinus. A strong and simple Platonic theism and morality, which had a great influence in the Middle Ages, was nobly expressed in the final work of the last great philosopher-statesman of the ancient world, Boethius (*c.* 480–524). This was the *Consolation of Philosophy*, written in prison while its author was under sentence of death. But the Christian Platonism that had the widest, deepest, and most lasting influence in the West was that of St. Augustine.

Augustinian Platonism. Each of the great Christian Platonists understood Platonism and applied it to the understanding of his faith in his own individual way; and of no one of them was this more true than of Augustine, with his extremely strong personality and distinctive religious history. Augustine's thought was not merely a subspecies of Christian Platonism but something unique—Augustinianism. Nonetheless, the reading of Plotinus and Porphyry (in Latin translations) had a decisive influence on his religious and intellectual development, and he was more deeply and directly affected by Neoplatonism than any of his contemporaries and successors.

Teachings of Augustine

In his anthropology, Augustine was firmly Platonist, insisting on the soul's superiority to and independence of the body. For him, as for Plotinus and Porphyry, it was axiomatic that body could not act on soul, for soul was superior in the hierarchy of reality, and the inferior cannot act on the superior. This affected both his ethical doctrine and his epistemology. On the other hand, he differed from the philosophers who influenced him in his insistence that not only man but higher spiritual beings as well are mutable and peccable, liable to sin and fall, and in his consequent stress on the necessity of divine grace. His crucial doctrine that man's destiny is determined by the right direction of love, though profoundly original, was a development rather than a contradiction of Platonism. His very original theology of history and his view of human society, however, owed little to Plotinus and Porphyry, whose interests lay elsewhere.

In his epistemology, Augustine was Neoplatonic, especially in the subjectivity of his doctrine of illumination —in its insistence that in spite of the fact that God is exterior to man, men's minds are aware of him because of his direct action on them (expressed in terms of the shining of his light on the mind, or sometimes of teaching) and not as the result of reasoning from sense experience. For a Platonist, as has been said, body cannot act upon soul. Sense experience, therefore, though genuinely informative on its own level, cannot be a basis for metaphysical or religious thinking. This must be the result of the presence to the soul of higher realities and their action upon it. In Plotinus the illumination of the soul by Intellect and the One was the permanent cause of man's ability to know eternal reality; and Augustine was at this point very close to Plotinus, though for him there was a much sharper distinction between Creator and creature, and the personal relationship between God and the soul was much more strongly stressed.

In his theology, insofar as Augustine's thought about God was Platonic, he conformed fairly closely to the general pattern of Christian Platonism; it was Middle Platonic rather than Neoplatonic in that God was not the One beyond Intellect and Being but the supreme reality in whose creative mind are the Platonic Forms, the eternal patterns or regulative principles of all creation. Perhaps the most distinctive influence of Plotinian Neoplatonism on his thinking about God was in his Trinitarian theology. He started with the unity of God and continually insisted upon it, unlike Greek Christian thinkers, who started with the Three Persons perfectly united, and because he thought that something like the Christian doctrine of the Trinity was to be found in Plotinus and Porphyry, he tended to regard it as a philosophical doctrine and tried to make philosophical sense of it to a greater extent than the Greek Fathers did. His last and most important and influential attempt to do so was in his treatise *On the Trinity*, with its discovery of analogies to the divine mystery in the self-directed, internal activities of the soul.

Medieval Platonism. With the gradual revival of philosophical thinking in the West that began in the Carolingian period (late 8th–9th centuries), the history of Platonism becomes extremely complex. Only a kind of sketch map that distinguishes the main streams of more or less Platonic tradition can be given here.

Influence of the *Timaeus*

In the 4th century the Christian exegete Calcidius prepared a commentary on Plato's *Timaeus*, which exerted an important influence on its medieval interpretation. A Christian Platonic theism of the type of which Boethius is the finest example thus arose, based on a reading of the *Timaeus* with Christian eyes, and continued to have a

strong influence in the Middle Ages, especially in the earlier period. This kind of theism, issuing in a strongly positive view of God's creation and a nobly austere but humane view of man's duty and destiny, was particularly apparent in the Christian humanism of the School of Chartres (12th century).

The widest, deepest, and most persistent Christian Platonist influence in the Latin West was that of Augustine (see above *Augustinian Platonism*). Augustinianism in a variety of forms—often stiffened, exaggerated, or distorted—persisted throughout the Middle Ages and survived the "recovery of Aristotle" (see below). In the later Middle Ages Augustine's influence was particularly strong in the Franciscan school, though not confined to it. But the greatest and most influential of medieval thinkers deeply influenced by Augustine was Anselm of Canterbury (1033–1109), the originator (probably on the basis of suggestions in Augustine) of the still much discussed "ontological argument" for the existence of God (see RELIGION, PHILOSOPHY OF) and a philosopher whose humility, openness, and readiness to consider objections had a genuinely Socratic quality.

John Scotus Erigena

One of the boldest and most original thinkers of medieval Europe was John Scotus Erigena (2nd half of the 9th century AD), who introduced to the West the Greek Christian Platonist tradition (see above *Patristic Platonism*), as it had been developed by the Pseudo-Dionysius and Maximus the Confessor. His views were much disapproved of by the Western Church; and his great philosophical work, the *Periphyseon* (usually known as *De divisione naturae*), was not much read and ceased to be copied after his condemnation in 1210. But a considerable part of the text circulated in the form of anonymous glosses to the Latin translations of the Pseudo-Dionysius (of which the first adequate translation was by Erigena himself); and in this way his thought influenced both the tradition of Western mysticism, which derived from the Pseudo-Dionysius, and 13th-century Scholasticism, for which St. Paul's supposed disciple was still a major authority.

The Platonism in "Aristotelian" Scholasticism

There is no more superficial and misleading generalization in the history of philosophy than that which sharply opposes "Christian Platonism" and "Christian Aristotelianism." To be sure, the recovery of the authentic thought of Aristotle through Latin translations of his works in the 12th and 13th centuries was indeed a major event in the history of philosophy. But Platonism and Aristotelianism have never been tidily separated in the history of European thought. There was already a strong Aristotelian element in Middle and Neoplatonism. Byzantine theologians (in the East) from the 6th century AD onward were as Aristotelian as anybody in western Europe in the 13th century. Thirteenth-century "Aristotelian" Scholastics, though much preoccupied with the new translations of Aristotle and their philosophical and theological implications, were still deeply influenced by Augustine, Boethius, and the Pseudo-Dionysius (with glosses derived from Erigena). And the Islāmic philosophy, to be mentioned below, with which they had to grapple, was as much Neoplatonist as it was Aristotelian. Further, they also were influenced by the *Liber de causis*, compiled by an Arab author from Proclus and attributed to Aristotle, as well as by Latin translations of some of the shorter works of Proclus (see above). It has been said that "Aquinas is closer to Plotinus than to the real Aristotle," and there is some truth in this judgment.

Islāmic and medieval Jewish philosophy. After the Muslim conquest of Syria and Egypt, there began a great work of translation of the texts that had been studied in the late Greek philosophical schools—including a number of dialogues of Plato and Neoplatonic treatises, as well as the works of Aristotle and a number of the Alexandrian Neoplatonist commentaries on them. The translations—partly from Greek, partly from Syriac versions of the Greek texts—were made between about 800 and 1000. On the basis of these translated texts an impressive development of Islāmic theology and philosophy (*q.v.*) took place, strongly influenced by Neoplatonism, though Aristotelian influence also became increasingly important. An interesting feature of this Islāmic philosophy, which distinguished it sharply from the familiar Neoplatonism, was its much greater interest in the political and social side of Plato's thought and its attempts to adapt Platonic political ideas to the conditions of the contemporary Muslim world. The tradition culminated in two great Muslim philosophers who deeply influenced the medieval West, Avicenna (Ibn Sīnā, 980–1037) from Persia and Averroës (Ibn Rushd, 1126–98) from Muslim Spain. Of these, Avicenna was perhaps the more Platonist, and Averroës, whose fame and influence rested primarily on his commentaries on Aristotle, was the more Aristotelian. But both knew, were influenced by, and developed in their own way both Platonic and Aristotelian ideas. Medieval Jewish philosophy (*q.v.*), which also developed within this Muslim intellectual tradition, especially in Spain, reflected—at least in its earlier phases—strong Neoplatonic influence. This is especially true of the thought of the first great Jewish philosopher of Muslim Spain, Ibn Gabirol (born *c.* 1021).

PLATONISM FROM THE RENAISSANCE TO MODERN TIMES

From the 15th century onward the *Dialogues* of Plato and a large number of Middle Platonist and Neoplatonist works, above all the *Enneads* of Plotinus, became available in the original Greek in western Europe and were studied by scholars, philosophers, and theologians. As a result of this new acquaintance with the original texts, Platonic influences on Renaissance and post-Renaissance thought became even more complex and difficult to recognize than those on medieval thought. Older Neoplatonically influenced traditions (notably Augustinianism) persisted, and new ones developed from the direct reading of the Neoplatonic texts. And, at least from the time of Leibniz (1646–1716), European thinkers realized that the Neoplatonic interpretation of Plato was in some ways a distorted and one-sided one; hence they sometimes developed their own allegedly more authentic understandings of Plato on the basis of direct readings of such of his varied works as they found to be philosophically congenial. Only a few of the more interesting Platonic influences can be indicated here.

The persistence of Christian Platonism

In spite of its deep influence on Greek Christian thinkers, Platonism was regarded with profound suspicion by the Byzantine Orthodox Church. The suspicion reflected its association in the Byzantine ecclesiastical mind with the militant paganism of the Athenian Neoplatonists (see above). Nonetheless, it survived in the Byzantine world—generally underground but with a notable overt revival in the 11th century, in which the most notable figures were the broadly erudite Michael Psellus, who did much to enhance the prestige of philosophy, and his rival, the syncretistic Aristotelian commentator John Italus. In the 15th century the last known Byzantine philosopher, George Gemistus Plethon, a passionate pagan Platonist in the manner of Proclus, travelled to Italy (1438–39) and persuaded Cosimo de' Medici to sponsor a Platonic Academy at Florence, of which the greatest figures to emerge were its founder, the Platonic commentator Marsilio Ficino, and the humanist Pico della Mirandola, author of the influential *Oration on the Dignity of Man*. In spite of the paganism of Plethon, the Platonism of the Florentine Academy was a Christian one of humane and liberal kind. This was probably at least partly due to the influence in Italy of the thought of Cardinal Nicholas of Cusa (1401–64), who at about the same time (the period of the short-lived reunion between Eastern and Western churches at the Council of Florence), worked out his own very original version of Christian Platonism, influenced by the Pseudo-Dionysius, Erigena, and the German mystical tradition (as in Meister Eckehart).

The influence of the Christian Platonism of the Florentine Academy was quite extensive and was particularly marked in England. Perhaps the most impressive development of this post-Renaissance movement lay in the works of the Cambridge Platonists (late 17th century). Since their time a tradition of liberal Christian Platonism has persisted in England, predominantly, though not exclu-

sively, among Anglican thinkers; and though at present unfashionable, it is by no means dead. Moreover, there have been other notable traditions of Platonically influenced Christian thought in Europe. One that deserves to be better known is that of the outstanding French philosopher of "action" Maurice Blondel (1861–1949), who found a prominent place in his system for the formation of ideas—interpreted as an important species of action that faithfully reflects the eternal order of reality. Blondel's philosophy has had a widespread influence, mainly among Catholic philosophers dissatisfied with Neoscholasticism.

Pagan Neoplatonic influence on German Idealism and on English literature

The rediscovery of Proclus by the great German Idealist G.W.F. Hegel (1770–1831) had an important influence on his thought and so on the whole history of 19th-century Idealist philosophy. The German Idealists, of course, understood Neoplatonism much in their own way; and Idealism should not be interpreted as revived Neoplatonism, nor Neoplatonism as an anticipation of Idealism. But the historical influence of a direct reading of the Neoplatonists on Idealist thought is indisputable. There was a strong reaction against Hegel's influence in some quarters in 19th-century Germany, and this reaction led to a corresponding depreciation of Neoplatonism. But 20th-century German (and French) scholars have been readier than English scholars to take a serious interest in Neoplatonism. The latter, with some notable exceptions, have maintained a traditionally rationalist attitude toward that philosophy, which, in all its forms, they have regarded as "decadent" and "mystical."

The influence of the sort of Christian Platonism mentioned above on English literature, and especially on English poetry, has been wide and deep. But there has also been a strongly anti-Christian Neoplatonic influence, that of Thomas Taylor "the Platonist" (1758–1835), who published translations of Plato, Aristotle, and a large number of Neoplatonic works in the late 18th and early 19th centuries. Taylor was as militant in his pagan Platonism as Gemistus Plethon. His ideas had a strong influence on the poetry of William Blake, one of the fathers of "visionary" British Romanticism, who eventually succeeded in reconciling Taylor's paganism with his own very original version of Christianity: much of the symbolism in Blake's poetry is Neoplatonic. The Platonism of the English Romantic poet and political radical Shelley also derives from Taylor. Thomas Taylor also deeply influenced Emerson and his circle in America. Later, in the early 20th century, the influence of Taylor's writings was again apparent in the Irish poet and dramatist, Yeats, especially in his later poems.

The foremost process philosopher A.N. Whitehead (1861–1947), perhaps because of his original and abiding concern with mathematical philosophy, was interested in Plato (though not, apparently, in the Neoplatonists); and the reading of the *Timaeus* in particular contributed something to the metaphysical system of his last period and especially to his concept of a God who does not timelessly transcend process but is in some way involved in it. Whitehead is an excellent example of a Platonically influenced thinker whose development of Plato's own thought proceeds along lines completely opposed to Neoplatonism.

EVALUATION OF PLATONISM

The essential point at issue between Platonists and their opponents through the centuries has been the existence (in some sense) of spiritual or intelligible reality, which is independent of the sense world, is normative, and is the source of values. This is a very rough generalization, which does not apply to the skeptical Academy (see above *Greek Platonism from Aristotle through Middle Platonism*) nor do full justice to the thought of the modern skeptical Platonist George Santayana. Platonists have understood this central doctrine in a great variety of ways and defended it with a great variety of arguments. But whenever it has been strongly held, it seems to have been by a faith depending on some sort of experience rather than simply as the conclusion of an argument. Its opponents have generally followed the lines of attack laid down by Aristotle (and to some extent anticipated by Plato in the first part of his *Parmenides*): that the doctrine involves an illegitimate duplication of reality and the postulation of entities for the existence of which no sufficent evidence or arguments can be offered and the relationship of which to the world of sense experience cannot be intelligibly stated.

The argument still continues and will perhaps never finally be settled, but there can be no doubt about the central importance of Platonism in the history of European thought. It has been a continual stimulus to philosophical discussion, and it has provided the central tradition of Western religious philosophy and so has deeply influenced Christian theology. Finally, it has done more than any other philosophy of ancient Greek origins to stimulate new developments in physical science, literature, and art.

BIBLIOGRAPHY. A good source of information on the Platonic and Neoplatonic philosophers up to and including Anselm is the *Cambridge History of Later Greek and Early Medieval Philosophy*, ed. by A.H. ARMSTRONG (1967; reprinted with revised bibliographies, 1970). The bibliographies of this work include a list of the editions of ancient and medieval sources (complete and fragmentary), with the more important translations and modern works. For medieval Platonism, see B. GEYER, *Die patristische und scholastische Philosophie*, in F. UEBERWEG, *Grundriss der Geschichte der Philosophie*, pt. 2, 12th ed. (1951); ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955), with valuable bibliographical material; D. KNOWLES, *The Evolution of Mediaeval Thought* (1962); G. LEFF, *Mediaeval Thought: St. Augustine to Ockham* (1960); and W. BEIERWALTES (ed.), *Platonismus in der Philosophie des Mittelalters* (1969), a collection of important articles published in German. All these works to some extent cover the later medieval period not dealt with in the *Cambridge History*. For medieval Jewish philosophy, see I. HUSIK, *A History of Mediaeval Jewish Philosophy* (1946). For Neoplatonic movements in Jewish Hellenistic and medieval philosophy, see J. GUTTMANN, *Philosophies of Judaism: The History of Jewish Philosophy from Biblical Times to Franz Rosenzweig*, trans. from the Hebrew by D.W. SILVERMAN (1964). There is no adequate and up-to-date account of Renaissance and post-Renaissance Platonism; reference must be made to general histories of philosophy. P. SHOREY, *Platonism, Ancient and Modern* (1938), still remains the best general introduction. On Byzantine Platonism, see J.M. HUSSEY, *Church and Learning in the Byzantine Empire, 867–1185* (1937); and B.N. TATAKIS, *La Philosophie byzantine* (1949). The most satisfactory short general account of Renaissance Platonism in English is that by F.C. COPLESTON, *History of Philosophy*, vol. 3, ch. 12 and 15 (1953). A good short introduction to Renaissance Platonism in England is ERNST CASSIRER, *Die Platonische Renaissance in England und die Schule von Cambridge* (1932; Eng. trans., *The Platonic Renaissance in England*, 1953). G.R. CRAGG (ed.), *The Cambridge Platonists* (1968), is an excellent anthology, with good introductions and notes. On English Christian Platonism, see also W.R. INGE, *The Platonic Tradition in English Religious Thought* (1926). On the influence of Thomas Taylor, see *Thomas Taylor the Platonist*, selected writings ed. by K. RAINE and G.M. HARPER (1969); and F.A.C. WILSON, *W.B. Yeats and Tradition* (1958).

(A.H.A.)

Platyhelminthes

The flatworms constitute the phylum Platyhelminthes, a group of invertebrate animals that are free-living as well as parasitic—*i.e.*, living on or in another organism and securing nourishment from it. They are bilaterally symmetrical (*i.e.*, the right and left sides are similar), usually flattened, and lack respiratory, skeletal, and circulatory systems; no body cavity (coelom) is present. The body is not divided into true segments; spongy connective tissue (mesenchyme) constitutes the so-called parenchyma and fills the space between organs. Flatworms, generally hermaphroditic—functional reproductive organs of both sexes occurring in one individual—are the lowest invertebrates to possess three embryonic layers—endoderm, mesoderm, and ectoderm—and to have a head region containing concentrated sense organs and nervous tissue (brain).

The phylum consists of five classes: Trematoda (flukes), Cestoda (tapeworms), Turbellaria, Monogenea, and As-

pidocotylea (or Aspidobothria). Members of all classes except Turbellaria are parasitic during all or part of the life cycle. Most turbellarians are exclusively free-living forms. About 13,000 species of flatworms have been described.

By courtesy of (bottom left) D.P. Wilson; from (top) Faust, Russell and Jung, *Craig and Faust's Clinical Parasitology* (1970), reproduced by permission of Lea & Febiger; (centre right) *Invertebrate Zoology* by Paul A. Meglitsch, Copyright © 1967 by Oxford University Press, Inc., reprinted by permission; (bottom right) © 1965 Time-Life Books, Time Inc.

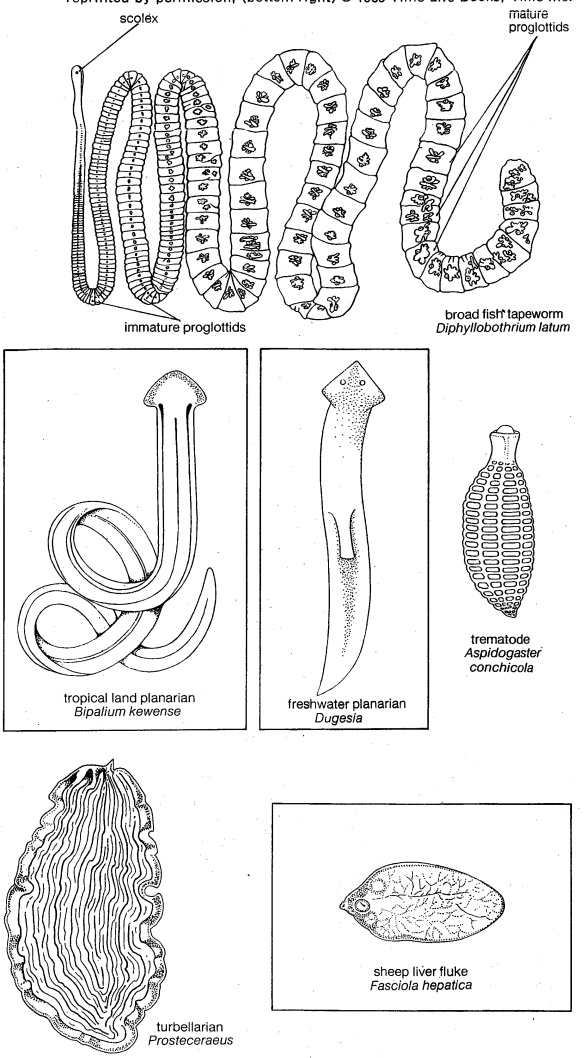


Figure 1: Representative body types of Platyhelminthes.

GENERAL FEATURES

Tapeworm, fluke, and trematode infestations

Importance. Platyhelminthes are of particular economic interest because many species of flukes are parasitic in man, in domestic animals, or in both. In Europe and in North and South America, tapeworm infestations of man have been greatly reduced as a consequence of routine meat inspection. But where sanitation is poor and meat eaten undercooked, the incidence of tapeworm infestations is high. In the Baltic countries much of the population is infested with the broad tapeworm (*Diphyllobothrium latum*); in parts of the southern United States a small proportion of the population may be infested with the dwarf tapeworm (*Hymenolepis nana*). In Europe and the United States the beef tapeworm (*Taenia saginata*) is not uncommon due to the habit of eating undercooked steaks or other beef products.

Immature stages (larvae) can seriously damage the host. A larval stage of the gid parasite of sheep (*Multiceps multiceps*) usually lodges in the sheep brain. Fluid-filled hydatid cysts (*i.e.*, sacs containing many cells capable of developing into new individuals) of *Echinococcus* may occur almost anywhere in the body of sheep. Hydatids of the liver, brain, or lung of man are often fatal. Infestation occurs only where man lives in close association with dogs, and the dogs have access to infested sheep for food.

Thirty-six or more species of flukes have been reported as parasitic in man. Endemic (local) centres of infection occur in virtually all countries; widespread infections occur in the Far East, Africa, and tropical America. Many species are ingested as cysts, called metacercariae, in uncooked food—*e.g.*, the lung fluke *Paragonimus westermani* found in crayfish and crabs, the intestinal flukes *Heterophyes heterophyes* and *Metagonimus yokogawai* and the liver fluke *Opisthorchis sinensis* in fish, and the intestinal fluke *Fasciolopsis buski* on plants. Free-swimming larvae (cercariae) of blood flukes penetrate the human skin directly. In man, these parasites and others listed below cause much misery and death. Control of certain flukes through the eradication of their mollusk hosts has been attempted but without much success.

Schistosomiasis (bilharziasis) is a major human disease caused by three species of the genus *Schistosoma*, known collectively as blood flukes. Africa and western Asia (*e.g.*, Iran, Iraq) are endemic centres for *S. haematobium; S. mansoni* also is found in these areas, as well as in the West Indies and South America. In the Far East, *S. japonicum* is the important blood fluke.

Among domestic animals, the sheep liver fluke (*Fasciola hepatica*) may cause debilitating and fatal epidemics (liver rot) in sheep. These animals become infected by eating metacercariae encysted on grass. Monogenea are common pests on fish in hatcheries and home aquariums (see also DISEASES OF ANIMALS).

Size range and diversity of structure. Most of the Turbellaria are less than five millimetres (one-fifth inch) long, and many are microscopic in size. The largest of this class are the planarians, which may reach 0.5 metre (about 20 inches) in length. Trematoda are mostly five to 100 millimetres (about one-fifth to four inches) long; some species, however, may grow to several centimetres. The smallest Cestoda are about one millimetre long, but some exceed 15 metres (50 feet) in length. The Monogenea range in length from 0.5 to 30 millimetres. Aspidocotylea are from a few millimetres to 100 millimetres in length.

Distribution and abundance. In general, free-living flatworms can occur wherever there is moisture. Except for one group of turbellarians, the temnocephalids, flatworms are cosmopolitan in distribution. They occur in both freshwater and salt water and occasionally in terrestrial habitats that are moist, especially in tropical and subtropical regions. The temnocephalids, which are parasitic on freshwater crustaceans, occur primarily in Central and South America, Madagascar, New Zealand, Australia, and islands of the South Pacific.

Some flatworm species occupy a very wide range of habitats. One of the most cosmopolitan and most tolerant of different ecological conditions is the turbellarian *Gyratrix hermaphroditus*, which occurs in freshwater at elevations from sea level to 7,000 feet (2,000 metres) as well as in saltwater pools. Adult forms of parasitic flatworms are confined almost entirely to specific vertebrate hosts; the larval forms, however, occur in vertebrates and in invertebrates, especially in mollusks, arthropods (*e.g.*, crabs), and annelids (*e.g.*, marine worms). They are cosmopolitan in distribution, but their occurrence is closely related to that of the intermediate host or hosts.

NATURAL HISTORY

Life cycle. *Reproduction.* With very few exceptions, platyhelminthes are hermaphroditic, and their reproductive systems are generally complex. Numerous testes but only one or two ovaries are usually present. The female system is unusual in that it is separated into two structures: the ovaries and the vitellaria, often known as the vitelline glands or yolk glands. In contrast, in most higher animals the yolk is part of the egg. The cells of the vitellaria form yolk and eggshell components. In some groups, particularly those that live primarily in water or have an aqueous phase in the life cycle, the eggshell consists of a hardened protein known as sclerotin, or tanned protein. Most of this protein comes from the vitellaria. In other groups, especially those that are primarily terrestrial or

Separation of the female reproductive system

have a terrestrial phase in their life cycle, the eggshells are composed of another protein, keratin, a tougher material that is more resistant to adverse environmental conditions.

The general structure of the reproductive system may be seen in Figure 2. In the tapeworms, the tapelike body is generally divided into a series of segments, or proglottids, each of which develops a complete set of male and female genitalia. A rather complex copulatory apparatus consists of an evertible (capable of turning outward) penis, or cirrus, in the male and a canal, or vagina, in the female. Near its opening the female canal may differentiate into a variety of tubular organs. Fertilized eggs are often stored in a saclike uterus, which may become greatly distended; in tapeworms, it may fill a whole segment.

Each male and female reproductive system may have its own external opening, or gonopore, or the terminal regions of each system may join to form a common genital atrium (Figure 2), or passage, and a genital pore.

Modes of fertilization

Either cross-fertilization (*i.e.*, involving two individuals) or self-fertilization may occur; self-fertilization is probably more common. Some free-living flatworms perform a type of copulation known as hypodermic impregnation, whereby the penis of one animal pierces the epidermis of another and injects sperm into the tissues. Some forms reproduce asexually—*i.e.*, they produce chains that eventually separate to form individuals.

Development. The life cycles of the free-living forms are relatively simple. Eggs are laid singly or in batches. Frequently they are attached to some object or surface by an adhesive secretion. After a period of embryonation (*i.e.*, the formation of the first larval stage) minute worms emerge, feed, develop genitalia, and become adult.

Parasitic platyhelminths undergo very complex life cycles, often involving several larval stages in other animals —the intermediate hosts; these hosts may be invertebrate or vertebrate.

The simplest cycle in parasitic platyhelminths occurs in the Monogenea, which have no intermediate hosts. The majority of the Monogenea are ectoparasitic (externally parasitic) on fish. The eggs hatch in water. The larva, known as an oncomiracidium, is heavily ciliated (has hairlike projections) and bears numerous posterior hooks. It attaches to a host, grows, and matures. In some species (*e.g.*, *Polystoma integerrimum*) parasitic in frogs, maturation of the genitalia is synchronized with maturation of the host and apparently is controlled by the endocrine system of the latter.

In the life cycle of flukes of the subclass Digenea, mollusks serve as the intermediate host. Eggs of Digenea usually hatch in water or in a snail host. The first larval stage, the miracidium, generally is free-swimming and penetrates a freshwater or marine snail, unless it has already been ingested by one. Within this intermediate host, the parasite passes through some or all of a series of further larval stages known as sporocysts, rediae, and cercariae. Cercariae are essentially young flukes with tails. They either penetrate the host or are ingested by it. Tapeworms of the subclass Eucestoda are generally transmitted from host to host by direct ingestion of eggs; by ingestion of intermediate hosts containing larval stages; and, very rarely, by passage of a larva from an intermediate host through a skin wound into another intermediate host.

Transmission to a human host through a skin wound is most likely to occur in Asia, where frogs infested with tapeworm larvae are sometimes used to treat wounds. The tapeworm, *Hymenolepis nana*, parasitic in rodents and man, can complete its life cycle without an intermediate host. Certain species of trematodes and cestodes show a tendency toward progenesis, in which adult features—such as the appearance of genital rudiments—appear in the larva. In some cases of progenesis the worm achieves sexual maturity in the intermediate host.

Regeneration. The ability to undergo tissue regeneration, beyond simple wound healing, occurs in two classes of Platyhelminthes: Turbellaria and Cestoda.

Turbellaria. Turbellarians, planaria particularly, have been commonly used in regeneration research. The greatest regenerative powers exist in species capable of asexual reproduction. Pieces from almost any part of the turbellarian *Stenostomum*, for example, can reorganize into completely new worms. In some cases regeneration of very small pieces may result in the formation of imperfect (*e.g.*, headless) organisms.

Polarity

In other Turbellaria, regeneration of the head is limited to pieces from the anterior region or to tissues containing the cerebral ganglion (brain). The region anterior to this ganglion is incapable of regeneration, but if cuts are made posterior to it, many species can replace the entire posterior region, including the pharynx and the reproductive system. In the cut pieces, polarity is retained; *i.e.*, the anterior zone of the cut piece regenerates the head and the posterior region regenerates the tail. If a region in front of the pharynx is transplanted into the posterior region of another individual, it influences that region to form a pharyngeal zone that eventually differentiates a pharynx. This new pharyngeal zone is now said to be determined and, if removed, will regenerate again into a new pharynx.

There is evidence that a special type of cell, a neoblast, is concerned in planarian regeneration. Neoblasts, rich in ribonucleic acid (RNA), which plays an essential role in cell division, appear in great numbers during regeneration. Similar cells, apparently inactive, occur in the tissues of whole organisms (see also REGENERATION, BIOLOGICAL).

Cestoda. Regeneration, although rare in the parasitic worms in general, does occur in the cestodes. Most tapeworms appear to be capable of regeneration from the head (scolex) and neck region. This property often makes it difficult to treat man for tapeworms; treatment may eliminate only the body, or strobila, leaving the scolex still attached to the intestinal wall of the host and thus capable of producing a new strobila, which re-establishes the infestation. Several species of cestode larvae can regenerate themselves from cut regions. A branched larval form of *Sparganum prolifer*, a human parasite, may undergo both asexual multiplication and regeneration.

Ecology. Turbellaria are adapted to a wide range of environments, and many species are resistant to extreme environmental conditions. Some occur in coastal marine habitats—in sand, on or under rocks, and in or on other animals or plants. Some marine species occur at relatively great depths in the sea; others are pelagic (*i.e.*, living in the open sea). Freshwater species are found in ponds, lakes, rapidly flowing rivers, and streams. Temporary freshwater pools may contain adult forms that survive periods of dryness in an encysted state. Some aquatic species exhibit considerable tolerance to osmotic changes —*i.e.*, to differences in salt concentrations of the water; a marine species *Coelogynopora biarmata*, for example, has also been found in freshwater springs.

Terrestrial species of Turbellaria occur in soil, moist sand, leaf litter, mud, under rocks, and on vegetation. Some have been found in pools in the desert and in caves. Cave-dwelling species tend to show loss of eyes and pigment.

Tolerance of temperature extremes

Some species are able to stand considerable temperature ranges; for example, *Crenobia alpina*, which occurs in alpine streams, apparently can survive temperatures of −40° to −50° C (−40° to −58° F). Remarkable heat tolerance is exhibited by *Macrostomum thermale* and *Microstomum lineare*, which are found in hot springs at 40°–47° C (104°–117° F). *M. lineare* can also tolerate temperatures as low as 3° C (37° F).

Many Turbellaria live in association with plants and animals. Marine algae, for example, frequently harbour many species, often in large numbers. Turbellarians most commonly associate with animals such as echinoderms (*e.g.*, starfish), crustaceans (*e.g.*, crabs), and mollusks. Less commonly, associations occur with sipunculids (tiny marine invertebrates), annelid worms, arachnids (*e.g.*, spiders), coelenterates (*e.g.*, jellyfish), other turbellarians, and lower vertebrates. An interesting feature of these associations is that species within a turbellarian family tend to associate with one type of organism; for example,

almost all members of the family Umagillidae associate with echinoderms.

In a few cases, the association is considered parasitic; *i.e.*, the turbellarians obtain all nourishment from the host. Most of these species belong to the order Rhabdocoela, in which the alimentary canal is either absent or reduced. In a few hosts they cause so-called parasitic castration—*i.e.*, suppression of gonad maturation in the host organism.

Among the turbellaria that are parasitic or commensal (*i.e.*, living in close association with but not harmful to another organism) the Temnocephalida are best adapted for attachment to other organisms. They have a large saucer-shaped posterior adhesive organ and anterior tentacles that are also used for adhesion. All temnocephalids occur on freshwater hosts, mainly crustaceans but also mollusks, turtles, and jellyfish.

The tendency to associate with other animals apparently represents a definite evolutionary trend among the platyhelminths; permanent associations essential to the survival of a species could develop from loose associations, which could also give rise to parasitic forms, such as the Trematoda and Cestoda. The free-living larval stages that frequently occur in these groups play a major role in the dissemination of the species.

The ecology of the parasitic groups (*i.e.*, Cestoda and Trematoda) is particularly complex, because as many as four hosts may be involved in the life cycle. In the case of the broad tapeworm, for example, man serves as the final (or definitive) host, various species of fish as one intermediate host, and species of a small water crustacean (*Cyclops*) as another intermediate host. It is clear that the broad tapeworm (*Diphyllobothrium latum*) can occur only where an intimate ecological association exists among the three host groups.

Adaptation to the microenvironment of the host

In addition to adapting to the general external environment, individuals at each stage of the life cycle of a parasite must adapt to the micro-environment provided by the host. Adaptations include not only obvious features, such as suckers or hooks for attachment, but also those associated with the biochemical and physiological conditions imposed by the host. Parasites frequently utilize the physiological and biochemical properties of a new host, especially those that differ markedly from the external environment, in order to trigger the next developmental stage—*e.g.*, several species of cestodes are stimulated to mature sexually by the high body temperature (40° C) of their bird host, which contrasts sharply with the low body temperature of the cold-blooded fish host of the larval stage. The unusually intimate association of certain flukes (subclass Digenea) with mollusks suggests that flukes were originally parasites of mollusks and that they later developed an association with other hosts.

Knowledge of the ecology both of a platyhelminth parasite and of its intermediate host(s) is essential if control measures against the pest are to be effective. Man's alterations of the environment sometimes increase the occurrence of platyhelminth diseases. The Aswan High Dam in Egypt, for example, has produced conditions especially favourable for the breeding of the snail that serves as the intermediate host of the blood fluke (*Schistosoma mansoni*). In this case, as with many trematode infestations, man exposes himself to the disease by bathing in water containing infective larvae (cercariae) released from infested snails; the cercariae enter directly through the skin. Certain other human diseases of platyhelminth origin—such as hydatid (cyst) disease, caused by the tapeworm *Echinococcus granulosus*—owe their survival and dissemination to man's close ecological association with dogs.

In the parasitic platyhelminth groups (*e.g.*, the Monogenea) that do not normally utilize intermediate hosts, there is a close ecological association between egg release and production of young of both the parasite and its host; infection of the next generation of host could not otherwise occur.

Many platyhelminths show highly specific adaptations to internal host environments. Many monogeneans, for example, show a marked preference for a particular gill arch in a fish. The scolex (head) of certain tapeworms of elasmobranch fishes (*e.g.*, sharks, skates, and rays) is highly specialized and can satisfactorily attach only to the gut of a fish possessing a complementary structure.

FORM AND FUNCTION

External features. Some Turbellaria are gray, brown, or black, with mottled or striped patterns. Others, which contain algae in the mesenchyme, are green or brown. Parasitic flatworms usually have no pigment, but cestodes may be coloured by food (*e.g.*, bile, blood) in their gut. Some parasitic forms may show masses of dark eggs through a translucent, creamy, or whitish tissue.

Reprinted with permission of The Macmillan Company from *Invertebrate Zoology* by R. Hegner and J. Engemann. Copyright © by The Macmillan Company, 1968

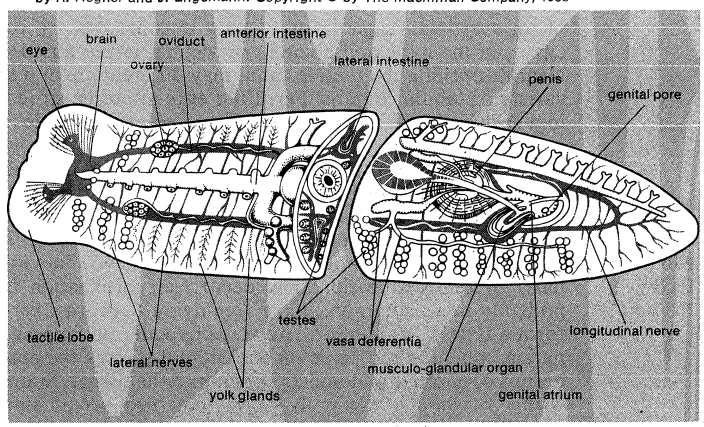


Figure 2: General structure of a platyhelminth, as exemplified by a freshwater turbellarian, *Procotyla fluviatilis*.

The flatworm body plan

The typical flatworm body is flatened and leaflike or tapelike (Figure 2). The head may be set off from the body or grade imperceptibly into it. The anterior (head) end can usually be distinguished from the posterior end in free-living forms by the presence of two pigment spots, which are primitive eyes. In the case of the tapeworm, the scolex is usually conspicuous for its breadth. The strobila (body) typically consists of proglottids, each of which is usually a self-sufficient reproducing unit with all of the sexual organs necessary to reproduce. The number of proglottids may vary from three in some species to several hundreds in others. Organs of attachment on the scolex may, in addition to suckers, consist of hooks, spines, or various combinations of these.

The structure and function of the body covering, or tegument, differs markedly between free-living and parasitic forms. In free-living forms, the body covering is typically an epidermis consisting of one layer of ciliated cells—*i.e.*, cells with hairlike structures—the cilia being confined to specific regions in some species. In the parasitic groups—flukes, tapeworms, and monogeneans—the tegument shows striking modifications associated with the parasitic way of life. It once was thought that the tegument is a nonliving secreted layer; it is now known, however, that the tegument of parasites is (Figure 3) metabolically active and consists of cells not separated from

From (Left) R. Hegner and J. Engemann, *Invertebrate Zoology*, (Right) J. Smyth, *The Physiology of Cestodes*

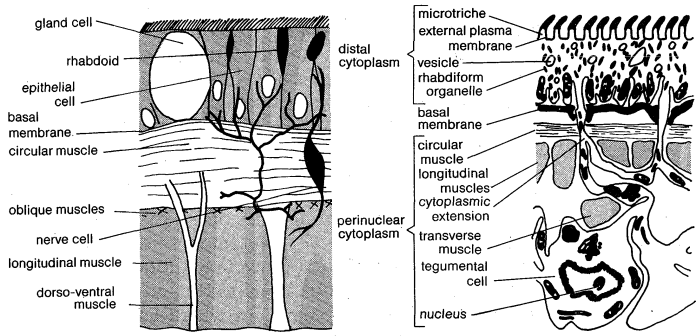


Figure 3: Comparison of the body wall in a (A) free-living and a (B) parasitic platyhelminth. A is a turbellarian, B is a tapeworm.

one another by cell walls (*i.e.*, a syncytium). The tegument itself consists of cytoplasmic extensions of tegumental cells, the main bodies of which lie in what may be described as the "subcuticular" zone, although a true cuticle is not present. A membrane separates the inner zone of the tegumental cells, the so-called perinuclear cytoplasm, from the surface syncytium, or distal cytoplasm (Figure 3).

The surface of tapeworms and monogeneans is drawn out into spinelike structures called microtriches, or microvilli. The microtriches probably help to attach the parasite to the gut of the host, absorb nutritive materials, and secrete various substances. In the flukes, microtriches are lacking, but spines are frequently present.

Embedded in the epidermis of turbellarians are ovoid or rod-shaped bodies (rhabdoids, see Figure 3) of several sorts; of uncertain function, the bodies frequently are concentrated dorsally or may be clustered anteriorly as rod tracts opening at the apex. Rhabdoids are absent in flukes and tapeworms.

Internal features. Beneath the epidermis of turbellarians is a homogeneous or lamellated basal membrane. Club-shaped mesenchymal gland cells, opening externally, generally are present in all flatworms. In turbellarians two major types of mesenchymal glands occur: one produces a slimy material upon which the organisms creep; the other secretes an adhesive substance for capture of prey, for adhesion, and for cementing egg capsules to a suitable surface. The larvae of parasitic forms generally possess similar glands whose secretions are used for adhesion, for producing cyst walls around resting stages, and for penetrating hosts; some adult parasites have glands for adhesion and, in trematodes, for softening and digesting host tissues.

Mesenchyme

The mesenchyme consists of fixed and free cells as well as a fibrous matrix. A fluid occupies the minute open spaces and serves for distribution of nutrients and wastes. The mesenchymal cells in certain groups may differentiate during growth to become sex cells or may function in asexual reproduction in repair or in regeneration.

Nervous system. The main ganglia, or nerve centres, of the nervous system and the major sense organs are generally concentrated at the anterior end. Typically, the primitive brain of the flatworm consists of a bilobed mass of tissue with lateral longitudinal nerve cords connected by transverse connectives, thus forming a rather ladderlike structure or grid running the greater length of the organism (Figure 2). Free-living forms commonly have two longitudinal cords, but some tapeworms have as many as ten. Sensory receptors occur in all groups.

Musculature. The well-developed muscular system present in flatworms is comprised of a subcuticular musculature consisting of layers of circular, longitudinal, and diagonal muscles close to the epidermis, and a mesenchymal musculature consisting of dorsoventral, transverse, and longitudinal fibres passing through the mesenchyme. In general, platyhelminths are capable of extensive body contraction and elongation.

Digestive and excretory systems. The blind-ending intestine of trematodes consists of a simple sac with an anterior or midventral mouth or a two-branched gut with an anterior mouth; an anus is usually lacking, but a few species have one or two anal pores. Between the mouth and the intestine are often a pharynx and an esophagus receiving secretions from glands therein. The intestine proper, lined with digestive and absorptive cells, is surrounded by a thin layer of muscles that effect peristalsis; *i.e.*, they contract in a wavelike fashion down the length of the intestine. In many larger flukes lateral intestinal branches, or diverticula, bring food close to all internal tissues. Undigested residue passes back out of the mouth.

The digestive system of Turbellaria typically consists of mouth, pharynx, and intestine. In the order Acoela, however, only a mouth is present; food passes directly from the mouth into the parenchyma, to be absorbed by the mesenchymal cells. Tapeworms lack digestive systems.

The excretory system consists of protonephridia. These are branching canals ending in so-called flame cells—hollow cells with bundles of constantly moving cilia.

Nutrition. *Free-living forms.* Free-living platyhelminths (Turbellaria), mostly carnivorous, are particularly adapted for the capture of prey. Their encounters with prey appear to be largely fortuitous, except in some species that release ensnaring mucus threads. Because they have developed various complex feeding mechanisms, most turbellarians are able to feed on organisms much larger than themselves, such as annelids, arthropods, mollusks, and tunicates (*e.g.*, sea squirts). In general, the feeding mechanism involves the pharynx which, in the most highly developed forms, is a powerful muscular organ that can be protruded through the mouth. Flatworms with a simple ciliated pharynx are restricted to feeding on small organisms such as protozoans and rotifers, but those with a muscular pharynx can turn it outward, thrust it through the tegument of annelids and crustaceans, and draw out their internal body organs and fluids. Turbellaria, with a more advanced type of pharynx, can extend it over the captured prey until the animal is completely enveloped.

Complex feeding mechanisms

Digestion is both extracellular and intracellular. Digestive enzymes (biological catalysts), which mix with the food in the gut, reduce the size of the food particles. This partly digested material is then engulfed (phagocytized) by cells or absorbed; digestion is then completed within the gut cells.

Parasitic forms. In the parasitic groups with a gut (Trematoda and Monogenea), both extracellular and intracellular digestion occur. The extent to which these processes take place depends on the nature of the food. When fragments of the host's food or tissues other than fluids or semifluids (*e.g.*, blood and mucus) are taken as nutrients by the parasite, digestion appears to be largely extracellular. In those that feed on blood, digestion is largely intracellular, often resulting in the deposition of hematin, an insoluble pigment formed from the breakdown of hemoglobin. This pigment is eventually extruded by disintegrating gut cells.

Despite the presence of a gut, trematodes seem able to absorb certain materials, such as glucose, through the metabolically active tegument covering the body surface. The tapeworms, which have no gut, must take in all foods through the tegument. The entry of amino acids (the structural units of proteins) and small molecules of carbohydrate (*e.g.*, sugars) through this tegument occurs by a mechanism known as active transport, in which molecules are taken up against a concentration gradient. This process, similar to that in the vertebrate gut, requires the expenditure of energy. There is also some evidence that cestodes may be able to digest materials in contact with the tegument by means of so-called membrane digestion, a little-understood process.

Metabolism. Both free-living and parasitic platyhelminths utilize oxygen when it is available. Most of the parasitic platyhelminths studied have a predominantly anaerobic metabolism (*i.e.*, not dependent upon oxygen). This is true even in species found in habitats—such as the bloodstream—where oxygen is normally available.

Parasitic platyhelminths are made up of the usual tissue constituents—protein, carbohydrates, and lipids—but, compared to other invertebrates, the proportions differ somewhat; *i.e.*, the carbohydrate content tends to be relatively high and the protein content relatively low. In larval and adult cestodes, carbohydrate occurs chiefly as animal starch, or glycogen, which acts as the main source of energy for species in low oxygen habitats. The level of glycogen, like other chemical constituents, can fluctuate considerably, depending on the diet or feeding habits of the host. In some species, more than 40 percent of the worm's dried weight is glycogen.

Because carbohydrate metabolism is important in parasitic flatworms, a substantial amount of carbohydrate must be present in the host diet to assure normal growth of the parasite. Hence the growth rate of the rat tapeworm, *Hymenolepis diminuta*, is a good indicator of the quantity of carbohydrate ingested by the rat. Experiments have shown that most parasitic worms can utilize only certain types of carbohydrate. All tapeworms studied thus far utilize the sugar glucose. Many can also utilize

galactose, but only a few can utilize maltose or sucrose.

An unusual constituent of both trematodes and cestodes is a round or oval structure called a calcareous corpuscle; large numbers of them occur in the tissues of both adults and larvae. Their function has not yet been established, but it is believed that they may act as reserves for such substances as calcium, magnesium, and phosphorus.

The chief proteins in cestodes and trematodes are keratin and sclerotin. Keratin forms the hooks and part of the protective layers of the cestode egg and the cyst wall of certain immature stages of trematodes. Sclerotin occurs in both cestode and trematode eggshells, especially in those that have larval stages associated with aquatic environments.

Requirements for egg hatching

Platyhelminth eggs hatch in response to a variety of different stimuli in different hosts. Most trematode eggs require oxygen in order to form the first larval stages and light in order to hatch. Light is thought to stimulate the release of an enzyme that attacks a cement holding the lid (operculum) of the egg in place. A similar mechanism probably operates in cestodes (largely of the order Pseudophyllidea) whose life cycles involve aquatic intermediate hosts or definitive hosts, such as birds or fish.

In many cestodes, especially those belonging to the order Cyclophyllidea, the eggs hatch only when ingested by the host. When the host is an insect, hatching sometimes is apparently purely mechanical, the shell being broken by the insect's mouthparts. In vertebrate intermediate hosts, destruction of the shell depends largely on the action of the host's enzymes. Activation of the embryo within the shell and its subsequent release depend on other factors, including the amount of carbon dioxide present, in addition to the host's enzymes. Factors involving a vertebrate host are also important in establishing trematode or cestode infections after encysted or encapsulated larval stages have been ingested. Under the influence of the same factors, tapeworm larvae are stimulated to evaginate their heads (*i.e.*, turn them inside out, so to speak), a process that makes possible their attachment to the gut lining.

EVOLUTION AND CLASSIFICATION

Evolutionary hypotheses. The origin of the platyhelminths and the evolution of the various classes have not been positively settled. There are, however, two main lines of thought. According to the more widely accepted view, the Turbellaria and an aberrant group, the Ctenophora, have been derived independently from the lower Coelenterata. In the alternative view, the Turbellaria are derived from the Ciliata and the Ctenophora from the Turbellaria.

Origin of Turbellaria

It is generally believed that the parasitic groups are derived from the Turbellaria, many of which form close associations with other animals. These associations often show host specificity, a characteristic of truly parasitic forms. There are a number of views regarding the evolutionary relationships among the various parasitic groups. One school of thought proposes that rhabdocoel turbellarians gave rise to monogeneans; these, in turn, gave rise to digeneans, from which the cestodes were derived. Another view is that the rhabdocoel ancestor gave rise to two lines; one gave rise to monogeneans, who gave rise to digeneans, and the other line gave rise to cestodes. A further modification of the latter view, based largely on the study of the larval forms, proposes that cestodes were derived from monogeneans.

In considering the evolution of the parasitic groups, the digeneans should be mentioned in particular. The life cycle of digeneans is dominated by mollusks, which invariably act as intermediate hosts. This condition has led to the widely accepted view that digeneans were originally parasites of mollusks and later formed an association with vertebrates; the vertebrates, in turn, became incorporated into the life cycle as definitive hosts.

Classification. *Distinguishing taxonomic features.* The phylum Platyhelminthes is grouped with several other phyla (Nemertea, Rotifera, Nematoda, Gastrotricha, Acanthocephala, and Nematomorpha) as acoelomate triploblasts—*i.e.*, lacking a body cavity and with three embryonic layers: endoderm, mesoderm, and ectoderm. These phyla also represent animals that are unsegmented and whose body space tends to be occupied by mesenchyme. In Platyhelminthes, mesenchyme cells completely fill this space.

In the classification of platyhelminths the principal criteria are: habitat of organism (*i.e.*, free-living or parasitic); the characteristics of the body covering; the form and position of organs for attachment to host (when present); the presence or absence of segmentation; the form of the reproductive system, especially with respect to vitellaria (yolk glands) and the number of testes; the presence or absence of an alimentary canal; the characteristics of the pharynx (when present); and the nature of protective egg membranes.

Annotated classification. There is no unanimity concerning the classification of platyhelminths. The following classification is based partly on that of Grassé, Hyman, Yamaguti, and La Rue (see below *Bibliography*).

PHYLUM PLATYHELMINTHES (flatworms)

Flat, unsegmented worms; gastrovascular cavity and respiratory, skeletal, and circulatory systems absent; excretion by means of flame-bulb protonephridia; mesenchyme fills all spaces between organ systems; a variable number of longitudinal nerve cords with transverse connectives; body structure triploblastic (*i.e.*, 3 embryonic layers); reproductive system hermaphroditic and complex.

Class Turbellaria

Epidermis usually ciliated at least in part, provided with rhabdoids (minute rodlike structures); body unsegmented; gut present except in order Acoela; life cycle simple; mostly free-living, some ectocommensal, endocommensal (*i.e.*, living, respectively, outside or inside another organism without harming it), or parasitic; about 3,000 species.

Order Acoela. Exclusively marine; mouth present; pharynx simple or lacking; no intestine; without protonephridia, oviducts, yolk glands, or definitely delimited gonads.

Order Rhabdocoela. Saclike intestine; protonephridia and oviducts usually present; gonads few, mostly compact; nervous system generally with 2 longitudinal trunks.

Order Alloecoela. Pharynx simple, bulbose or plicate (many ridges); intestine may have short diverticula, or pockets; protonephridia paired; testes usually numerous; penis papilla generally present; nervous system with 3–4 trunks.

Class Monogenea

Oral sucker lacking or weakly developed; posterior end with large posterior adhesive disk (opisthaptor) usually provided with hooks; excretory pores paired, anterior and dorsal; parasites of the skin and other superficial locations, especially on the gills of fish; life cycle simple, no alternation of hosts; about 1,350 species.

Class Aspidocotylea (Aspidobothria)

Oral sucker absent; main adhesive organ occupying almost the entire ventral surface and consisting of suckerlets arranged in rows; excretory pore single and posterior; endoparasites of vertebrates (especially turtles and fish), mollusks, and crustaceans; about 100 species.

Class Cestoda (tapeworms)

Elongated endoparasites with alimentary canal lacking; epidermis modified for absorption and secretion; usually divided into segments (proglottids); adhesive organs limited to anterior end; except in Cestodaria, adult stages almost entirely parasites of vertebrates; life cycles complicated with 1 or more intermediate hosts; about 3,500 species.

Subclass Cestodaria

Unsegmented tapeworms containing 1 set of genitalia; parasites of the body cavity or intestine of annelid worms or fish.

Order Amphilinidea. Uterus long and N-shaped; genital pores at or near posterior extremity; intestinal parasites of teleosts (bony fish).

Order Caryophyllidea. Uterus a coiled tube; genital pore well separated from posterior extremity; intestinal parasites of teleosts, occasionally in annelids.

Order Gyrocotylidea. Testes confined to anterior region; genital pores near anterior end; parasitic in intestine of fish of the genus *Chimaera*.

Subclass Eucestoda

Polyzoic tapeworms with scolex (head) of varying structure; body usually with distinct external segmentation; parasitic in intestine of vertebrates.

Order Tetraphyllidea. Scolex with 4 bothridia (leaflike muscular structure); vitellaria located in lateral margins of proglottids; genital pores lateral; parasites of elasmobranchs.

Order Lecanicephalidea. Reproductive system similar to

Tetraphyllidea, but scolex divided into an upper disklike or globular part and a lower collar-like part bearing 4 suckers; mainly parasites of elasmobranchs.

Order Proteocephalidea. Scolex with 4 suckers, sometimes a 5th terminal one; vitellaria located in lateral margins; genital pores lateral; mainly parasites of cold-blooded vertebrates.

Order Diphyllidea. Two bothridia, each sometimes bisected by a median longitudinal ridge; large rostellum (cone-shaped or cylindrical projection) armed with dorsal and ventral groups of large hooks; cephalic peduncle (fleshy stalk on head) with longitudinal rows of T-shaped hooks; genital pore median, parasitic in elasmobranchs; 1 genus, *Echinobothrium.*

Order Trypanorhyncha. Scolex with 2 or 4 bothridia; vitellaria in continuous sleevelike distribution; parasites of elasmobranchs.

Order Pseudophyllidea. Scolex with 2 elongated, shallow bothria, 1 dorsal and 1 ventral; genital pore lateral or median. Vitellaria lateral or extending across proglottid and encircling other organs; parasites of teleosts and land vertebrates.

Order Nippotaeniidea. Scolex bears 1 apical sucker; parasites of freshwater fish; 1 genus, *Nippotaenia.*

Order Cyclophyllidea. (Taenoidea) Scolex with 4 suckers; no uterine pores; 1 compact vitellarium behind ovary; mainly parasites of birds and mammals.

Order Aporidea. No sex ducts or genital openings; parasites of swans.

Order Spathebothriidea. Scolex without true bothria or suckers; strobila with internal segmentation but no external segmentation; parasites of marine teleosts.

Class Trematoda (flukes)

Ectoparasites or endoparasites; no ciliated epidermis; body undivided; adhesive organs well-developed; life cycles generally complex with 2 or more hosts, about 6,250 species.

Subclass Aspidogastrea

Oral sucker absent; main adhesive organ occupying almost the entire ventral surface, consists of suckerlets arranged in rows; excretory pore single and posterior; endoparasites of vertebrates, mollusks, and crustaceans.

Subclass Digenea

Oral and ventral suckers generally well-developed; development involves at least 1 intermediate host; usually endoparasites of vertebrates.

Superorder Anepitheliocystidia. Excretory bladder thin walled (*i.e.*, not epithelial in any stage); excretory pores open to exterior in tail; stylet always lacking.

Order Strigeatoidea. Cercaria (immature form) fork-tailed; penetration glands present; 1-2 pairs of protonephridia.

Order Echinostomida. Cercaria with simple tail and many cyst-producing glands; life cycle with 3 hosts.

Superorder Epitheliocystidia. Cercaria completely lacking caudal excretory vessels at any stage of development; stylet (piercing structure) present or lacking.

Order Plagiorchida. Cercaria typically armed with a stylet; encystment in invertebrates, rarely vertebrates; excretory vessels not open to the exterior.

Order Opisthorchiida. Cercaria never armed; excretory pores open on margins of tail.

Critical appraisal. There is considerable disagreement on many aspects of the taxonomy of Platyhelminthes, especially regarding class divisions. Among the free-living or semiparasitic forms, the temnocephalids are placed as an order among the class Turbellaria (as here) by some authorities but are treated as a separate class by others. It is now generally recognized that the monogeneans, previously placed in a class of the Trematoda, are sufficiently divergent from the digeneans, nearer to the cestodes and turbellarians, to be placed in a separate class (Monogenea). The classification of the digeneans that has been used here is based on the development of the excretory bladder and associated structures. Such a scheme, however, clearly raises taxonomic problems of identification for species whose life cycles are not yet fully known.

BIBLIOGRAPHY. B. DAWES, *The Trematoda* (1946), a classic work, largely taxonomic with some general biology, but somewhat out of date; P.P. GRASSE, (ed.), "Plathelminthes, mésozoaires, acanthocéphales, némertiens," in *Traité de zoologie*, vol. 4 (1961); and L.H. HYMAN, *The Invertebrates*, vol. 2, *Platyhelminthes and Rhynchocoela* (1951), two monumental classics; G.R. LA RUE, "The Classification of Digenetic Trematoda: A Review and a New System," *Expl. Parasit.*, 6:306–344 (1957); J.D. SMYTH, *The Physiology of Trematodes* (1966) and *The Physiology of Cestodes* (1969), detailed monographs for students; R.A. WARDLE and J.A. MCLEOD, *The Zoology of Tapeworms* (1952), a classic work, largely taxonomic, now somewhat out of date; S. YAMAGUTI, *Systema Helminthum*, vol. 1, *The Digenetic Trematodes of Vertebrates*, 2 pt. (1958), vol. 2, *The Cestodes of Vertebrates* (1959), vol. 4, *Monogenea and Aspidocotylea* (1963), classic taxonomic works with detailed keys.

(J.D.Sm.)

Plautus

Little is known for certain about the life and personality of Plautus, who ranks with Terence as one of the two great Roman comic dramatists. His work, moreover, presents scholars with a variety of textual problems, since the manuscripts by which his plays survive are corrupt and sometimes incomplete. Nevertheless, his literary and dramatic skills make his plays enjoyable in their own right, while the achievement of his comic genius has had lasting significance in the history of Western literature and drama.

According to the grammarian Festus (2nd or 3rd century AD), Plautus was born at Sarsina, Umbria, in northeastern central Italy. He is customarily assigned the dates *c.* 254–184 BC. These dates, however, are largely based on statements made by later Latin writers, notably Cicero in the 1st century BC. Even the three names usually given to him—Titus Maccius Plautus—are of questionable historical authenticity. Internal evidence in some of the plays does, it is true, suggest that these were the names of their author; but it is possible that they are stage names, even theatrical jokes or allusions ("Maccus," for example, was the traditional name of the clown in the "Atellan farces," a long-established popular burlesque, native to the Neapolitan region of southern Italy: "Plautus," according to Festus, derives from *planis pedibus, planipes* (flat-footed) being a pantomime dancer. There are further difficulties: the poet Accius (170–*c.* 86 BC), who made a study of his fellow Umbrian, seems to have distinguished between one Plautus and one Titus Maccius. Tradition has it that Plautus was associated with the theatre from a young age. An early story says that he lost the profits made from his early success as a playwright in an unsuccessful business venture, and that for a while afterward he was obliged to earn a living by working in a grain mill.

Approach to drama. The Roman predecessors of Plautus in both tragedy and comedy borrowed most of their plots and all of their dramatic techniques from Greece. Even when handling themes taken from Roman life or legend, they presented these in Greek forms, setting, and dress. Plautus, like them, took the bulk of his plots, if not all of them, from plays written by Greek authors of the late 4th and early 3rd centuries BC (who represented the "New Comedy," as it was called), notably by Menander and Philemon. Plautus did not, however, borrow slavishly; although the life represented in his plays is superficially Greek, the flavour is Roman, and Plautus incorporated into his adaptations Roman concepts, terms, and usages. He referred to towns in Italy; to the gates, streets, and markets of Rome; to Roman laws and the business of the Roman law courts; to Roman magistrates and their duties; and to such Roman institutions as the Senate. Not all references, however, were romanized: Plautus apparently set little store by consistency, despite the fact that some of the Greek allusions that were left may have been unintelligible to his audiences. Terence, the more studied and polished playwright, mentions Plautus' carelessness as a translator and upbraids him for omitting an entire scene from one of his adaptations from the Greek (though there is no criticism of him for borrowing material, such plagiarism being then regarded as wholly commendable). Plautus allowed himself much other freedom in adapting his material, even combining scenes from two Greek originals into one Latin play (a procedure known as *contaminatio*).

Sources of the plays

Language. But even more important was Plautus' approach to the language in which he wrote. His action was lively and slapstick, and he was able to marry the action to the word. In his hands, Latin became racy and collo-

quial, verse varied and choral. Whether these new characteristics derived from now lost Greek originals—more vigorous than those of Menander—or whether they stemmed from the established forms and tastes of burlesque traditions native to Italy, cannot be determined with any certainty. The latter is the more likely. The result, at any rate, is that Plautus' plays read like originals rather than adaptations, such is his witty command of the Latin tongue—a gift admired by Cicero himself. It has often been said that Plautus' Latin is crude and "vulgar," but it is in fact a literary idiom based upon the language of the Romans in his day.

Themes of the plays

Plot. The plots of Plautus' plays are sometimes well-organized and interestingly developed, but more often they simply provide a frame for scenes of pure farce, relying heavily on intrigue, mistaken identity, and similar devices. Plautus is a truly popular dramatist, whose comic effect springs from exaggeration, burlesque, often coarse humour, rapid action, and a deliberately upside-down portrayal of life, in which slaves give orders to their masters, parents are hoodwinked to the advantage of sons who need money for girls, and the procurer or braggart soldier is outwitted and fails to secure the seduction or possession of the desired girls. But Plautus can also recognize the virtue of honesty (as in *Bacchides*), of loyalty (as in *Captivi*), and of nobility of character (as in the heroine of *Amphitruo*).

Verse. Plautus' plays, almost the earliest literary works in Latin that have survived, are written in verse, as were the Greek originals. The metres he used included the iambic six foot line (*senarius*) and the trochaic seven foot line (*septenarius*), which Menander had also employed. But Plautus varied these with longer iambic and trochaic lines and more elaborate rhythms. The metres are skillfully chosen and handled to emphasize the mood of the speaker or the action. Again, it is possible that now lost Greek plays inspired this metrical variety and inventiveness, but it is much more likely that Plautus was responding to features already existing in popular Italian dramatic traditions. The *Senarii* (conversational lines) were spoken, but the rest was sung or chanted to the accompaniment of double and fingered reed pipes. It could indeed be said that, in their metrical and musical liveliness, performances of Plautus' plays somewhat resembled musical comedy of the 1950s and '60s.

Influence. Plautus' original texts did not survive. Even by the time that Roman scholars such as Varro, a contemporary of Cicero, became interested in the playwright, only acting editions of his plays remained. These had been adapted, modified, cut, expanded, and generally brought up-to-date for production purposes. Critics and scholars have ever since attempted to establish a "Plautine" text, but 20th-century editors have admitted the impossibility of successfully accomplishing such a task. The plays had an active stage life at least until the time of Cicero and were occasionally performed afterward. Whereas Cicero had praised their language, the poet Horace, writing in the early years of the empire under its founder, the emperor Augustus, is a more severe critic and considered the plays to lack polish. There was renewed scholarly and literary interest in Plautus during the 2nd century AD, but it is unlikely that this was accompanied by a stage revival, though a performance of *Casina* is reported to have been given in the early 4th century. St. Jerome, toward the end of that century, says that after a night of excessive penance he would read Plautus as a relaxation; in the mid-5th century, Sidonius Apollinaris, a Gallic bishop who was also a poet, found time to read the plays and praise the playwright amid the alarms of the barbarian invasions. During the Middle Ages, Plautus was little read—if at all—in contrast to the popular Terence. By the mid-14th century, however, the humanist scholar and poet Petrarch knew eight of the comedies. As the remainder came to light, Plautus began to influence European domestic comedy after the Renaissance poet Ariosto had made the first imitations of Plautine comedy in the Italian vernacular. His influence was perhaps to be seen at its most sophisticated in the comedies of Molière (whose play *L'Avare*, for instance, was based on *Aulularia*), and it can be traced up to the present day—witness such adaptations as Jean Giraudoux's *Amphitryon 38* (1929) or Cole Porter's musical *Out of This World* (1950). Plautus' stock character "types" have similarly had a long line of successors: the braggart soldier of *Miles Gloriosus*, for example, became the "Capitano" of the Italian commedia dell'arte, is recognizable in Nicholas Udall's *Ralph Roister Doister* (16th century), in Shakespeare's Pistol, and even in his Falstaff, in Rostand's *Cyrano de Bergerac* (1897), in Bernard Shaw's Sergius in *Arms and the Man* (1894), while a trace of the character perhaps remains in Bertolt Brecht's Eilif in *Mother Courage and her Children* (1941). Thus, Plautus, in adapting Greek "New Comedy" to Roman conditions and taste, also significantly affected the course of the European theatre.

Influence on European domestic comedy

MAJOR WORKS

Twenty-one comedies (most dates uncertain) have survived: *Amphitruo* (*Amphitryon*); *Asinaria* (*The Comedy of Asses*); *Aulularia* (*The Pot of Gold*); *Bacchides* (*The Two Bacchises*); *Captivi* (*The Captives*); *Casina; Cistellaria* (*The Casket Comedy*); *Curculio; Epidicus; Menaechmi* (*The Two Menaechmuses*); *Mercator* (*The Merchant*); *Miles Gloriosus* (*The Braggart Warrior*); *Mostellaria* (*The Haunted House*); *Persa* (*The Persian*); *Poenulus* (*The Little Carthaginian*); *Pseudolus* (first produced 191 BC); *Rudens* (*The Rope*); *Stichus* (first produced 200 BC); *Trinummus* (*Three Bob Day*); *Truculentus; Vidularia*, fragmentary (*The Tale of a Travelling Bag*).

There is an English translation of all the extant plays in the "Loeb Series" (1916–38) and a French translation in the "Budé Series" (1932–40). Several of the plays have been translated in the "Penguin Classics Series." Many other translations of specific plays are available.

BIBLIOGRAPHY. Full bibliography is in G.E. DUCKWORTH, *The Nature of Roman Comedy*, pp. 447–464 (1952); a more recent but selective one may be found in MASON HAMMOND, ARTHUR M. MACK, and WALTER MOSKALEW, *Plautus: Miles Gloriosus*, rev. ed. pp. 59–66 (1970). Still fundamental are the collections of ancient evidence and the discussions in MARTIN SCHANZ, 4th ed. by CARL HOSIUS, *Geschichte der römischen Literatur I: Die römischen Literatur in der Zeit der Republik*, pp. 55–86 (1927); and the article "Maccius" by A.F. SONNENBERG in *Pauly–Wissowa Realencyclopädie*, vol. 14, col. 95–126 (1928). Useful is the briefer article "Plautus" by G.W. WILLIAMS in the *Oxford Classical Dictionary*, 2nd ed., pp. 843–844 (1970). In addition to Duckworth (above), a good account is in WILLIAM BEARE, *The Roman Stage*, 3rd ed. (1964). A standard text is that by M.W. LINDSAY, 2 vol. (1904 and reprints). The Loeb edition with English translation by PAUL NIXON, 5 vol. (1916–38 and reprints), unfortunately used the text of FRIEDRICH LEO, 2nd ed. (1895–96).

Playas, Pans, and Saline Flats

Playas, pans, and saline flats are the flattest of all landforms. Each owes its almost perfect flatness to periodic past and present inundation by water, and their surfaces appear similar for this reason. The processes that control their individual development vary greatly, however. They are ubiquitous features of arid and semi-arid environments, where evaporation greatly exceeds precipitation; consequently they are dry most of the time. Storm runoff or the flooding of coastal lowlands occasionally fills playas, pans, and coastal saline flats to shallow depths. Groundwater levels beneath them actively affect surface properties, especially the amount of saline minerals accumulated by evaporative concentration.

Definitions and general distinctions

A playa occupies the lowest area in a basin of interior drainage, usually consists of fine-grained sediment, and is dry most of the time. When flooded it is called a playa lake. Playa is the Spanish word for shore or beach, but in the English-speaking world it is now applied in a general sense to a wide range of topographic depressions and desiccated former lakes. Playa nomenclature is highly variable, often because of a range in both individual and regional characteristics. In North America, "salt flat," "salt marsh," "salt pan," "alkali flat," and "salina" may be used to describe a saline playa; a salt-free playa may be termed a clay pan, hardpan, dry lake bed, or an alkali flat. Playas are called lakes and pans in Australia, *takyr*s in the U.S.S.R., *sabkha*s in Arabia, and *kavīr*s in Iran. All are playas, but the terms used do not necessarily repre-

sent identical features, because geology and climate have varied with time in each locale, and the effects of these variations usually create distinctive attributes. Thus, terminology varies from place to place and within specific regions.

In South Africa, topographic depressions that periodically collect water and subsequently evaporate are called pans. Pans are generally smaller than playas, but the larger ones are equal to or greater in area than playas occurring elsewhere and that have had similar origins. Thus the terms pan and playa are used interchangeably by many authorities. Although differences do exist in playas in South Africa, they are essentially no greater than the variations that occur in other arid regions. Thousands of small depressions in the southern High Plains of Texas and New Mexico, and in Western Australia, are also termed pans by many.

Saline flats are located adjacent to coasts, shrunken inland lakes, and deltas and are usually topographically higher than the zone of tidal or lake fluctuation. They are flooded infrequently by storms or may fill to shallow depths from wave surges associated with storms. Saline crusts on the surfaces of such flats are very similar in appearance to those on saline playas. Barrier ridges often separate coastal saline flats from the open sea. These ridges have been formed by dune accumulation and, in some places, by the lowering of sea level. Saline flats extend for hundreds of miles along the coastal zones of North Africa, India and Pakistan, and the Persian Gulf.

Importance to man

Man has used playas and saline flats to his benefit for centuries. His need for salt in many areas has been satisfied only because of the existence of saline playas and coastal salt flats. Brine from several saline playas in California and Nevada contains potassium, borax, lithium, sodium, and magnesium, and these elements are extracted in commercially useful quantities. Following storms, playas and pans hold water and some ranchers in North America have constructed catchment channels and ponds for livestock watering. Persons hunt wild game that congregate at water-filled pans in Africa and Australia. The hard salt crust at Bonneville Salt Flats, Utah, has made possible construction of an automobile speed course, and several clay playas are designated as alternate aircraft landing fields. Playas have been considered for spacecraft recovery areas as well as sites for installing large phased-array antenna systems and for solar energy collectors. Numerous attempts have been made to use playas for agriculture but few successfully, usually because of excessive clay or salinity in the soil.

This article treats the physical characteristics and modes of origin of playas, pans, and saline flats and describes type examples from the western U.S., the Kalahari, the Persian Gulf area, and the Rann of Kutch in the Indian subcontinent. For further information on factors important in the development of these landforms, see WEATHERING; RIVERS AND RIVER SYSTEMS; CLIMATIC CHANGE; LAKES AND LAKE SYSTEMS; and WIND ACTION. See also EVAPORITES; and GEOCHEMICAL EQUILIBRIA AT LOW TEMPERATURES AND PRESSURES for relevant information on the kinds of surficial minerals that are formed by evaporative processes and the conditions of formation of these minerals.

PHYSICAL CHARACTERISTICS

Size, shape, and distribution. The smallest pans are measured in tens of feet, whereas the largest playas that were once permanent lakes measure tens of miles across and sometimes contain secondary depressions within a larger basin. Saline flats may be rather localized, extending across only a few hundred yards, but at their maximum development an uninterrupted length of several hundred miles, with widths of 20 miles (32 kilometres), is not uncommon.

There are probably about 50,000 playas and pans in the world, the greater majority being small deflation (wind-eroded) pans. Although large in number, they are small in area. Conversely, the largest playas are few in number and large in area. The number of playas in the world that exceed 25 square miles (65 square kilometres) in area is probably less than 1,000, and fewer than 100 are larger than 200 square miles (520 square kilometres).

The shape of playas is variable. Larger playas generally are oriented with the trend of the basin in which they occur. Smaller playas tend to be more nearly circular, but many of the smaller playas and pans have originated from deflation, and this may cause an orientation parallel to the strongest wind direction. It is highly unusual for playas to have length–width ratios that exceed 10:1; more typically a 2:1 or 3:1 ratio is observed. Playas resting on horizontal rocks tend to be more irregular in their shape, reflecting in part the regional drainage patterns. Saline flats are characteristically elongated parallel to shorelines, but this depends upon topographic and other variations.

Relation to flooding

Playas, pans, and saline flats owe their unique properties and geographic location to climate—specifically to flooding and to long periods of drought in between. They are almost exclusively situated in arid and semi-arid climates, but a small percentage are located in the dry side of adjacent climatic zones including tropical savanna, Mediterranean, and humid subtropical climates. Drained and desiccated lakes and elevated tidal flats are known to occur in polar regions, but they are not considered to belong in the same category of features. Distinction can be based on flooding ratios, wherein basins that contain water less than one-fourth of the time (and usually less than one-tenth) should be called playas. Lakes should contain water more than three-fourths of the time, and those intermediate between playas and lakes can be termed intermittent lakes. It is not unusual for average evaporation–precipitation ratios in the world's arid zones to exceed 10:1; this is the chief factor producing low flooding ratios and, in consequence, playas.

Host rock and thickness of deposits. Most playas, pans, and saline flats are floored with alluvium or unconsolidated sediment, but a few are rock-floored and some contain rock islands. Bedrock beneath playa surfaces is more likely to be sedimentary because flat-lying or gently undulating sedimentary deposits often favour playa formation. In a few areas, such as Jordan, basalt flows have served as dams and these have aided shallow-basin formation. The amount and thickness of sediments in playas and saline flats depend largely on whether the dominant formational processes and history have been depositional or erosional and on the position of the groundwater table.

Sediment thickness is not necessarily related to the age of the basin or to the rate of deposition. The 3,700-square-mile (9,600-square-kilometre) Lake Eyre (playa) in Australia, for example, contains silts and sands and evaporite deposits that are no more than 15 feet (4.5 metres) thick in places, and it has received sediment for about 80,000 years. High relief in adjacent highlands and the amount of precipitation are probably responsible for the greatest playa thicknesses. The total sedimentary fill in the basins of some former lakes in the western United States is more than 5,000 feet (1,500 metres) thick, and the Pleistocene (from about 2,500,000 to 10,000 years ago) lacustrine thickness can be as much as 1,000 feet (300 metres). Deflation pans in western Texas and southern Africa are typically shallow; many contain only a few feet of sediment, and bedrock is exposed in others. Because of the lack of a widespread pluvial (Pleistocene moist period) climate combined with a general lack of relief, the sedimentary sequences in the basins of North Africa are usually not deep. Areas of sand accumulation may be several hundred feet thick; interdune playas are often situated at the water table in these sand plains.

Surficial minerals, brines, and other materials. Playa surfaces display a range of composition and physical properties that is largely dependent on the basin origin and the hydrologic framework (precipitation, evaporation, and runoff). The depth and degree of circulation of groundwater controls the amount of capillary discharge (loss) occurring through the surface. The accumulation of evaporite minerals at the playa surface results from dissolved solids in the groundwater. The amount of clastic sediment (transported particles of sand, silt, and clay)

Importance of groundwater

is a result of prior lake sedimentation or of more recent infilling in the present environment. Clastic materials in playas are chiefly fine grain clays, silt, and sand, and these may contain carbonate that serves to bind them together.

Groundwater in arid basins follows several patterns. In northern and southern Africa, Australia, Central Asia, and Arabia it is common to have many playas or pans situated within regional basins that may cover thousands of square miles. In the United States, Iran, and Chile, smaller basins predominate, and the groundwater divides usually occur within the topographic basins. If the basin has no outlet, or if groundwater is near the surface, discharge will occur through the surface by evaporation and by transpiration from plants. Most playas and saline flats discharge some groundwater, but there are many places where there is no evidence for it; at these locations the groundwater depth is usually more than 50 feet (15 metres). These playas often contain admixtures of clay, silt, and carbonate, and there is evidence that fine-grained sediments inhibit groundwater movement.

The chemical quality of groundwater in arid basins varies greatly. Even though brines associated with saline-crusted playas are often saturated at the surface, potable water can usually be found underlying the alluvial fans and other sedimentary deposits not far from the playa. In some areas entire basins have no potable water, however.

Dissolved solid constituents have their sources in rock weathering, volcanic emanations and thermal springs, and pre-existing lake beds. It is possible to classify groundwater basins on the basis of the dominant anion (negatively charged ion)—*e.g.*, chloride, sulfate, carbonate. More attention has been given to saline mineralogy and solution geochemistry than any other aspect of playa geology, undoubtedly because of the value of agricultural, industrial, and household chemicals. Searles Lake, California (a playa), is well known for its 25 evaporite species, 10 of which were first discovered there in a natural state.

Evaporite zonation

Evaporite precipitation in a desiccating (drying) saline lake follows a solubility ordering and zonation; chlorides are the most soluble and preferentially occur in the pan or playa centre, followed by sulfates under and outside the chlorides, and carbonates deeper still and fringing the sulfates and chlorides. Other zonations that involve minor minerals also occur.

Saline flats have mineral assemblages similar to saline playas but can receive much of their brine directly from flooding by seawater during storms or unusually high tides. Saline flats receive an additional sediment source in the form of fecal pellets and guano, derived chiefly from nearshore bird populations. In a number of places this material is harvested for fertilizer.

Many authorities have attempted to classify playas according to their several characteristics; predominant sediment composition and basin hydrology have been used most often. It is desirable to know whether a basin discharges groundwater because this affects the surface properties of playas. Surfaces that appear on nondischarging playas also can appear on discharging playas, however, and this creates confusion. For this reason it is desirable to describe most playas individually and to qualify the more generalized, regionally occurring classes or types.

By courtesy of J.T. Neal



Hard evaporite crust at Sīrjān Kavīr, Iran.



Saline mud flat at Sīranda Lake, Pakistan.
By courtesy of R.E. Snead

Surficial structures and textures. Hard, dry crusts result from the drying of fine-grained sand, silt, and clay admixtures and have little microrelief except for the occurrence of mud cracks, which are widespread. These surfaces are nearly always hard enough to support automobiles but when wet they become slippery, sticky mud. Because of the compact clay structure, floodwater is sometimes unable to penetrate more than a few inches downward. Drying mud often flakes and curls and is then more susceptible to deflation—removal by wind. The characteristic colour of these crusts is very pale tan, sometimes with tinges of pink. They are usually barren, but coarser sediments in or adjacent to them permit vegetative growth. A tendency for grasses to grow occurs in the cooler fringes of the semi-arid climatic zone.

Mud cracks and desiccation polygons

Mud cracks are widespread in desiccated fine-grained playa soil and form in saturated mud. Cracks intersect essentially at right angles to form a polygonal network that varies in size from a few inches to a foot across individual polygons. In some clay-rich playas that have undergone long-term drought periods, the depth of drying and sediment contraction is great enough to permit larger mud-crack polygons to form; the term giant desiccation polygon describes those up to 300 feet (90 metres) across, a yard wide, and 50 feet (15 metres) deep in individual cracks.

Soft, friable surfaces contain some salt at the surface as a result of evaporative concentration of groundwater that rises by capillary action. These surfaces normally have groundwater within 10 or 15 feet (3 or 4.5 metres) of the surface, but the important factor is salinity rather than depth of the water. Microrelief generally ranges from two to six inches, creating a puffy appearance resembling a plowed field. These surfaces have been termed self-rising ground in the U.S. and are called *solonchaks* in the U.S.S.R. Upon flooding, these surfaces can release much of their salt by dissolution and, upon drying, can form relatively salt-free surfaces.

Salt crusts and salt karst topography

Evaporite crusts are usually associated with near-surface groundwater levels and result from rigorous evaporative concentration. These glistening white crusts range from less than an inch to many feet in thickness. Some



Wah Wah Valley playa, Utah.
By courtesy of J.T. Neal

authorities believe that the thicker crusts do not develop solely by evaporative concentration but may derive their salinity from the desiccation of pre-existing (pluvial) lakes. Where salt is thin the usual condition immediately beneath the crust is one of dark-brown sticky-wet mud. When salt is thick it may form rugged crusts, such as the "Devils Golf Course" on the playa floor near Badwater, Death Valley, California, or if flooded at regular intervals a smooth "racecourse" will form, such as the Bonneville Salt Flats, Utah. The crusts of saline coastal flats often originate from a combination of seawater flooding and of the evaporative concentration of saline groundwater adjacent to coasts. The thickness of salt is gradational in playas and saline flats; the greatest thicknesses tend to occur in the lowest portion of the depressions.

Salt crusts form expansion ridges that intersect in polygonal outlines that normally do not exceed a few feet, but giant forms have been noticed in unusually thick crusts, where the ridges are 2 feet (0.6 metre) high and 75 feet (23 metres) apart.

In the thicker salt crusts downward percolation of groundwater occurs where water tables are several feet below the surface or which fluctuate seasonally. Because of the high solubility of these crusts, dissolution occurs during percolation and a "salt karst" topography (a surface that is pitted by solution cavities) can develop on a small scale, leaving minor depressions as a result.

All of these surfaces are subject to change during flooding, and more than one, and sometimes all, may occur on a single playa. A distinct boundary between the surface types does not always exist and a considerable variety of surface types occurs. Despite these facts, they are much the same everywhere in the world.

Surface drainage occurs on playas, often in the form of small birdfoot deltas that extend inward from the periphery, and spread outward approaching the interior. Shallow channelways called grooves are common and at their maximum development are 50 miles (80 kilometres) long, for instance, at the Warburton Groove, Lake Eyre, Australia. Because of the flushing action of water the grooves are often salt free and hard most of the time and provide trafficable routes across many of the soft-surfaced playas.

Artesian (free-flowing) spring discharge on playas is common because the water table frequently slopes toward playas and because the relatively less permeable playa sediments cause the subsurface water to rise and flow near the playa margin. Both hot and cold springs occur and sediment and travertine frequently are deposited around them. Some of these spring mounds are relict features, but a number of them are still actively forming today.

MODES OF ORIGIN

Role of climate and climatic change. Climatic variation that occurred during the Quaternary Period (the last 2,500,000 years, approximately) of geologic time had great influence on the hydrologic regimen in the areas that contain playas, pans, and saline flats. Although continental glaciation during this period extended only to latitude 47° N in Europe and 38° N in North America, the effects from climatic change that caused the glaciation were noticeable far beyond the margins of maximum ice advance. Increased precipitation and cooler temperatures are generally thought to be responsible for the development of hundreds of pluvial lakes that existed in many parts of present-day arid zones. A pluvial lake is one that has had considerable, and usually periodic, fluctuation in lake level, and which may or may not contain water today. The pluvial lakes are restricted to arid region basins throughout the world, and all of them are believed to be of Quaternary age.

Pluvial lakes and the occurrence of playas today

Among the better known pluvial lakes are Lake Baikal and the combined Caspian-Aral seas in the U.S.S.R., the Dead Sea bordering Israel and Jordan, and Lakes Bonneville and Lahontan in Utah and Nevada. At its maximum extent, Lake Bonneville was nearly the size of Lake Michigan, but the modern remnant, Great Salt Lake, is only about 10 percent of its former size. South of Salt Lake City, Utah, Pleistocene glacial moraines (sedimentary deposits that formed at the margins of a glacier) that originated in the Wasatch Mountains extend beyond the old shorelines of pluvial Lake Bonneville. This suggests that some of the glacial intervals were synchronous with pluvial periods. Interpluvial periods existed with interglacial periods, and the climatic conditions of the intervals may have been similar to the modern climate.

Many playas and some saline flats that are adjacent to present-day lakes were the sites of these pluvial lakes, and the sedimentary record beneath several of them has been studied and dated by radiocarbon methods (see DATING, RELATIVE AND ABSOLUTE). Searles Lake, California, is among 120 pluvial lakes in the western United States, and it has been intensively cored and studied. Alternating layers of mud and salt correspond to pluvial and interpluvial periods and Recent (the last 10,000 years, approximately) time. These units record the former presence of two deep lakes followed by at least near-desiccation and evaporite precipitation. Radiocarbon dates on samples from these layers show that the more recent of the two pluvial periods lasted from before 23,000 to about 10,000 years BP (before present), and it was therefore contemporary with the Wisconsin stage of glaciation in the central and eastern U.S. The older pluvial sequence was well es-

tablished by about 46,000 BP and started to wane before about 32,000 BP.

A somewhat similar sequence of salt and mud layering has been discovered at the Masileh Kavīr, Iran, suggesting an alternating pluvial and interpluvial climate, but the layers have not been dated. It is possible that this occurrence represents more local conditions adjacent to the Elburz and Zagros mountains because there is little evidence for a marked pluvial period in most of Iran. The numerous playas in pluvial lake basins in western North America at one time prompted speculation on similar correlations in other regions. Studies of pluvial lakes and past climates in other regions have failed to show a general scheme of climatic fluctuation that is acceptable in more than one region.

Meteorological variations in North America

The worldwide occurrence of pluvial lakes did not depend entirely upon runoff or on glacial meltwater but rather on the combined effects of local and secular variation of temperature, precipitation, wind, and the seasonal distribution of each, which together regulate the hydrologic balance. Most of the world's pluvial lakes were situated beyond the drainage of glaciers and were thus hydrologically unrelated, although exceptions did exist. Precipitation and temperature variation during the pluvial periods change considerably from one region to another. Studies of pluvial lake environments in western North America have suggested precipitation increases of about seven to nine inches (18 to 24 centimetres) and temperature reductions of 5° to 9° F (3° to 5° C), relative to modern values. The last true pluvial period occurred during Wisconsin time, reaching maximum development about 20,000 years BP. Minor climatic fluctuations have occurred since the last major phase and will continue to occur in the future. About 5,000 years BP a period of slightly higher temperature and increased wind activity is thought to have occurred. Many deflation pans in western Texas probably formed at this time.

During the last hundred years glacier advances have been observed in the western U.S., concurrent with a rise in levels of some enclosed lakes in Oregon and Washington. Climates since 1850 are thought by many to have become warmer, with less precipitation and greater evaporation. The drying up of several permanent lakes during the past 100 years may be related to arid conditions that followed the "little ice age," a widespread period of renewed glacial activity that lasted from about 1650–1850. These short-term fluctuations are not true pluvial episodes, but nonetheless they have important local effects on basin hydrology and geologic processes. Playas and pans are sensitive to meteorological variations of even small magnitude, and the longer term events associated with and following the Pleistocene glaciations have profoundly affected their emplacement.

Role of tectonism. Tectonism, or movements of the earth's crust, affects proportionately few playas as a dominant playa-forming process, but those that are affected include the larger ones; thus the total area involved is great. There are noticeable effects of tectonism that probably influence all playas to some extent. The very existence of many regional desert basins depends on mountain barriers, which essentially reflect the global tectonic framework. Individual basins in Nevada, Iran, Chile, and elsewhere are down-dropped, fault-block basins or grabens, and in some places they are associated with volcanism. It is the relief of the adjacent mountains combined with precipitation that provides the sedimentary framework for the thick lacustrine (lake bed) sequences in many of these basins. Large regional basins in North Africa, Central Australia, and the Dasht-e Kavīr of Iran have resulted from regional downwarping and faulting and folding. Coastal warping and uplift undoubtedly have been responsible for the initial formation of some coastal saline flats.

Association of faults, earthquakes, and springs

Faults transect numerous structurally controlled playas in the western U.S., Iran, Chile, and Australia and are represented by low scarps (ledges) and hydrologic effects in the form of springs and wet zones marked by vegetation. Earthquakes and active faulting occur in Nevada, and hot springs deposit travertine at several locations.

Pleistocene Lake Bonneville reveals tilted shorelines today, showing that the weight of water in the largest pluvial lakes was capable of depressing the crust of the earth. Isostatic rebound, the recovery of the crust from this load following lake desiccation, is still going on today, and it is estimated that it is 75 percent complete after some 12,000 years. The crustal depression of Lake Bonneville was more than 200 feet (61 metres) in the basin centre, where the water was about 1,100 feet (335 metres) deep. This process demonstrates that large basins can effectively perpetuate themselves by further deepening their basin once they have achieved a certain size.

Role of wind action. Wind has an important role in at least four aspects of playa and saline-flat development: (1) the initial depression-forming processes; (2) modification and orientation of the shapes of basins; (3) removal of sediments by deflation; and (4) depositional features leeward of deflated playas and saline flats.

Wind is a universal agent that seems to have affected all playas to some degree. It may be the principal agent of small basin formation in West Texas and the Kalahari. There are more than 20,000 of these basins—large in number, but small in total area compared with larger playa basins. Many of the larger basins in the Sahara, South Africa, and Australia are thought to have had considerable material removed by deflation, perhaps at an early stage in their development.

A principal role of wind that occurs during flooding on playas and saline flats lies in assisting water as a levelling agent. It agitates the water and increases its capacity to hold and move suspended sediment. It moves water in sheets and piles it up in local concentrations, regardless of level. Dissolved salts can be removed from playas by this process; Harper Playa, California, was effectively washed clean in this manner. Prior to flooding its surface was soft and friable and contained substantial amounts of salt, but upon drying a hard, dry crust appeared.

Dispute regarding the importance of deflation

There is lack of agreement in many regions on the relative importance of deflation on playas. It is known that the amount of wind erosion ranges widely over the world and that it is controlled by aridity and the intensity of local winds. Danby Playa, California, has been described as an example of a playa that has had approximately ten feet (three metres) of sediment removed by deflation. Estancia Valley, New Mexico, contains numerous deflation pans that experienced wind erosion in the past. In the Sahara the evidence is often more direct and widespread.

Once barrier ridges are formed and tidal salt flats are isolated from the sea, wind becomes an important agent of change. As material is deposited by sedimentation and evaporative concentration of groundwater, wind serves to carry it away, thus destroying marine constructed features.

Sand dunes in the Kara-Kum and Kyzyl-Kum deserts of Turkistan are interspersed with depressions (playas) that are frequently saline and that are similar to widespread areas of interdune playas in many other areas, including the Algerian Sahara, the Dasht-e Lut desert of Iran, and western Rhodesia. They usually have a linear orientation and are small, but a few are a mile long and a half mile wide. Wind action is essential in the formation of these playas, but a near-surface water table is present in many of them, serving as a base level (lower limit) of deflation.

Many playas in Australia and the United States exhibit transverse crescentic foredunes called lunettes on their leeward side and may contain substantial amounts of silt and clay. Similar features are forming today along the saline flats of the Gulf Coast of Texas and have been termed clay dunes. Their widespread occurrence in association with playas and saline flats, and yet general inactivity at present, has caused several researchers to believe that many are relict features, possibly representative of a drier and windier period during earlier Quaternary time.

Some elongated lakes can be oriented perpendicular to the winds that produce them. In the development process, winds blowing across shallow circular lakes produce 90° longshore currents that have maximum erosive power at the lake ends.

At many playas vegetation serves as a windbreak and sand accumulates on the lee of the plants. Often the

vegetation will continue growing, thereby perpetuating the sand-accumulation process and forming conical mounds up to 20 feet (6 metres) high. The roots of the plants frequently extend downward to phreatic (subsurface) water; thus the term phreatophyte mounds has been coined to describe these features.

Summary of occurrences and origins. Playas, pans, and saline flats have similar surface characteristics; they are extremely flat because of flooding and usually consist of fine-grained sediment, or evaporite minerals. The resemblance is superficial because great variation exists in basin origin, sediment thickness, and hydrologic balance, each lending individual or regional character to them.

Larger playas usually were associated with pluvial periods during Pleistocene time; the runoff often depended on nearby mountain precipitation sources, and it was responsible for increased basin sedimentation. Structural warping and faulting created the basins for many of these lakes. Ancient river systems have been responsible for the initial location of numerous playas and pan systems. Past climatic fluctuation has affected nearly all playas and pans to some degree. Most important were the pluvial and interpluvial periods of the Wisconsin stage, and subsequent postglacial events. Deflation occurred on many pans and small playas at this time. A combination of processes, including solutioning and animal wallowing, is locally important. The amount of capillary discharge and the depth and salinity of groundwater are especially significant in controlling the evaporite buildup in all playas. Coarser grained playa sediments tend to increase capillary discharge.

Coastal saline flats have originated as depositional features that have been cut off from the sea by sand-barrier ridges, by reef building, or by coastal uplift. The slight elevation of some coastal flats, such as the Sirte Gulf of Libya, is believed to have resulted from recent sea-level oscillations, where at high-water stands these features were shallow bays or lagoons. Although similar to playas in many respects, saline flats are much less stable features of the landscape because of their precarious location.

Current uncertainties and gaps in knowledge

Several aspects of playas, pans, and saline flats remain poorly understood. Very few of these features have been explored beneath the upper few feet. Knowledge of the depositional sequence and chronology contained in the sedimentary record would provide detailed knowledge of past and present climates and ultimately help to predict future trends. Age dating of sediments and analysis of evaporites is required to correlate absolute age with the depositional environment and to reconstruct late Pleistocene history and the worldwide variation in climatic change. More regional information on playa basins, surface conditions, and hydrologic factors would help in understanding changes through time and their cause.

TYPE EXAMPLES

Playas of the western U.S. Arid-land specialists have remarked that the U.S. arid zone is atypical in comparison to other deserts of the world. A chief reason for this difference is that mountain precipitation sources are so close to the arid basins that runoff is greater, lessening the geologic impact of aridity. Consequently, western North America had the greatest concentration of pluvial lakes during Pleistocene time. It has much less sand than most deserts and less evidence of wind activity. During the Pleistocene, south of about latitude 34° N, the amplitude of climatic change showed a marked decrease because of increased distance from the belt of westerly storm tracks. This fact, in combination with higher temperatures at more southerly latitudes, resulted in reduced pluvial lake fluctuation. For this reason the Mojave Desert shows less evidence of pluvial lake development than Nevada, but at the same time more sand and consequent wind activity. Searles Lake derived its pluvial lake water from overflow of the Owens River system, which originates in the high Sierra Nevada. There are some 300 playas exceeding five square miles (13 square kilometres) in the Basin and Range Province, which covers western Utah, Nevada, southern Oregon, Idaho, and Wyoming and extends into California, Arizona, and New Mexico. Because the majority of these playas were formerly pluvial lakes, the thickness of sediments is great. The pluvial history of Searles Lake shows that more than 800 feet (240 metres) of sediment were deposited over a period of some 100,000 years.

More than half of the U.S. playas contain hard, dry crusts, which is a higher ratio than that occurring elsewhere. This is partly the result of thick lacustrine sediment sequences that help retard capillary discharge, because of interbasin circulation of groundwater in some places, and because of periodic flooding that helps wash away saline accumulations. Man's intensive use of groundwater has affected playa surfaces in many basins; lowered groundwater levels and reduced capillary discharge have been one result. Subsidence of basins and the occurrence of giant desiccation polygons are other effects that have been observed.

The thousands of deflation pans on the southern High Plains are different in origin than Basin and Range playas and are more similar to the pans of the Kalahari. A number of workers believe that U.S. playas are essentially in equilibrium, with deposition and deflation being nearly equal.

By courtesy of the Air Force Cambridge Research Laboratories

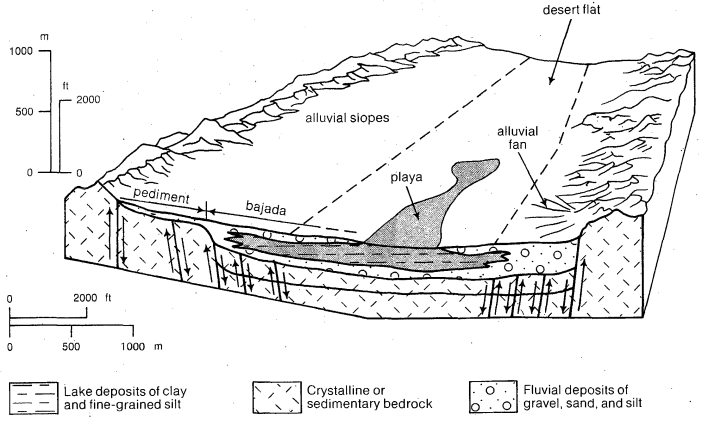


Figure 1: A typical playa in the Basin and Range region showing the nature of the deposits and related features.

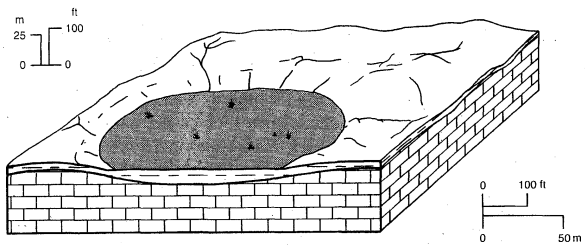


Figure 2: Deflation pan, representative of the Kalahari, southern Africa, and of western Texas.

Kalahari pans. The Kalahari region of southern Africa is a gently undulating plain, elevated 3,000 to 3,600 feet (900 to 1,100 metres) above sea level. Much of it is covered with sand and has a vegetative cover of trees, grasses, and scrub. Sand dunes are inactive, and movement occurs only where vegetation has been destroyed.

There are some 10,000 pans over 100 yards (90 metres) across in the Kalahari south of the Okavango and Zambezi rivers. The flat-floored depressions vary considerably in their extent and may lie 50 feet (15 metres) or more below the surrounding plain. They are not peculiar to the Kalahari but more are located there than in the rest of South Africa. There are several varieties of pans: (1) large pans like Haakschein Vley that are 80 square miles (210 square kilometres) and more; (2) sand pans a mile (1.6 kilometres) or more across, but often smaller; and (3) smaller limestone pans. The first group contains shallow sandy mud situated on sandstone or shale bedrock, whereas the sand pans in sandy terrain may overlie several hundred feet of sand. Limestone pans are apparently

Varieties of pans and relation to former drainage systems

peculiar to the Kalahari and are smaller than the others; sandy limestone is usually encountered within a few feet of the surface. A considerable amount of variation exists in all of these pans; many are saline, or grass covered where salinity is low.

Wind erosion is thought to have played a prominent role in the formation of the larger pans, and sand and interdune pans, but it is also essential that an initial topographic depression existed. It is believed that the majority of these Kalahari pans are aligned along sand-covered ancient drainage systems. Limestone pans are generally smaller and may originate in animal wallows where groundwater fluctuates near the surface during alternating wet and dry seasons. A combination of processes seems to be involved in the formation of most of the Kalahari pans.

The Makgadikgadi Depression in Botswana is not a single basin but rather a series of salt pans within a larger basin that measures some 110 by 60 miles (117 by 96 kilometres). The larger region of internal drainage covers about 200,000 square miles (517,998 square kilometres) and includes the basin of the Okavango River. Numerous fossil river valleys, termed *mokgacha* locally, lead into the depression. The margins of the basin pass gradually from grassland and isolated pans to panland with grassy islands, and finally to unbroken pan in the interior. The main floor is not completely level, and the interior pans occupy the lowest part of an area of internal drainage. Artesian springs are seen in the depression and probably originate in higher ground to the north. The depression is structural in origin and is deflated to some extent by easterly winds during the dry season. Little deposition is occurring at present because the Okavango River deposits its sediment in the marshes of Ngamiland.

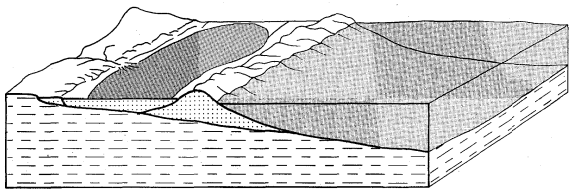


Figure 3: A coastal saline flat separated from the open ocean by a barrier ridge of dune sand.

Saline flats of the Persian Gulf and the Rann of Kutch. The saline flats that border most of the Persian Gulf coast of Abu Dhabi, a sheikhdom of the Trucial States, are known locally as *sabkhas*. These flats extend a distance of almost 200 miles (322 kilometres) and are nearly 20 miles (32 kilometres) across at their widest part. The flats stand just above the normal high-tide level and are everywhere bordered on their landward side by low bluffs of Tertiary (2,500,000 to 65,000,000 years ago) and Quaternary rock exposures. On the seaward side the *sabkhas* face broad intertidal flats, and in some places lagoons. Outside the outer beach ridge, the upper parts of the intertidal flats resemble the *sabkha* surface. Storms may create new beach ridges and cut these *sabkha* flats off from the sea. Low linear ridges cut much of the *sabkha* plain and provide vehicle routes across an otherwise treacherous surface. They are undoubtedly old beach ridges and often consist largely of gastropod shells, similar to many of the present outer beach ridges. Some parts of the outer coastal ridge system are capped by low dunes; when these are breached during storms with concurrent high tides, much of the *sabkha* plain can be covered with water.

The barren *sabkha* surface is often encrusted with salt up to several inches thick, and gypsum crystals and anhydrite nodules can be seen projecting from the surface, indicating that deflation is operative. In places Quaternary limestone is exposed just beneath the surface, but elsewhere the sediment is much thicker. The Tertiary and Quaternary limestone inland provides the source of the unconsolidated sediment, which is predominantly carbonate, with minor amounts of quartz.

Some authorities believe that the *sabkha* plain is very young and, because it shows no major breaks in level across its width, that it has not experienced a change in sea level during its development. Its outward growth is the result of extensive and prolonged intertidal flat sedimentation, together with chemical deposition and the growth of calcareous organisms. The Rann of Kutch in northwestern coastal India (Gujarāt state) is an area of extensive tidal mud flats and salt flats that are periodically inundated by seawater. It extends inland from the coast for a distance of about 250 miles (400 kilometres), with an average width of about 50 miles (80 kilometres). Much of the "Great Rann" and the extensive lowland at the eastern end of the Gulf of Kutch known as the "Little Rann" contain local saline efflorescences and salt crusts that range in elevation from sea level to 10 or 12 feet (3 or 4 metres) above sea level. Together, the Great Rann and the Little Rann comprise a saline flat desert 9,000 square miles (23,300 square kilometres) in area. The Rann of Kutch is a remnant of an arm of the sea that previously connected the Narmada Rift with Sind and separated Kutch from the mainland.

The Rann is mostly covered with water and is virtually impassable to any type of vehicular traffic during the southwest monsoon, but when dry much of it will allow some traffic by drivers who are aware of its dangers. These routes must generally follow the higher elevations, as the lowest parts of the Rann remain flooded throughout the year. Saline areas are usually avoided, but when a salt crust is thick or dry enough it may be trafficable. Everywhere groundwater is within a few feet of the surface and it is saline. Vegetation is practically absent, except in a few areas that are slightly raised and less saline, or which are naturally favourable for mangrove growth or salt-tolerant plants.

BIBLIOGRAPHY. G.K. GILBERT, "Lake Bonneville," *Monogr. U.S. Geol. Surv.*, vol. 1 (1890), a classic work on one of the major pluvial lakes of the western U.S.; C.C. REEVES, JR., *Introduction to Paleolimnology* (1968), a discussion of playa characteristics and origin in western Texas and the High Plains of the U.S.; J.T. NEAL, "Playa Variation," in W.G. MCGINNIES and B.J. GOLDMAN (eds.), *Arid Lands in Perspective*, ch. 2, p. 13–44 (1969), a survey of playa characteristics and theories of origin, together with a bibliography that serves as a guide to the world literature; C. BOOCOCK and O.J. VAN STRATEN, "Notes on the Geology and Hydrogeology of the Central Kalahari Region, Bechuanaland Protectorate," *Trans. Geol. Soc. S. Afr.*, 65:125–171 (1962), a basic paper on the Kalahari pans.

(J.T.Ne.)

Pleistocene Epoch

The sixth of the seven epochs that constitute the Cenozoic Era of geologic history (the last 65,000,000 years), the Pleistocene immediately antedates the most recent or Holocene Epoch (*q.v.*). The time boundaries are not fixed precisely, but the Pleistocene is believed to have begun about 2,500,000 to 3,000,000 years ago and to have ended about 10,000 years before the present time. Although originally defined in 1839 as the time when certain marine strata containing fossil mollusks were deposited, the Pleistocene Epoch also was a time during which extensive ice sheets and other glaciers formed repeatedly on the lands. For that reason it was long informally called the Ice Age. It is now evident, however, that large glaciers formed on the Antarctic Continent, Arctic North America, and possibly other areas as early as 10,000,000 to 13,000,000 years ago, in the later part of Miocene time. Following an unknown number of fluctuations of these "early" glaciers, ice sheets developed over vast areas of northern and middle-latitude North America and northwestern Eurasia. This invasion of middle latitudes seems to have occurred within the Pleistocene Epoch proper. It was followed by the melting away of the ice sheets and their later renewal, an oscillating process that occurred several times. These events, evidently responses to fluctuation of the earth's climate, are widely known as glacial ages and interglacial ages. The Pleistocene Epoch is frequently referred to as the Glacial Age, and the usage is reasonably accurate when large ice sheets extending into middle latitudes are implied. Earlier gla-

Glacial and interglacial ages

cial ages, however, resulted from earlier climatic fluctuation and were characterized by the presence of numerous glaciers in higher latitudes.

Although the spreading of glaciers, including large ice sheets, was a conspicuous result of climatic fluctuation during the Pleistocene, it was accompanied by other results of great importance. Among these were notable fluctuations of the level of the world ocean, the creation of temporary lakes in regions that today are dry, and changes in the composition and areal ranges of plants and animals, both on the lands and in the ocean. In addition, the earth's crust subsided to considerable depths beneath the loads created by the spreading ice sheets. In high latitudes the ground froze deeply and remained frozen through entire glacial ages. The great natural process of erosion of rocks and deposition of sediment was temporarily altered in a number of ways. All these changes were directly or indirectly the result of changes in the world's climate during the Pleistocene. This article describes these direct and indirect effects of Pleistocene glaciation and climatic change. For additional detail on the earth processes and features involved, see CLIMATIC CHANGE; ICE SHEETS AND GLACIERS; GLACIATION, LANDFORMS PRODUCED BY; PERMAFROST; LAKES AND LAKE SYSTEMS; RIVERS AND RIVER SYSTEMS; COASTAL FEATURES; WIND ACTION; SOILS. See also TERTIARY PERIOD and HOLOCENE EPOCH for allied information on the previous and subsequent intervals of geological time, and CENOZOIC ERA for a general overview of the last 65,000,000 years of earth history.

PLEISTOCENE CHRONOLOGY

The framework of Pleistocene events has been determined primarily from sequences of strata, deposited in many different environments, that contain assemblages of fossils and bear absolute dates measured by a variety of methods (see DATING, RELATIVE AND ABSOLUTE). The study and dating of Pleistocene strata is in progress in all countries but is farthest advanced in parts of North America and Europe. An example of the accumulated information is the generalized succession of Pleistocene tills (direct glacial deposits) recognized in the Sierra Nevada, California (Table 1), each till representing a glacial episode. Some of the tills are separated from adjacent ones by ancient soils, and others by volcanic sediments. They have been dated by various means and the dates given represent varying degrees of uncertainty. Each indicates an episode, or a longer period, when glaciers in the Sierra Nevada region were much larger and more extensive than they are today.

Table 1: Chronology of Glacial Deposits in the Sierra Nevada, California

glacial till	approximate date in years BP* (uncertainty increases downward)
Hilgard Till	10,000
Tioga Till	
Tenaya Till	20,000
Tahoe Till	50,000
Mono Basin Till	80,000
Donner Lake Till	200,000 to 500,000
Sherwin Till	750,000
McGee Till	1,500,000
Deadman Pass Till	3,000,000

*Before present.

North American sequences

In central North America, from the Appalachians through the Great Lakes region to the Great Plains, extensive and detailed sequences of strata have been studied and described. These strata represent a succession of major glaciations by an ice sheet; the glaciations have been identified, and each is named after the state in which it was first clearly recognized (Table 2).

The Wisconsin major glaciation occurred in at least three pulses or stades, separated by interstades during which the ice sheet shrank. Each of the three earlier major glaciations was likewise multiple and so consisted of stades, but less is known of their number and their dates; hence, their local names are not included in the Table. The sediments representing each major glaciation are separated from the layers overlying them by ancient soils or by sediments of nonglacial character, some of them containing fossil organisms that imply temperate rather than glacial climate. The intervals corresponding to such features are interglaciations or, simply, interglacials.

The late Wisconsin Stade in central North America began some 25,000 years ago, reached its maximum some 15,000 to 20,000 years ago, and ended around the end of Pleistocene time, although remnants of the ice sheet persisted in Canada until after 7000 BP (before present, where present means AD 1950). During this glaciation most of the deposits and other features were formed that led 19th-century geologists to recognize the extent of glaciation in North America. The middle Wisconsin Stade can be placed near 50,000 to 40,000 BP in round numbers, and the early Wisconsin Stade around 70,000 to 60,000 BP, although with more uncertainty. Between the times of these glacial events, climates were cooler than those of today, as indicated by the fossil pollen of plants that constituted the vegetation.

In the central region there has been no direct measurement of the age of any of the older glacial sediments, so that time comparison of the sequence as a whole with that in the Sierra Nevada is not yet possible. In most places Nebraskan sediments unconformably overlie much older rocks, no strata from the preceding epoch, the Pliocene, being present, so that direct evidence of the position of the Nebraskan relative to the Pliocene is lacking. On indirect evidence, however, the Nebraskan Glaciation is supposed to belong within the Pleistocene Epoch, and on very indirect reasoning a tentative age of around 700,000 years has been suggested for it.

The Pleistocene units in central North America are grouped, on the basis of their content of fossil mammals, into the three mammal ages, shown in Table 2. These

Table 2: Chronology of Glacial and Interglacial Episodes in Central North America

approximate date (years BP)	glaciations		interglaciations	mammal age
10,000				
20,000		late		
30,000				
40,000	Wisconsin	middle		
50,000				Rancholabrean
60,000		early		
70,000				
			Sangamon	
	Illinoian			
			Yarmouth	Irvingtonian
	Kansan			
			Aftonian	
700,000?	Nebraskan			Blancan

same ages have been identified in Pacific North America, although not in the Sierra Nevada, because the tills are not in contact with strata that contain fossil mammals. The Blancan Mammal Age reaches much farther back in time than the Nebraskan Glacial Age is thought to do. The latter age may have begun less than 1,000,000 years ago. Potassium-argon dates suggest that the Blancan Mammal Age may have begun perhaps as long aso as 4,000,000 years BP.

European sequences

Sequences representative of those developed in Europe are set forth in Table 3, in which interglaciations and temperate-climate layers are italicized. A somewhat similar sequence of glacial deposits is recognized in the Alps region in Europe but is insufficiently well known and insufficiently controlled by dates to be represented in a comparable table. These sequences are more complicated than those in North America. Most of the units designated as cold-climate units include glacial sediments and therefore clearly indicate glaciation. Some, however, record merely low temperature, through the kinds of fossils (mostly of plants) they contain. The temperate-climate units are identified principally through their fossils.

In the two western European columns in Table 3, the lowest units directly overlie, in places, strata of Pliocene age, and so are clearly of lower Pleistocene age. Based on

fossil evidence, chiefly fossil mammals, the units are grouped into lower, middle, and upper Pleistocene categories. Indeed, fossil mammals are a chief means of correlating the temperate and even the glacial units between one part of Europe and another.

Table 3: Representative Pleistocene Sequences in Europe*

	Britain	Netherlands, Northern Germany, Denmark	European U.S.S.R.
Holocene			
Pleistocene			
Upper	Weichsel (Devensian)	Weichsel	Valdai
	Ipswich	*Eem*†	*Mikulino*
	Gipping (Wolstonian)	Saale	Moscow
			Odintsovo
			Dnepr
	Hoxne	*Holstein*†	*Likhvin*
	Lowestoft		
Middle	Anglian Corton	Elster	(possibly glacial sediments)
	Gunton		
	Cromerian	"*Cromer*"	
	Beestonian		
	Pastonian		
	Baventian†	Menap	
	Antian†	*Waal*	
	Thurnian†	Eburon	
Lower	*Ludhamian*†	*Tegelen*	
	Waltonian†	Brüggen	
Pliocene	*Coralline Crag*†	Reuver Clay	

*Correlations, particularly in the lower part of the sequence, are uncertain and provisional. †Marine units.

The dates of the youngest major cold-climate unit, the Weichsel/Valdai, fall within the last 70,000 years and so are comparable with the Wisconsin unit in North America. Like that unit, the Weichsel/Valdai includes repeated periods of low temperature, although the details of its earlier part are still obscure. The time relationships are inadequately fixed by absolute dates, although the relationship of the base of the sequence to the Pliocene strata is more evident than it is in central North America.

Although lower Pleistocene and Villafranchian are regarded by some authors as synonymous, the latter is properly a mammal age, like the Blancan in North America. Radiometric dates, however, suggest that the Villafranchian does not reach as far back in time as the Blancan does. It may have begun only some 3,000,000 years ago.

PLEISTOCENE CLIMATES AND CLIMATIC CHANGE

Snow lines and lake levels. Although the former presence of great glaciers in middle latitudes is perhaps the most obvious evidence of cold Pleistocene climates, less conspicuous indications likewise show convincingly that cold climates prevailed. One of the most widespread concerns the snow line, the elevation above which snow and ice remain on the ground throughout the year. Plots of the altitude of the snow line during glacial ages, measured in various ways, reveal that it was everywhere lower than is the snow line of today. The plots indicate further that the two snow-line positions are broadly parallel, although the vertical distance between them is greater in middle than in very high latitudes, and is greater in moist rather than dry regions. In moist middle latitudes the vertical interval between former and modern snow lines amounts to between 1,000 and 1,500 metres (3,300–4,900 feet).

Temperature differences

Although there exists no precise way of determining the related temperature difference, the opinion that the glacial-age summer temperatures were 5° to 7° C, or more, below those of today is rather widely held. If today's temperatures in southern Norway were lowered by only half that amount over a long period, it is believed that the resulting growth of glaciers would again cover that country with an ice sheet.

Another indication of the colder climates of glacial ages is provided by the existence of former lakes. Saline lakes and saline flats that today occupy basins in the world's dry regions are surrounded by abandoned beaches and other shoreline features. Such features demonstrate that larger lakes formerly occupied the basins, some of them to great depth, and that a different climate must have prevailed when the lakes existed.

Oxygen isotope ratios. Further indication of Pleistocene climatic fluctuation has been derived from a core of glacier ice, bored from the Greenland Ice Sheet in 1969. The water vapour in the atmosphere contains two isotopes of oxygen, ^{16}O and ^{18}O. The ratio between these two isotopes depends on the temperature at which atmospheric moisture condensed to form falling snow. In the region where the snow fell, changes in this ratio, from layer to layer in fallen snow preserved through many years, should therefore reflect changes in atmospheric temperature through those years. The isotope ratios from the core of the whole thickness of the northern part of the Greenland Ice Sheet yield a curve with variations that are similar to those constructed from other climatic data from North America and Europe. In 1970 analogous measurements in a core from the Antarctic Ice Sheet produced compatible results. Although these measurements demonstrate fluctuation of climate in very high latitudes, they do not reveal the absolute amount of fluctuation at any given place.

Measurement of the $^{18}O/^{16}O$ isotope ratio in sediment cores obtained from beneath the ocean floor in low latitudes gave results from which curves were plotted as early as 1955. The substance measured was carbonate, from the shells of planktonic Foraminifera that accumulated year by year on the ocean bottom. The ratio depends on both temperature and the isotopic composition of the seawater from which those organisms secrete carbonate. The curve shows changes between warm and cold times, and in that respect it resembles broadly the curves developed from the ice cores. It does not, however, discriminate between changes in temperature of seawater and changes in isotopic composition of the water. Isotopic measurements of sediment cores, therefore, do not give independent temperature data.

Faunal and floral evidence. Another indication of Pleistocene climatic change, likewise derived from analysis of cores from beneath the ocean floor, consists of changes in the proportions of constituents of fossil microfaunas between one layer of sediment and another. Many groups of planktonic species can be labelled warm, transitional, or cold according to their areal distribution in the oceans today. When cold faunas systematically occur as fossils in a particular layer, in cores from lower latitudes in which only warm elements live today, it can be inferred that when the fossil fauna was living the temperature was lower than it is now. Applied with proper techniques, this principle has yielded results that agree broadly with temperature data derived from other sources.

Microfauna as index of climatic change

A different kind of temperature indication is obtained from counting and estimating the absolute numbers of individual fossil planktonic organisms in particular layers within a core. In many parts of the ocean the numbers of such organisms decrease with decreasing temperature, and a decrease therefore indicates a colder climate. In certain areas, however, numbers increase with increased nutrients, brought to the surface in greater amount by more rapidly upwelling cold water during glacial ages. A curve developed from a series of such measurements on a seven-metre core from the eastern equatorial Pacific shows pronounced fluctuations, which are of greater amplitude in the upper than in the lower part of the section. The fluctuations probably were climatically caused, but because they are not dated radiometrically they cannot yet be compared closely with climatic changes inferred from features on land. When such comparison can be made, curves from long sea-floor cores will be especially useful, because in many of them the sediments were deposited continuously. Continuous sedimentation represented in cores taken from land areas is less likely because the processes of erosion on land are everpresent.

The Pleistocene sediments of the land as well as those of the ocean contain fossils from which important climatic data are extracted. The fossils include both plants and animals, but the data most suitable for the interpretation of former climates are derived from fossil pollen that

Analysis of fossil pollen

accumulated in lakes and bogs. The kinds of vegetation that flourished in the vicinities of such places, compared with the distribution of similar kinds that are living today, supply sensitive information about former climates. It has been found that in glacial ages, in Eurasia and North America at least, distinctive groups of middle-latitude plants lived somewhat nearer the Equator than their counterparts do today, and in mountainous terrain they lived at lower altitudes. The inference that temperatures were then generally reduced is obvious. Where the inference can be quantified, it appears that the amplitude of temperature reduction is roughly comparable with that calculated independently from snow-line altitudes—specifically, a few degrees Celsius.

In the regions overrun by former ice sheets, many of the pollen data come from exposures of interstadial and interglacial sediments sandwiched between layers of till or other sediments of glacial character. Beyond the limits of the former ice sheets, data have been obtained from long cores taken from thick sediments deposited in lakes and bogs. Such sediments, unlike those in the glaciated regions, may represent deposition that was continuous throughout the latest major glaciation, and perhaps through an even longer time. The longest cores of this kind have been taken from lake sediments in southern Spain and in Macedonian Greece and are dated radiometrically, back through about the last 50,000 years.

The most basic result of the comparison of curves derived from pollen data is that most of the curves are similar. This similarity strengthens the inference that changes of climate have occurred in similar succession throughout wide regions—at least in the Northern Hemisphere, where most of the existing data have been derived.

Another important source of climatic inference consists of fossil land mammals, collected systematically since the 18th century. In the 19th century it was realized that the nature of some of them implied climatic differences. Pleistocene fossil mammals furnish a routine basis for climatic inferences today, however, and much has been learned from them. Land mammals are a less sensitive indicator of environment than plants because they are able to withstand a wider variety of environmental influences. In consequence the value of fossil mammals largely lies in confirming the environmental data provided by the more sensitive fossil plants, especially in the form of pollen. Such confirmation is abundant, and few conflicts in testimony from the two sources have appeared.

Fossil invertebrate land animals also have played a part in the reconstruction of Pleistocene environments. Nonmarine mollusks, for example, have proved important in the discrimination between forest and steppe in central and eastern Europe, a distinction closely related to the change between interglacial and glacial climate. Insects, too, are useful, partly because they are very sensitive to temperature and respond promptly to a change of climate by migration.

Relict frost-climate features. Among other kinds of widespread evidence from which Pleistocene climates have been inferred are frost-action effects. In many areas such features are comparatively obscure; nevertheless close study of them has produced informative results.

Frost-wedged debris and cave rubble

Debris created by frost wedging of rocks is one widespread effect. The debris forms taluses (sloping masses of unconsolidated fragments), block fields, and other patchy mantles on the land surface, especially in moist climates in which freeze-and-thaw occurs frequently. Although forming today at high latitudes and altitudes, frost-wedged debris occurs also in an inactive state in lower positions. Its relict character is attested by direct measurement as well as by its state of weathering and by the growth of lichens on it. In Europe relict taluses occur as far south as Gibraltar. In the western United States most taluses in the south are inactive, but toward the north taluses increase in size and abundance. It is thought that former colder, moister, glacial-age climates, reflected in the pluvial lakes of that region, promoted frost wedging that today is decadent.

Relict debris formed in this way, but now found interlayered in the floor sediments of caverns, is particularly significant because in some caverns it is associated with artifacts, bones of fossil mammals including man, and radiometrically dated material. Two layers of rubble derived from the cave roof at different times are included in floor sediments of a cavern in coastal Libya. The layers correspond in radiometric age with two glacial stades in northern Europe. Therefore, they support evidence from fossils that European glaciations were marked, in some places at least, by frost climates in a belt of country that extended far beyond the glaciers themselves.

Other features indicative of former frost climate are relict mantles of solifluction (soil-flowage) sediment. Common in moist Arctic areas today, such sediment is activated by spring thawing and moves downslope over frozen ground as thin tongues and sheets. Wide areas of relict solifluction sediment are common in such regions as western Europe, the central Appalachians, and southeastern Australia and testify to glacial-age frost climates.

Still another kind of frost-action effect that occurs in relict form in middle latitudes consists of ice-wedge casts, which today are observed in higher latitudes in all stages of active development. Sudden winter cold snaps cause permafrost (permanently frozen ground) to crack at the surface. Frost accumulating in the crack, over the years, fills the crack with a wedgelike body of ice and continually widens it. When the climate warms sufficiently, the wedge melts and the vacated space gradually fills with washed-in or blown-in sediment. Patterned ground and frost-stirred ground are additional frost-climate features that occur commonly as relicts. Networks of patterned ground occur at high altitudes on ridge tops in the middle Appalachians but are inactive today.

Some of the frost-action effects mentioned above are related to permafrost; others are not. Permafrost is widespread in Arctic regions today and during glacial ages certainly extended to lower latitudes. Its former extent, however, is difficult to fix. In general, permafrost occupied belts of country bordering the major ice sheets. Indications of its former presence seem to be most abundant in continental central Europe, where winter temperatures were much lower than in the maritime climates of western Europe. Apparently the permafrost belt across the eastern half of North America was narrow and discontinuous, possibly because warm air from the Gulf of Mexico was carried northward with little hindrance, thus maintaining a less rigorous climate than that of central Europe.

Soils and eolian sediments. Soils, which result basically from processes of weathering of pre-existing materials, constitute distinctive markers (layers) in a sequence of Pleistocene sediments. Thus two layers of till may be separated by a soil or zone of weathering, developed on the surface of the lower till under a nonglacial climate and later overridden by the glacier that deposited the upper till. Hence the soil records a nonglacial interval between two colder times. The maturity and thickness of some such soils suggest that interglacial times were rather long but do not in themselves indicate how long they were.

Loess as a paleoclimatic indicator

Eolian sediments, such as dunes of wind-blown sand and broad blankets of the wind-blown silt known as loess (*q.v.*), occur as relicts over wide regions. Much of the loess is of glacial age, having been derived from the bare ribbons of sand and silt deposited in the valleys of major rivers by streams of meltwater as these flowed away from the ice sheets. Such loess, in itself, tells little of former temperatures; neither do the patches of relict sand dunes of glacial age that are widespread in North America and Europe. Some types of dunes, however, record wind directions; some of them indicate that effective winds were northwest in regions where today the prevailing winds are southwest. Although this relationship says nothing about temperature, it does suggest that in glacial ages the belts of westerly winds that characterize middle latitudes were somewhat nearer the Equator than they are today.

Runoff and carbonate dissolution. The behaviour of streams in nonglaciated regions through parts of Pleistocene time has been studied in an effort to learn the climatic history of such regions. Persistent study has been given especially to streams in the dry southwestern quarter of

the United States. Repeated fluctuation in the regimens of such streams, expressed in alternating cutting and filling, has been discerned. Very likely some such changes have resulted from climatic changes related to the glacial–interglacial sequence in higher latitudes. The factors that determine the regimen of a stream are so complex, however, that it is difficult to extract uncomplicated data on former climate from studies of this kind.

Dissolution and precipitation of carbonates appear to offer a means of determining former fluctuation of rainfall. The features of some caverns in carbonate rocks indicate former alternations between dissolution and refilling with either chemically or mechanically deposited sediments. Stalactites and other cave deposits (speleothems) in caverns have been measured for their temperature implications. In one attempt, $^{18}O/^{16}O$ isotopic ratios were measured at short intervals along radii of stalactites in order to determine the fluctuation of temperature during deposition through the last 30,000 years or more. In another, caverns throughout a wide region were examined for the crystal systems to which the calcium carbonate in their speleothems belongs. Calcium carbonate crystallizes as aragonite at temperatures above 17° C; below that temperature it crystallizes as calcite. In the northern part of the western United States speleothems consist of calcite, whereas in the southern part they consist of aragonite, and in an intermediate belt they include both. The development of such measurements could lead to useful data on the thermal history of a region.

Temperature and precipitation fluctuation. Many attempts have been made to learn the extent to which temperatures at the height of a glacial age differed from those prevailing today. The differences were much greater in some areas than in others, and knowledge is still very much generalized. Calculations based on the altitude of the inferred glacial-age snow line compared with that of today suggest that mean summer temperatures in high places were lower than they are now by some 5° to 9° C. Inferences drawn from the distribution of fossil plants are broadly comparable. Inferences from frost-action effects in European areas of continental climate suggest that mean annual temperatures there may have been more than 12° C below today's values in some places. Scattered data based on areal distributions of planktonic microfaunas in the ocean suggest that temperatures of surface seawater differed from those of today by amounts comparable with those derived from terrestrial sources.

Mean temperatures and precipitation

Data of the same kinds, based on features of interglacial rather than glacial age, seem to indicate mean temperatures comparable with, or at most no more than about 2° C higher than those of today. From this the conclusion is drawn that at present the earth's climates are closer to interglacial than to glacial conditions. In this connection it is worth noting that the entire range of mean temperatures cited is much less than today's range of mean annual values over the earth as a whole.

Precipitation, likewise, underwent changes during the Pleistocene. Physical and organic data suggest that in some areas it was greater than modern values for long periods; in other areas it was less. Nothing, however, compels the supposition that precipitation increased or decreased overall, although overall change is not excluded. Nor does increased precipitation seem necessary to explain the building of the Pleistocene glaciers. On reasonable assumptions it appears that all of the ice in the Northern Hemisphere, at its greatest extent, could have been built in a period of 50,000 years from only about 3 percent of today's annual precipitation on lands in the Northern Hemisphere. In this calculation the Southern Hemisphere can be neglected, because the Antarctic Ice Sheet was already in existence before the beginning of the Pleistocene.

Some authorities believe world precipitation during a glacial age is likely to have been as much as 20 percent less than that of today, but others believe that it may have been greater than modern values. Some evidence points to relatively dry glacial ages, but it is scattered and has not been fully evaluated; for this reason, the matter remains unresolved.

DIRECT EFFECTS OF GLACIATION

Pleistocene glaciers. Glaciers have existed on mountains and other highlands at many times during geologic history, and both the Antarctic Ice Sheet and other extensive glaciers are known to have existed in high latitudes during the last 10,000,000 years or so. But the Pleistocene seems to have been distinguished principally by a rather rare event, the building of large ice sheets in middle latitudes (Figure 1). The largest of these, the Laurentide

From R.F. Flint, *Glacial and Quaternary Geology* (1971); John Wiley and Sons

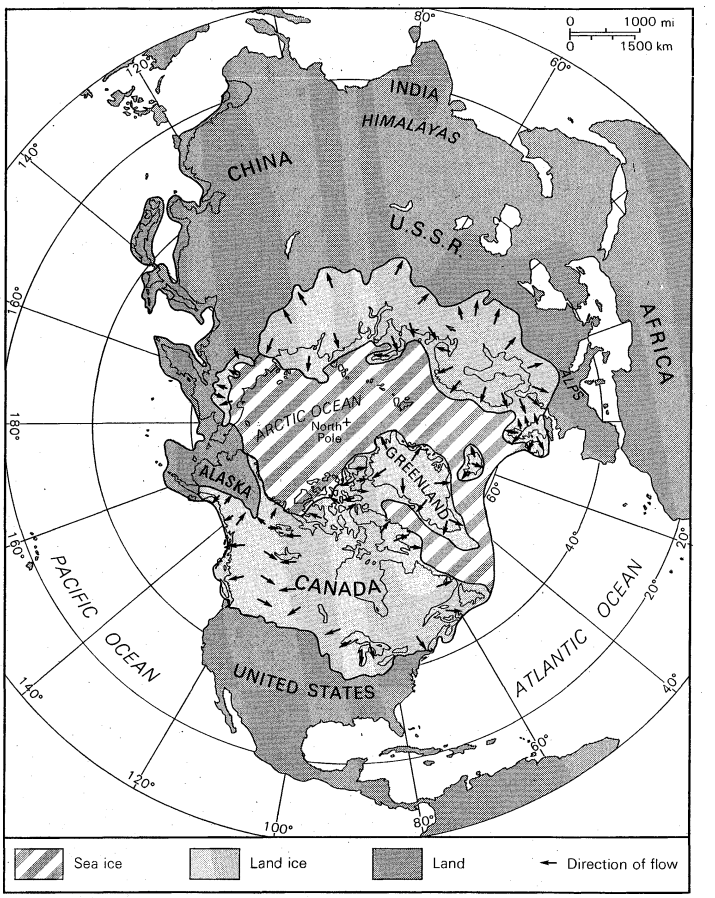


Figure 1: Northern Hemisphere showing areas glaciated during glacial ages.

Ice Sheet, covered an area of more than 13,000,000 square kilometres (5,000,000 square miles) and in terms of area was larger than the existing Antarctic Ice Sheet. Its thickness, unknown through direct measurement, may have attained a maximum of 3 kilometres (1.9 miles) or more. This ice sheet stretched from the Atlantic Ocean at New York City to the Rocky Mountains in Montana, and from the Arctic Ocean southward as far as latitude 37° N in the vicinity of the Mississippi River. Its southern margin stood near Cincinnati, Ohio; St. Louis, Missouri; Kansas City, Missouri; and Pierre, South Dakota. This great ice mass flowed outward in all directions, except where buried mountains distorted its flow. The rate at which its margin crept over the Ohio–Indiana region has been measured by means of the carbon-14 dates of spruce trees it knocked over; the rate averaged between 100 and 200 feet (30–60 metres) per year.

Laurentide and Scandinavian ice sheets

Second in order of area covered, the Scandinavian Ice Sheet (including all British glaciers) spread over about 6,600,000 square kilometres (2,500,000 square miles) according to present interpretation. There is fairly good reason to believe, however, that the ice sheet spread northward across the shallow sea floor to merge with the glaciers of Spitsbergen. If that did in fact happen, the maximum area covered probably would have exceeded 7,000,000 square kilometres (2,700,000 square miles). The southern margin of the Scandinavian Ice Sheet, at its maximum extent, reached the vicinities of London,

Amsterdam, and central Germany. Continuing through southern European Russia, it swung northward to the Arctic Ocean. Like its American counterpart, this ice body attained a probable maximum thickness of 3 kilometres (1.9 miles) or more. It was adjoined on the east by smaller ice sheets, while in eastern Siberia, China, and the rest of Asia all major high mountain ranges were capped with smaller glaciers of various sizes and shapes.

Conditions in North America were broadly similar. Although the Laurentide Ice Sheet did not overtop the Rocky Mountains in Canada, it was adjoined by, and in places connected with, a smaller ice sheet (some 2,300,000 square kilometres [890,000 square miles] in area) that extended from the Rocky Mountains to the Pacific Ocean, and from northern Alaska to Washington, Idaho, and Montana. South of it, as in Asia, high mountains as far south as southern California were dotted with local glaciers. The Arctic Ocean was covered with floating sea ice that surrounded glacier-covered Greenland and Iceland. The southern limit of sea ice stretched across the Atlantic from Newfoundland to Britain.

Ice in the Southern Hemisphere

In the Southern Hemisphere the Antarctic Ice Sheet was the only glacier of comparable size, mainly because a large part of that hemisphere consists of water. The southern Andes, however, were covered by a narrow ice sheet that subdivided northward into many separate groups of glaciers reaching right to the Caribbean Sea. Together, the South American glaciers covered an area of around 800,000 square kilometres (300,000 square miles). The South Island of New Zealand likewise was extensively covered with ice.

Apart from these large bodies of ice, a host of smaller glaciers occupied mountains in southern Europe and Africa, Japan, Tasmania, and other areas, some of them directly on the Equator. The combined area of all Pleistocene glaciers totals nearly 45,000,000 square kilometres (17,000,000 square miles), equivalent to about 30 percent of the world's land area, and more than three times the area occupied by glaciers today. A striking feature of formerly glaciated areas on mountains is that their asymmetry (slope locations) with respect to the mountains has the same orientations as the asymmetry of today's glaciers on the same mountains. This similarity implies in part that the wind systems bringing snowfall to the glaciers reach the mountains from broadly the same direction today as during the glacial ages.

Much of the terrain that formerly was covered by glaciers displays changes wrought by the glaciers themselves. Huge areas, including much of Canada and Fennoscandia (Norway–Sweden–Finland), were laid bare by the removal of regolith (rock mantle), and the underlying bedrock was ground down, smoothed, scratched, or polished by stones embedded in the bases of the glaciers, in a manner recognized by geologists as peculiar to glaciers. Other types of terrain, including areas of erodible bedrock and areas near the margins of former glaciers, were covered with heterogeneous deposits of earth, stones, and boulders, together known as glacial drift, spread rather thinly over the surface or heaped into prominent curving ridges at the terminus of a given glacier, in which case it is known as an end moraine. Such drift constitutes the basis for some of the best agricultural soils in the north central United States, but where its content of boulders is very great, as in most parts of New England, the soil developed in it is probably much poorer than the soil that existed in that region before glaciation. Glacial erosion and deposition likewise created the basins of the Great Lakes and most of the other lakes in northern North America and northern Europe.

Meltwater flowing away from the great ice sheets and the smaller glaciers followed pre-existing valleys and transported enormous quantities of gravel and sand through them to the sea. The Mississippi River was preeminent as a carrier of such glacial outwash, with which it built a massive delta into the Gulf of Mexico, a delta still being added to by the modern river.

Not only have rivers carried huge amounts of glacial sediment to the sea but the floors of the northeastern Pacific and the northern Atlantic oceans (to mention only two wide regions) have been liberally spread with sediments, including boulders, dropped to the bottom from floating ice shelves, or icebergs and pack ice (*q.v.*). In some of the cores taken from beneath the sea floor, two or more layers of such sediment have been recovered, separated by layers of normal, nonglacial origin. This reflects the alternation of cold climates with more temperate ones.

Glacial retreat and crustal recovery. The building up of thick glaciers on the earth's surface during glacial ages creates new, heavy, yet temporary loads, which the earth's crust is not rigid enough to sustain. The weak material of the substratum below the rocky crust flows plastically outward from beneath the overloaded areas as they subside, in basin-like fashion, to depths theoretically equal to about one-third the thickness of the overlying glacier. When the climate changes and the glacier begins to melt, the load on the crust is reduced, the direction of plastic flow of the subcrust material is reversed, and the rocky crust "recovers." It bulges upward as the overlying glacier melts away. The rate of uplift is very rapid at first, and then decreases nearly exponentially. Eventually, long after the ice has wasted away completely, the original, preglacial condition is restored. Such basin formation and later recovery of the crust is of the kind called isostatic.

Isostatic basin formation

Two classical examples of large-scale isostatic basin formation followed by rebound or recovery are Fennoscandia and glaciated eastern North America, where large ice sheets first depressed the crust and then permitted upward movement when they melted away. In both regions the updoming is still in progress many thousands of years after the ice sheets disappeared. The recent part of the movement is recorded by tide gauges along coasts. The earlier part is indicated by instrumental measurement of beaches and other shorelines that have been lifted above sea level and deformed. In Fennoscandia the outer limit of movement is a line that trends from southern Scandinavia northeastward toward Leningrad. South of that line the crust was not deformed isostatically. North of it, any strandline that represents a water plane made during or after deglaciation is tilted up in the northward direction. The tilt is clearly shown by contours that represent the altitude of the sea (Figure 2). It is predicted that a further 11 metres

From *Atlas over Sverige*

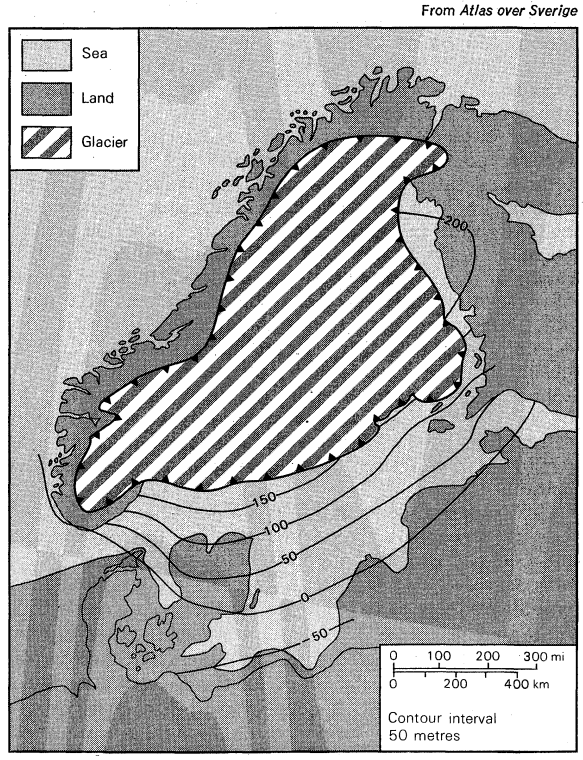


Figure 2: Distribution of former shorelines (contours) around the wasting Scandinavian ice sheet.

(36 feet) of the uplift still in progress will convert the Baltic Sea into a lake, which will gradually shrink to very small size, occupying only the northern part of the Gulf of Bothnia.

A similar history is reflected in the glaciated eastern part of North America. The dome in this region is much larger than the one in Fennoscandia, reflecting the larger size of the Laurentide Ice Sheet, but it has not yet been studied as thoroughly. It, too, is recorded both in the tide-gauge records and in measurements of deformed strandlines. The southern limit of deformation swings through a wide, irregular arc from Nova Scotia southwest to New Jersey. It is unknown across the Appalachians but is identified across the southern part of the Great Lakes region, where it trends northwest, passing near Cleveland, Detroit, and Milwaukee. Thereafter it continues northwestward and passes across northeastern South Dakota. Inside this limit, contours on various strandlines form a generally concentric pattern, with the centre of the pattern lying in the general region of Hudson Bay.

The parallelism between the outer limits of deformation and of glaciation is striking, both in North America and in the Baltic region. In both regions the concentric contours on deformed strandlines are related to areas, in North America to the Hudson Bay region and in the Baltic region to the Gulf of Bothnia, in each of which the related ice sheet is believed to have been centred during deglaciation.

In both regions continuing deformation is causing lakes to encroach on their southern shores and to recede from their northern shores. It has been predicted that in the future Hudson Bay, like the Baltic Sea, will disappear almost completely, although in the process it is not expected to become a lake. This is because of the fact that Hudson Bay's connection with the ocean is both deeper and differently placed.

Uplift similar to the examples described, although less spectacular, has been identified in the British Isles, East Greenland, the coastal region of British Columbia, and Patagonia. Indeed it is to be expected in any region in which former glaciers were massive enough to depress the crust beneath them.

INDIRECT EFFECTS OF GLACIATION

Fluctuation of sea level. The aggregate quantity of water at or near the earth's surface varies but little. Nearly 98 percent of it resides in the ocean, and a little more than 2 percent resides on land in the form of streams, lakes, ice, and snow; only a trace resides in the atmosphere. Hence, because most of the water on land resides in glaciers, the critical elements in the exchange between ocean and land are seawater and glaciers. Under today's climates with 10 percent of the land area covered with ice, the two elements are in equilibrium. In a glacial age they were in equilibrium, with the aggregate area of glaciers, and probably also their volume, increased by a factor of three. The glacial-age sea level, therefore, must have been much lower than today's sea level. Estimates vary, but according to an estimate that takes into account several complex factors, including isostatic adjustment, glacial-age sea level stood 100 metres (330 feet), in round numbers, below that of today. In essence, an ocean-wide layer of water was withdrawn by evaporation to build glaciers, and this water was later returned to the ocean, mainly through rivers, during the process of deglaciation.

Glacial-age sea level

The process of deglaciation has not rid the lands of glacier ice; 10 percent of the lands are still ice covered. Complete deglaciation of the lands would return to the ocean a volume of water that, after all factors are accounted for, would probably cause the level of the sea to rise a little more than 40 metres (130 feet). It appears, however, that the land has not been completely deglaciated during at least the last 10,000,000 years, for the Antarctic Ice Sheet seems to have been in existence during that much time. It is likely that at the interglacial maxima glaciers were less extensive than they are now. In that case, the sea at such times would have contained more water than it does now and its surface would have stood higher against land that was "stable." One reasonable estimate that has been made has placed a limiting height for interglacial sea levels at a figure of +20 metres (66 feet).

The difficulty here is that any part of a continent might move upward or downward through forces that have nothing to do with exchange of water between ocean and land. Hence a sinking coast could become submerged or a rising coast could emerge from the sea even while such exchange was in equilibrium. This is a principal reason why ideas as to the fluctuation of sea level are based only on estimates. Fluctuation cannot be determined closely at any locality along a coast where movement of the crust cannot be measured also.

Most broad movements of the crust, however, seem to be much slower than was the last great rise of sea level, evidence of which is found around most of the world's coasts. The evidence consists of a variety of features that were formed at or near sea level and that are now submerged to various depths. They include physical features such as beaches as well as fossils of many kinds. Where such things are datable by the radioactive carbon method it becomes possible to construct curves of time versus depth (Figure 3).

From F.P. Shepard, *Science*, vol. 143, pp. 574—76, figs. 1 and 2, Feb. 7, 1964

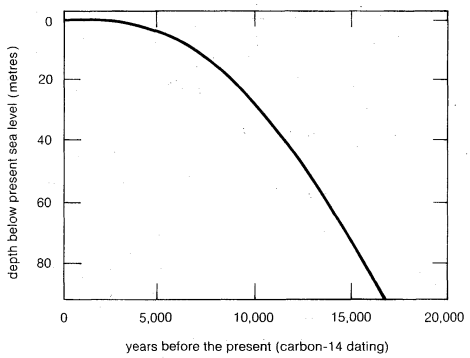


Figure 3: Generalized curve showing position of sea level, relative to land, since peak of last glaciation.

Curves of time versus depth, mostly spanning the last 10,000 years or so, agree broadly with each other where they pertain to coasts believed to be "stable." They indicate that the rise of sea level was rapid at first, then gradually slowed with the passage of time. The rise seems to have continued to the present. Authorities are not agreed as to whether a worldwide time–depth curve would be smooth or marked by fluctuation. Some, indeed, favour the view that the rising sea level reached above its present position at some time or times during the last few thousand years, although this concept is not readily supported by unequivocal evidence and consequently has few advocates.

Land bridges during glacial stades

If, as seems probable, the level of the sea was lowered during glacial ages by as much as 100 metres (330 feet), the effect on the positions of coasts would have been conspicuous. Along all coasts the area of the land would have been increased. The increase would have been slight along a steep, mountainous coast, but along a coast fringed by a continental shelf the increase of land would have been spectacular. The land area of the world might have been increased temporarily by as much as several percent.

An additional effect in parts of the world would have been the conversion of former islands into parts of mainlands, and even the connection of one continent with another. Thus, during one or more glacial ages, England was connected with France, Alaska with Siberia, Japan with Korea and perhaps also with Siberia, Borneo and Sumatra with the mainland of southeast Asia, Australia with Tasmania and with New Guinea, and Ceylon with India, to cite only a partial list. Such connections are postulated on a basis of the depths (less than 100 metres [330 feet] of the shallow shelves that now separate these pairs of land areas, and they are confirmed by certain striking

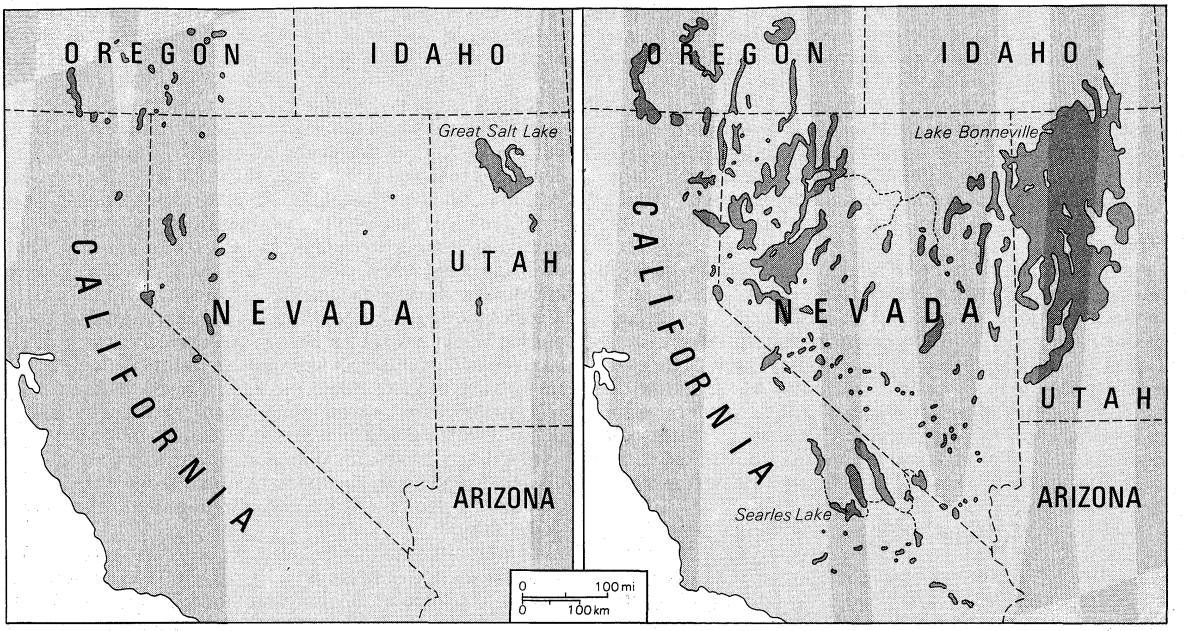


Figure 4: Distribution of (left) existing lakes and (right) pluvial lakes in the western United States. Dashed lines indicate probable drainage connections.
Courtesy of the Geological Society of America

similarities in the fossil faunas as well as by similarities in the living faunas of the land areas.

One of the largest and best known of the former "land bridges" is that now represented by the Bering Strait, which separates Siberia from Alaska. It was not merely a bridge, but a broad land nearly as wide as Alaska itself, that was created and re-created again and again, first by movement of the crust and later by glacial lowering of sea level. The evidence that animals travelled across it from one continent to the other, in both directions, is conclusive (see below *Pleistocene life: Migration of plants and animals*).

Even where no bridge was involved, the addition of a wide strip to an existing land mass influenced organisms in a significant way. The falling sea level in glacial ages may have added land as much as 170 kilometres (100 miles) wide to the Atlantic coastal region of the United States. This afforded an enlarged refuge for plants and animals driven southward by the encroaching ice sheet, as well as a convenient north–south migration route for various organisms.

Fluctuation of lake levels. The water economy of a lake is determined chiefly by precipitation and evaporation—in other words, by climate. In a moist climate the water level of a lake is controlled by the level of the outlet. Any increase in the water received by the lake simply increases the volume of the outlet stream. In a warm, dry climate, however, most lakes lack outlets. Under such conditions water level is controlled by the ratio of water received to water lost by the processes of evaporation and by leaking into the ground.

Because climates fluctuate, the balance between water received and water lost is rarely maintained for long; hence the levels of lakes in dry regions continually change. The Great Salt Lake in Utah, for example, has fluctuated in level through a range of about 18 feet (6 metres) since 1850, and the changes are known to correspond with measured changes of air temperature and rainfall.

Lakes responded to the changed climate in a glacial age in different ways, according to their environment. Most lakes in moist climates already possessed outlets, and those that received increments of meltwater from glaciers, particularly ice sheets, responded by passing the excess water along and discharging it through these outlets. The increased discharges eroded the outlets and lowered them somewhat, thus causing an enduring reduction of lake levels. Meanwhile, the meltwaters entering the lakes deposited much of their load of glacial sediment in the lake basins, thus creating depositional records of the glacial event.

In contrast, lakes in dry climates, which largely lack outlets, experienced a dual change. Under lowered glacial-age temperatures, evaporation diminished. At the same time the rainfall on many such lakes increased, as is shown by the kinds of plants, evident in the fossil record, that grew near the lakes. The increased rainfall, accompanied by decreased evaporation, caused lake levels to rise until they came into balance with the new climatic conditions.

Pluvial lakes in North America and Africa

In the basin of the Great Salt Lake, a water body that is very shallow today, the water rose to form a lake (called Lake Bonneville) at least 300 metres (985 feet) deep in its deepest part and nearly 500 kilometres (310 miles) in greatest length. It spilled over temporarily through a mountain pass. Subsequently it fell and rose again repeatedly, in response to changes in temperature and rainfall, before reaching its present shrunken condition. Its various levels are recorded by beaches and other strandlines that can be traced along the slopes of the enclosing mountains. Its fluctuations are reflected also in fine-grained sediments that are recovered from cores bored from the floor of the former lake.

Another striking example is Searles Lake, now a dry salt-encrusted flat, between high mountains in southeastern California. In the latest glacial age the changed precipitation–evaporation ratio permitted a lake to form and rise to a height of at least 230 metres (755 feet) above its floor. At that height it overflowed through a mountain pass into an adjacent basin. Subsequently, it too diminished, leaving many strandlines along the mountainsides. Many core borings, one of them more than 250 metres (820 feet) deep, reveal layers of two kinds. One consists of clay and silt, deposited in a deep lake and in places containing fossil fishes. The other consists of mixtures of fine sediment with salts precipitated on the lake floor by evaporation of the water. Radioactive carbon dates of these deep-lake and dry-lake sediments agree closely with those of glacial and nonglacial deposits in other parts of North America and demonstrate that the deep, pluvial lakes existed during times of glaciation. Records of pluvial lakes are abundant in the western United States (Figure 4), northern Mexico, and the dry Pacific slope in middle South America. One of the pluvial lakes that existed in Bolivia was nearly as large as the American Lake Bonneville.

The former Dead Sea in Jordan and Israel, 200 metres (656 feet) deep in the latest glacial age, was the first plu-

vial lake to be recognized as such; it was identified in 1865. One of the largest known pluvial lakes is the former Lake Chad, now a shallow saline lake (latitude 13° N, longitude 14° E) in the southern Sahara. In its pluvial phase the lake is thought to have attained an area of some 300,000 square kilometres (115,800 square miles; roughly three times the area of Lake Bonneville); it may have been the largest pluvial lake in the world. Other pluvial lakes have been recorded in northern Africa, western and central Asia, and central and southern Australia.

Pluvial lakes seem to be related in time to glacial ages in most regions for which information exists. There appear to be exceptions among the several lakes that occupy basins in the Rift Valley system in eastern Africa, however. Certain lakes seem to have remained at high levels much longer than those in other parts of the world, indeed until after the great ice sheets in the Northern Hemisphere had wholly disappeared. The cause of this difference is not presently known but may be related to peculiarities in the history of the climates of northern and eastern Africa.

Fluctuation of stream regimen. Although the fluctuation of lakes in Pleistocene time gives a fairly clear indication of the relationship to changing climates, changes in the regimens of streams are more difficult to identify and to interpret in climatic terms. Two stream parameters in particular that probably changed with changing climate are (1) discharge and (2) rate of transport of sediment. Both of these parameters are known to depend on a number of factors, of which rainfall is one. But no known direct relationship exists between either one of these parameters and the quantity or time distribution of rainfall.

Although today's discharge in a stream can be measured, former discharge is very difficult to determine. A means by which attempts have been made to approximate former discharge is measurement of the geometry (principally width, depth, and gradient) of stream channels. When discharge increases, the geometry of the channel changes accordingly. In some streams glacial-age changes were apparently so great, and have been so well preserved, that inferences as to former discharge could be attempted.

Terraces and stream meanders

One example is the Valley of the Nile in Nubia. Studies of ancient alluvium left as terraces along the sides of the river show that through a time corresponding with the latest glaciation of northern Europe the Nile repeatedly aggraded its valley floor with alluvium, building up its floodplain to as much as 30 metres (98.4 feet) above its present position. At the same time local tributaries, whose discharges today are zero, brought to the Nile Valley abundant sediment. Evidently both the Nile and its tributaries in Nubia were formerly characterized by greater discharge, and probably also by a more rapid rate of transport of sediment.

An example of a different kind consists of streams that meander in valleys that are themselves meandering on a larger scale. The ratio of wavelength of the former (valley) meanders to existing meanders is large, being as much as 10:1. Still other streams meander on floodplains beneath which are buried meandering channels; again the wavelength ratio between the two is large. Because wavelength is related systematically to channel width, and hence to discharge, it is held that in the streams mentioned above, discharge (at least at flood peaks) was formerly greater. Calculations based on wavelength ratios in valleys of the kind mentioned have yielded former (peak) discharges as great as 50 times existing discharges. In some cases the geologic relationships point to the later part of the latest glacial age as the time of last great discharge. Such large discharges may have been attained in only a few very exeptional floods and may have been aided by the existence, beneath the stream, of frozen ground that prevented a loss of water from the stream by infiltration. Although calculations of this kind provide very interesting material for speculation, they have thrown little light on the problem of rainfall during glacial ages.

Attempts to determine glacial-age rates of transport of sediment by streams likewise encounter difficulties. Streams that originated at the margins of ice sheets during glacial ages consisted mainly of meltwater and carried large volumes of outwash sediment. Such sediment was deposited continuously and so diminished in both volume and grain size in the downstream direction. The deposited material is generally thick, at least in upstream positions. In the Mississippi River drainage basin it extends from Montana on the west and Pennsylvania on the east continuously to the Gulf of Mexico. In this and other drainage basins, the streams have trenched deeply into outwash sediments in postglacial time, leaving them as terraces along valley sides. Such trenching, however, was not the direct effect of a change from glacial-age climate to postglacial climate. It was only an indirect effect, and a remote one at that, because in each stream the change in regimen occurred at the time when melting reached a point at which glacier ice ceased to deliver meltwater and sediment to the stream. With consequent changes in both discharge and load, regimens changed from one of predominant upbuilding to one of predominant erosion of valley floors.

Sediment transport in glacial streams

In summary, analysis of outwash sediment in a stream valley does not indicate whether glacial-age stream discharge was greater or less than today's nonglacial discharge. It does, however, suggest that rates of transport and deposition of sediment were greater in a glacial age than they are today.

Whether such changes in rates of transport characterized all streams, even those that were not influenced directly by glaciers, is not known with certainty. Sediment cores raised from beneath the floor of the Atlantic Ocean, which receives the drainage from about two-thirds of the world's watershed area, throw some light on this matter. The part of the sediment that was derived from the land rather than from the activities of marine organisms, and which is so situated as not to have been interfered with by sediment brought in by bottom currents, seems to have been deposited at considerably more rapid rates during glacial than during nonglacial times as calculated from radioactive carbon dates. This relationship suggests that transport of sediment from land to sea (mostly the work of streams) increased during glacial ages, although the cores reveal nothing about discharge.

Fills and cuts are common in the valleys of streams of nonglacial origin, particularly those in dry regions, and some of them are dated by means of radioactive carbon. Such features reflect changes in the regimens of the streams, but the changes are difficult to interpret in climatic terms. Factors in climate, such as amount of precipitation, distribution of precipitation through the year, and annual and seasonal temperature cannot easily be singled out. Other factors that affect stream regimen include (1) drainage area, range of altitude, and slopes in the watershed; (2) topographic texture; and (3) character of the soils and vegetation. A change in only one of the factors in climate might lead to different responses in two different streams, and even in two different segments of a single long stream. Because of this complexity, it has not yet become possible to establish a general relationship between Pleistocene changes of climate and the regimens of streams.

Wind action and wind direction. Just as the snow line, the extent of glaciers, and the volumes of lakes in dry regions fluctuated between glacial and nonglacial ages, so the direction of winds changed, at least in some regions. The changes are recorded in windblown sediments that accumulated in former times, and that since those times have been protected from erosion.

Winds move grains of sand by saltation (bouncing contact) close to the ground, and particles of silt in suspension through much greater heights. These sediments are picked up from sandy and silty surfaces not protected by vegetation, are carried downwind, and are deposited. The sand normally is deposited close to its source, in the form of dunes or, less commonly, in thin blankets; the silt accumulates in broad sheets that extend a much greater distance from the source area. Common sources include

ocean and lake beaches, alluvium in stream valleys, and basins in dry regions.

Eolian deposits and wind direction

Fields of modern windblown sand and silt are comparatively small. Most if not all of the large fields originated in Pleistocene time or even earlier. Among the large fields that consist primarily of sand is the one called the Nebraska Sand Hills in the Great Plains region of North America. In it sand covers an aggregate area of more than 35,000 square kilometres (13,500 square miles) to an average depth of about eight metres (26 feet), and includes at least three groups of dunes of different late Pleistocene ages. A far larger field, or group of fields, is represented by the great dune systems that cover most of central Australia. Still others are the several fields that occupy basins in northern Africa.

Many of the fields of windblown sediment contain indications as to the direction of the predominant winds under which they accumulated. Windblown silt, when sampled systematically and analyzed, shows a remarkably orderly decrease of both thickness and grain size in the direction downwind from the source. Windblown sand, most of which occurs in the form of dunes, some of them more than 100 metres (330 feet) in height, reflects wind direction in two ways. One is the direction of slope of the stratification within the downwind faces of dunes. The other is the external form of the dune itself. Dunes have several characteristic shapes, most of which are related systematically to the direction of the prevailing wind.

Few Pleistocene dune groups have been studied closely, but in some instances the indicated wind directions are different from those of today's winds in the same areas. In middle latitudes in both North America and Europe, where prevailing winds generally have a westerly component, some Pleistocene dunes built close to former ice sheets indicate effective winds of north and even northeast direction. Others indicate northwest winds where prevailing winds are southwest today. The indicated former directions generally are compatible with a glacial-age pattern in which wind belts occupied positions somewhat nearer the Equator than today, and ice sheets generated local winds having northerly components along the southern margins of the glaciers. Observations are still too few for good synthesis, but this method is likely to help substantially in the reconstruction of glacial-age wind belts.

Windblown silt (loess) occurs mainly in two situations: (1) In the lee of arid basins. The extensive loess in China is chiefly of this kind. The source material consists mainly of alluvium washed onto basin floors from the erosion of rocks that form the sides of the basins. Most such loess is unrelated to glacial outwash; probably most of it accumulated at times when the regional climate was dry. (2) In the lee of valleys that were being actively filled with glacial-outwash sediments during glacial ages. Floods of meltwater added annual increments of sediment that overwhelmed the local vegetation and thus maintained extensive bare surfaces of sediment subject to erosion by winds. Most of the extensive loess of the central United States is of this kind; it accumulated only as long as glaciers were present in the drainage basins of the source streams. Thereafter vegetation gained the upper hand on valley floors and so inhibited further wind erosion. Related wind directions are determined in the manner mentioned above.

Soil formation. Soils of Pleistocene age are widely preserved by burial beneath coverings of glacial drift, loess, or volcanic rocks, and others lie unburied at the present surface. Because soils have characteristics impressed on them by the environments in which they were formed, these ancient soils afford clues to the former environment. Factors in the environment that have been recognized in some ancient soils include topography, vegetation, and climate.

Characteristics of Pleistocene soil

If an ancient buried soil can be traced through closely spaced exposures, the topography at the time of burial can be inferred. Such a soil that today is confined to hills and is lacking in the intervening valleys indicates that cutting of the valleys occurred after the soil developed. The change could have resulted from movement in the earth's crust or alteration of the climate, particularly as regards rainfall; either change could have caused valley cutting.

A different sort of change is represented by a soil of one zonal type superposed on or partly obliterating a soil of another zonal type, in a relationship referred to as degraded. An ancient middle-latitude soil, for example, may be partly converted to a soil with remnants of secondary carbonate that is characteristic of the first type. The relationship would suggest a change of climate from semiarid to humid, with invasion of grassland by forest, such as might occur at the beginning of a glacial age. Another example consists of red and yellow soils, on the fringes of the subtropics, that have been podzolized. This degradation, too, suggests change to a cooler climate accompanied by the incursion of trees into grassland.

In parts of central North America and central Europe there are exposed stratigraphic sequences in which layers of loess alternate with ancient soils. Such occurrences represent the fluctuation of climate between glacial times of loess deposition and less cold times of relative stability of the surface, when loess was not accumulating. Several such fluctuations could occur within a single glacial age. Interglacial ages are marked in many areas by the presence of thick, mature, ancient soils.

It is not easy to generalize on the time required for soil making, both because many variable factors are involved and because few reliable dates between which measurements can be made are available. Indirect evidence suggests that some chernozem soils have developed to maturity within less than 200 years, that a thin podzolic soil in New England developed in about 1,500 years, and that many podzolics in moist-temperate environments have matured in less than 4,000 years. The dated stratigraphy of one buried soil in Illinois indicates that it formed between 22,000 and 28,000 years BP, although the proportion of the intervening 6,000 years that was involved in its development is not known.

In North America, soils developed in Wisconsin time differ, as a general rule, from those developed earlier. The latter are considerably more mature, with thicker horizons and greater development within them, including the formation of red and yellow iron oxides. In Europe, similar differences exist between soils related to young drift and those related to older drift.

Distribution of permafrost

Under the climatic regimen of today, permafrost in the Northern Hemisphere extends southward, even in discontinuous patches to only 52° N in North America and to 66° N in Europe, positions at which mean air temperatures are a little above 0° C. In the latest glacial age permafrost is believed to have extended much farther south. This statement is based principally on the following reasoning. Frost cracks, and the ice wedges created from them, are active today, not in the peripheral part of the permafrost region but only in the more central part where the mean air temperature is −6° C or lower. Ice-wedge casts, a relict form of ice wedges, are present in drift of the latest glaciation in central Europe, penetrating to depths as great as eight to ten metres. Because mean temperatures at least as low as −6° C seem to be required for the development of such structures, glacial-age mean temperatures in parts of central Europe are thought to have been lower than they are now by 15° to 18° C, and the southern limit of permafrost is thought, by analogy, to have stood farther south than central Europe. In central North America, likewise, relict ice-wedge casts of glacial age are known. They imply former mean temperatures 10° to 15° C below those of today. Permafrost probably did not extend as far beyond the southern limit of glacier ice as it did in Europe, however, because the configuration of the North American continent results in warmer climates than those of Europe.

PLEISTOCENE STRATIGRAPHY

Pleistocene strata represent sediments of the same kinds as those that are accumulating today. They differ from today's sediments principally in their areal distribution. Sediments deposited during glacial ages, for example, include more widespread glacial strata, and at the same

time more restricted marine strata on continental shelves, than those now accumulating. Such differences constitute primary clues to changes in environment. Pleistocene strata differ from Pliocene and older strata in being generally much thinner and little consolidated, mainly because of the comparatively short segment of geologic time they represent.

Like modern sediments, Pleistocene strata are commonly very discontinuous. The discontinuity increases the difficulty of correlating strata from one place to another. Correlation is effected through traceable physical features and relationships, through similarities in the fossils the strata contain, and through similarities in geopolarity and in radiometric dates. Extensive, sound correlation yields the relationships necessary for the construction of paleogeographic maps of regions in particular units of time. The small map that accompanies this article is essentially such a map, though with little detail. More detailed examples can show the distribution of natural features, environments, and groups of organisms at a given time. The former distribution can thus be compared with that of today.

Stratigraphic data are likewise the means of reconstructing historical sequences, in which are shown changes in natural environment through time. Both maps and developmental changes are, of course, best known from the youngest and highest strata, which are least subject to loss through erosion and to concealment through subsequent burial, and most amenable to dating by means of radioactive carbon. Thus the events of the last 25,000 years, which include the latest great glaciation, are those that are understood in greatest detail.

PLEISTOCENE LIFE

Pleistocene plants and animals differed in some respects from those of today, although many were identical with living forms. The differences are the result of at least three factors: (1) the unending process of organic evolution, (2) changes of range because environments changed, and (3) mass extinctions. Each of these factors will be considered in turn.

Evolution and adaptation to cold climates

Pleistocene evolutionary changes in living things represent an unbroken continuation that had been taking place throughout earlier geologic history. Under the influence of continuously changing environments, existing species became extinct and new species evolved. Among mammals, for example, only 6 of the 119 species that now live in Europe and adjacent Asia are present as fossils in Pliocene strata. All the others first appeared in the Pleistocene, or even later. The Pleistocene evolutionary history of plants is more conservative. Some 80 percent of the plant species now living in northwestern Europe were already in existence at the beginning of the Pleistocene Epoch.

Changes in environment that were principally responsible for the evolutionary changes stemmed in large part from uplifts of the earth's crust and from localized mountain building that together have been conspicuous since the middle part of the Cenozoic Era. These movements resulted in generally lowered temperatures and in physical compartmentation of the land. Uplifts and the creation of mountain barriers to rain-bringing winds in turn enhanced the continentality of climates. Apart from these changes arising from earth movements, there seem to have been other changes in temperature and rainfall that are thought to have been caused independently. Both, however, affected evolution by promoting extinctions, adaptations, and speciation.

Mammalian fauna. Specific adaptations to cold climate, as represented in large mammals, have attracted much attention because they are conspicuous. Most of these, however, affected fleshy and other soft parts and did not necessarily affect the skeleton. Examples include the development of wool (a thermal insulator) in elephants, rhinoceroses, and musk-oxen; of blubber (an insulator and a source of energy in winter) in polar bears; and of food- and water-storage structures in some camels. Apart from such adaptations, which in any case did not appear until rather late in the Pleistocene, the majority of the changes in mammals are not related to low temperature in an evident way. The repeated changes of temperature that characterized the later part of Cenozoic time could, however, have stirred and mixed populations by forcing many changes of range and thus might have induced evolutionary changes of other kinds.

Villafranchian fauna

The earliest Pleistocene fauna of land mammals in Europe, widely termed Villafranchian, differed from the preceding Pliocene fauna in that it included new groups such as true elephants, zebrine horses, and cattle, but it contained no known elements that were obviously or exclusively of cold-climate character. It endured through a long time, possibly amounting in the aggregate to more than 2,000,000 years. A somewhat comparable fauna in North America took form even earlier and so endured even longer.

Within the time between the Waal and the Cromer Interglaciations, the Villafranchian fauna was so greatly modified by extinctions and new additions that a new fauna can be said to have appeared, one generally labelled middle Pleistocene. It included mammals of "Arctic" affinity such as reindeer, musk-oxen, and lemmings, followed after a time by woolly mammoths, woolly rhinoceroses, and moose.

Following the Saale Glaciation another marked change occurred among European mammals. Certain large, non-"Arctic" mammals, including the straight-tusked elephant and the hippopotamus, became extinct and the saiga antelope appeared among the existing "Arctic" forms. The result was the late Pleistocene fauna.

In North America the earliest Pleistocene fauna was the Blancan, which resembled the European Villafranchian fauna in some respects but which seems to have begun earlier, dating from some 4,000,000 years BP (as compared with perhaps 3,000,000 years BP for the Villafranchian). The Blancan included many relict Pliocene forms but was marked by the appearance of zebrine horses from Asia and by modern camel-like forms.

The succeeding Irvingtonian fauna seems to have appeared late in the Kansan Glaciation. Its most distinctive characteristic was the presence of a large number of immigrants from Asia, including mammoths, caballine horses, woodland musk-oxen, antelopes, and many rodents. During the succeeding glaciation, the Illinoian, further extinctions and new arrivals created the Rancholabrean fauna. Among the new elements were bison, skunks, bats, and many other rodents. Decimation of that fauna occurred near the end of the Wisconsin Glaciation.

Migration of plants and animals. More conspicuous than evolutionary changes, yet springing mainly from the same causes, were alterations in the ranges occupied by species and indeed whole populations of animals and plants. These consisted of the displacement of populations from their accustomed areas, mainly through the agency of changing climate. When a glacial age set in, organisms were displaced downward on the sides of mountains and plateaus and moved toward the Equator in broad regions of lesser relief. The process was one of slow spreading toward more hospitable areas. The result for the organisms was minimal disturbance of their environment, a result that for many organisms was critical for their survival. The displacements were accompanied by little apparent evolutionary change, but this may have been only because glacial ages were comparatively short, permitting the early return of most organisms to their normal ranges.

Both the spreading of organisms at the beginning of a glacial age and their return at its end would have been very irregular. Some elements of a community would spread while others remained behind, depending on their particular tolerances. Ranges would overlap and mixing of populations would occur. That these things happened is indicated by the observation that local floras and faunas were not alike in successive glacial and interglacial ages, even where there is little reason for believing that the successive climates differed from each other.

Apart from plants and land animals, marine organisms underwent similar displacements. Planktonic microfaunas display this history particularly clearly because they

can be recovered as fossils in ocean-floor sediments by means of coring and because their large numbers are well adapted to statistical treatment. Boreal species, for example, have been repeatedly displaced toward the Equator at times of reduced temperature, and vice versa.

Displacement of marine organisms

In the vicinities of large ice sheets, shifting of plant and animal populations was the result of invasion by the ice as well as of general cooling of the earth's atmosphere. In regions unaffected by ice sheets, amplitude of shifting was generally less and was the result wholly of secular rather than local causes. The distribution of glacial drift, together with fossil pollen and fossil mammals, yields the following rough sketch of environments south of the ice sheet in the region below the existing Great Lakes, 17,000 to 20,000 years ago: The margin of the ice lay across southern Ohio, central Indiana, and central Illinois, with streams of meltwater flowing away toward the south. Cloudy weather was common, with much snow and rain resulting from interactions between warm, moist air masses from the Gulf of Mexico and the high, cold body of ice. In a narrow zone close to the glacier, intense frost activity occurred, and the vegetation cover consisted of barren, tundra-like grassland with scattered clumps of spruce trees. Toward the south this vegetation merged rapidly into spruce forest with birch, alder, and tamarack. A boreal forest of this kind extended far southward, but with ever increasing proportions of deciduous trees characteristic of temperate climate. The stations at which fossil mammals of that time have been collected are not spaced closely enough to reveal animal distribution in detail but show clearly enough that the relatively barren zone close to the ice sheet was inhabited by, among others, woolly mammoths, caribou, and collared lemmings, while the forest south of it was occupied by mastodons, ground sloth, and moose. In addition there were horses, bison, woodland musk-oxen, snowshoe hares, and many kinds of smaller mammals.

Mass extinctions. The remarkable difference between today's fauna of large land mammals and the far richer fauna that existed at the height of the latest glaciation is the direct result of a series of extinctions. These were both limited in time and apparently without precedent in earlier glacial ages, and they left many ecological niches unfilled, a most unusual phenomenon. Estimates of the proportion of large-mammal genera that became extinct in North America toward the end of the latest glacial age run as high as 70 percent, although so high a number is not accepted universally. The number in Eurasia is considerably smaller, as is the case also in Africa and South America. The extinctions involved not only large mammals but also large flightless birds, as in New Zealand, and small mammals, particularly on islands. Most of the activity occurred between 15,000 and 9,000 BP, although in New Zealand it occurred much later.

Causes of faunal extinction

The two most likely causes of the mass extinctions, which have been much debated, are (1) changing environment with changing climate and (2) the direct or indirect influence of early man. In its common form the climatic hypothesis specifies that the warming and drying of middle-latitude climates that accompanied deglaciation brought with it changes in ecology, especially vegetation. These changes are presumed to have created crises in the kinds and abundance of available food, with which large herbivores could not cope.

Three arguments urged against the climatic hypothesis, mainly in connection with the North American scene, are as follows: (1) independent evidence is lacking that the climatic warming in late Wisconsin time was greatly different from the warming that introduced earlier interglaciations, in which extinctions were fewer; (2) the change of climate involved more time overall than did the majority of extinctions, which clustered near 11,000 BP; (3) application of the hypothesis to a number of the mammals that became extinct is not clear: desert herbivores such as camels and desert ground sloths, for example, should have flourished in consequence of the enlargement of arid lands, instead of becoming extinct; steppe grazers such as bison had only to move northward in order to keep within their favoured environment, yet only one species survived; with few exceptions browsers such as mastodons and forest musk-oxen should have been able to do likewise, yet both succumbed.

The hypothesis of human influence embraces not only the direct slaughter of large mammals by Paleolithic hunters but also the pressure exerted by man as part of the whole ecologic pattern. Human disturbance of the environment might have included selective killing of young individuals and of genera having exceptionally long periods of gestation.

The most sweeping and frequently heard arguments against the hypothesis of human influence is that man was not sufficiently numerous to have caused such extensive extinction, despite the evidence of stone-tipped, high-velocity projectiles and of group drives in which hundreds of large mammals at a time were stampeded and driven over cliffs.

Probably no single cause can explain the known data on the extinctions, and probably the circumstances differed from one region to another. Overall, however, it seems likely that human influence was a significant factor in these remarkable events.

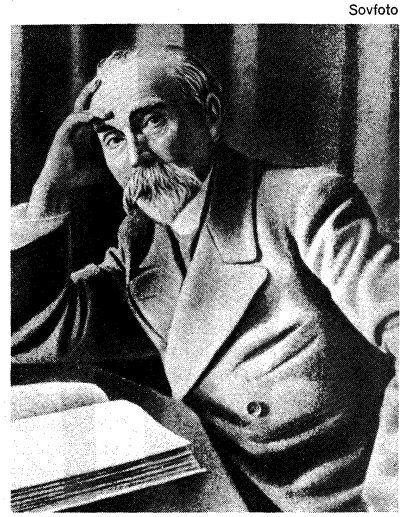
BIBLIOGRAPHY. Summary works on the Quaternary, with extensive bibliographies, include: R.F. FLINT, *Glacial and Quaternary Geology* (1971); P. WOLDSTEDT, *Quartär* (1969); and J.K. CHARLESWORTH, *The Quaternary Era*, 2 vol. (1957). All place emphasis on physical features. R.G. WEST, *Pleistocene Geology and Biology* (1968), emphasizes biologic features, with special attention to the history of plants as revealed by fossil pollen, and with the discussion confined mainly to Britain. B. FRENZEL, *Die Klimaschwankungen des Eiszeitalters* (1967), is concerned primarily with the glacial-age distribution of plants in Eurasia. For both physical and biologic aspects of the subject, see H.E. WRIGHT and D.G. FREY (eds.), *The Quaternary of the United States* (1965). P.S. MARTIN and H.E. WRIGHT (eds.), *Pleistocene Extinctions*, (1968), is confined entirely to the extinctions of mammals that occurred in many parts of the world at the end of the Pleistocene Epoch.

(R.F.F.)

Plekhanov, Georgy Valentinovich

The founder and for many years the leading exponent of Marxism in Russia, G.V. Plekhanov was born into a family of the minor gentry on November 29 (according to the Russian old style calendar, then 12 days behind the Western Gregorian calendar), 1856, in Gudalovka, in Tambov province. In 1873, after completing his studies at the Voronezh Military Academy, he enrolled in St. Petersburg's Konstantinovskoe Military School, with the intention of becoming an army officer.

Sovfoto



Plekhanov.

He soon transferred to the Mining Institute, and then, during his second year there, abandoned his studies to devote himself entirely to a populist revolutionary movement. Although this movement aimed to foment a peasant upheaval and to establish an agrarian socialist society,

Plekhanov's activity involved him particularly with urban factory workers. After becoming a leader of the populist organization Land and Freedom in 1877, he engaged in underground political agitation. His writings appeared in Land and Freedom's publication of the same name; some of them showed Marxist influence. When Land and Freedom turned increasingly to terrorist methods, Plekhanov formed an antiterrorist splinter group to continue mass agitation. This faction was, however, short-lived, and in 1880 Plekhanov went abroad to avoid arrest. He did not return to Russia until 1917.

For the greater part of his exile Plekhanov resided in Geneva, Switzerland. In 1883, with several friends, he established the first Russian Marxian revolutionary organization, Liberation of Labour. In two major works, *Socialism and Political Struggle* (1883) and *Our Differences* (1885), he launched a destructive critique of populism and laid the ideological basis of Russian Marxism. Russia, he argued, had been caught up in a capitalistic development that was altering its social structure and creating the conditions for the overthrow of Russian autocracy and the establishment of a bourgeois-democratic regime. The Marxists, according to Plekhanov, were obliged to organize the emerging industrial proletariat for the struggle against autocracy. After its overthrow further capitalistic development would multiply the numbers of the proletariat; and, under the leadership of a social democratic labour party, the working class would ultimately liberate itself through a socialist revolution. This two-phase revolutionary scheme lay at the centre of Russian Marxian thought through several decades.

Birth of Russian Marxism

During the ten years or so after the publication of *Socialism and Political Struggle*, Plekhanov's group produced a mass of socialist literature, but the Marxists remained isolated from the Russian working class. Toward the mid-'90s, however, after a renewal of antigovernment activity, the group won some notable followers, including Vladimir Lenin. Revolutionaries from the intelligentsia established relationships with workingmen to promote labour struggles against management. Plekhanov and his friends joined with these groups, whose activities culminated in 1898 in the formation of the Russian Social Democratic Workers' Party.

In the '90s Plekhanov continued his polemics against the populists, most importantly with his book *On the Development of the Monistic Conception of History*. In 1898 he began publishing a series of tracts defending Marxian orthodoxy from those who proposed modifications or deviations. His principal targets were the reformist revisionism propounded by the German Social Democrat Eduard Bernstein, and a Russian tendency called "economism," which purportedly deprecated the importance of the direct political struggle and, rather, encouraged efforts to improve the position of the working class through labour pressure on the factory owners. These campaigns were waged in concert with Lenin and some others, who, in 1900, had joined hands to publish the militant journal *Iskra* ("The Spark"). By 1903 their foes in the Russian movement had apparently been routed, and the second congress of the Russian Social Democratic Workers' Party was held. Unexpectedly, a controversy over the character of the party split the congress into what became known as the Bolshevik and Menshevik factions. Plekhanov initially sided with Lenin, the Bolshevik leader, but soon drew away and joined the Mensheviks in attacks upon him. For the remainder of his political life, and especially between 1906 and 1914, he strove in vain to reunite the party.

Break with Lenin

In the last two decades of his life, Plekhanov was deeply absorbed in philosophical, artistic, literary, and historical studies. In all these fields he made significant contributions. But he did not completely retire from political life. A prominent member of the Second International, he assumed a "defeatist" stance toward his country in the Russo-Japanese War of 1904–05. By contrast, he supported the Entente during World War I, as he believed that the victory of German militarism would unquestionably spell disaster for the progressive workers' movement in Russia and elsewhere.

The Revolution of 1905 tested the revolutionary scheme Plekhanov had devised in the 1880s and found it wanting. The bourgeoisie did not act as he had anticipated they would. The peasantry, whose role he had underestimated, proved to be a potent revolutionary force. Yet he did not modify his theory in any fundamental respect. His influence declined noticeably between the two revolutions.

In 1917 Plekhanov greeted the February upheaval as the long-awaited "bourgeois" revolution. Returning to Russia in 1917, he urged continuation of the war to victory. The soldiers heeded him no more than did the peasants and workers whom he urged to subordinate their political demands to the more pressing need for national survival. Though critical of the Provisional Government, he supported many of its key policies. He was impotent to stop the Bolsheviks' march to power, and as a consequence of his attacks on them he was harassed by overzealous Red Guards as "an enemy of the people." He died on May 30, 1918, in Terioki, Finland.

BIBLIOGRAPHY. For biographical information, see S.H. BARON, *Plekhanov: The Father of Russian Marxism* (1963). The most complete edition of Plekhanov's works is his *Sochineniya*, 24 vol. (1923–27). Among Plekhanov's works translated into English are: *Art and Social Life* (1953), *The Development of the Monist View of History* (1947), *Fundamental Problems of Marxism* (1929), and *History of Russian Social Thought* (1938, 1967).

(S.H.B.)

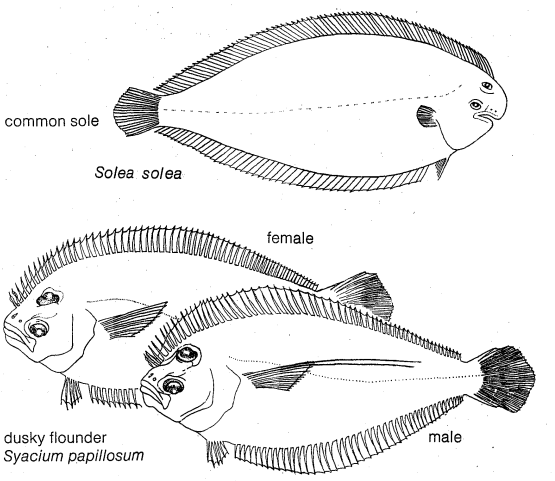
Pleuronectiformes

Members of the fish order Pleuronectiformes, commonly called flatfishes, flounders, or soles, are unique among fishes in being asymmetrical (strongly compressed, with both eyes on one side); other fishes and vertebrates in general are bilaterally symmetrical. The asymmetry is believed to have evolved from a generalized, symmetrical percoid (sea bass) body pattern in a fish that habitually rested on its side. Larval flatfishes have an eye on each side of the head, but during a period of rapid body change (metamorphosis) one eye migrates to the other side, after which the larvae settle to the bottom. Osteological changes resulting from the eye migration are responsible for the asymmetry in the flatfish skull.

Flatfishes of the family Pleuronectidae are commercially important in northern waters, and members of other families are taken in limited quantities. Some Bothidae and Soleidae (soles) are exploited in tropical and temperate waters, but no other flatfishes are utilized to the extent that Pleuronectidae are. In 1969, 1,400,000 tons of flatfishes were landed by the world's fisheries, and 85 percent of these were Pleuronectidae.

Natural history. Flatfishes generally spawn offshore, but some spawn in estuaries. Fecundity is high, females generally releasing at least several hundred thousand eggs (large female halibut have between 2,500,000 and

By courtesy of (centre and bottom) U.S. Fish and Wildlife Service; from (top) S.F. Harmer and A.E. Shipley, *Cambridge Natural History* (1904); Macmillan & Co., London and Basingstoke



Body plan and sexual dimorphism in the flounder.

3,000,000 eggs). The eggs are small and float freely (pelagic) or sink to the botton (demersal), with or without oil globules. Newly hatched larvae are 1.5 to three millimetres long (0.06 to 0.12 inch). Active feeding begins shortly after hatching, and mortality of newly hatched larvae is extremely high. Larvae drift with currents (planktonic) until metamorphosis, or shortly afterward, and then settle to the bottom to assume their adult bottom-living (benthic) existence.

Habitat

Flatfishes lie on the bottom, generally covered by sand or mud, with only their eyes protruding. The eyes can be raised or lowered and moved independently. Flounders feed primarily on crustaceans, other bottom invertebrates, and small fish. When feeding they remain motionless until their prey ventures too close and then literally leap off the bottom in pursuit. Flatfishes in turn fall prey to a variety of large fish and cetaceans (whales, porpoises, etc.), but the primary predator of flatfishes is man.

Swimming is accomplished by undulating movements of body and fins, rather than by sculling with the caudal fin as in most other fish.

Flatfishes are primarily found in temperate and tropical seas, with some species extending northward into the Arctic. Sizes range from about 100 millimetres (four inches) to the large Atlantic halibut, which attains a length of more than two metres (nearly seven feet) and a weight of about 325 kilograms (716 pounds). Most species are marine, but some spend all or part of their life in freshwater. Flatfishes are found in depths up to 1,000 metres (3,300 feet), but most occur on the continental shelf in less than 200 metres of water.

Form and function. Many species display sexual dimorphism, with the male having one or several of the following characteristics: elongated pectoral-fin rays, wider interorbital bones, spines on the head, tentacles on the eyes, more elaborate pigmentation. Flounders have a long dorsal fin extending from the head to caudal (tail) fin and an anal fin extending from vent (anus) to caudal fin in most species. Pectoral fins are present on all larval flatfishes but are lost or reduced in adults of the families Soleidae and Cynoglossidae. Caudal-fin rays and their supporting structures are variable. Scales are ctenoid (rough edged) or cycloid (smooth). Dentition is variable and corresponds to feeding habits of the species. Active predators have large, well-developed teeth in both jaws, whereas those living primarily in mud and feeding on bottom invertebrates have teeth only on the lower jaw of the blind side. Sexes are easily distinguished because the ovaries extend posteriorly from the body cavity beneath the skin and a thin layer of muscle immediately above the muscles of the anal-fin rays. Testes cannot be seen without dissection. The stomach and intestines curl within the body cavity to form a loop.

Migration of the eye during metamorphosis

The main feature of metamorphosis is the migration of the eye around or through the head. This is accomplished as a movement either over the middorsal ridge or through the head, in a depression between the supraorbital bars (over the eye) and ventral edge of the dorsal fin. The supraorbital bars extend forward from the cranium to the ethmoid region of the skull (the area in front of the eye), gradually shifting ventrally and coming to lie next to one another. As the eye migration begins, the dorsal edge of the supraorbital bar is reabsorbed to make room for the eye moving through the head. The supraorbital bars ossify and become the interorbital bone after the eye has completed its migration. The blind- (bottom-) side frontal bone shifts to the ocular (upper) side and forms a portion of the optic-capsule floor for the upper eye. Torsion (twisting) of the frontals, ethmoid, and mouthparts is the essential feature of the flatfish skull.

Normal pigmentation on adult flatfishes consists of a coloured ocular side and an unpigmented (white) blind side. The ocular side is variable in pigment pattern and intensity. Flatfishes can mimic their background by assuming a similar coloration. Partial or complete albinism is known in some species, but a more common colour variation is ambicoloration (coloration on both sides). Ambicoloration can be partial or complete and is often associated with incomplete migration of the eye (in which the migrating eye stops on middorsal ridge) and a hooked appearance, caused by the unattached origin of the dorsal fin. Reversal (eyes and pigmentation on the side that normally is unpigmented) is fairly frequent in some species but quite rare in others.

Classification. Flatfishes are divisible into three suborders and seven families.

ORDER PLEURONECTIFORMES

(Heterosomata of some authors)

Allied to Perciformes but asymmetrical, compressed, both eyes on one side of head; pelvic bones attached directly to cleithrum. Swim bladder absent in adults. Fossil records for this group of fish are limited, extending from Paleocene to Recent, about 65,000,000 years.

Suborder Psettodoidei

The least specialized (most primitive) flatfish. Spines present in dorsal, anal, and pelvic fins; dorsal fin not extending onto head; eyes on either right (dextral) or left (sinistral) side; maxillary (upper-jaw) bone with well-developed supplemental bone; vertebrae 24–25 (10 precaudal, 14–15 caudal).

Family Psettodidae (primitive flatfishes)

Same characters as given for the suborder. Two species, one from Indo-Pacific and one from Africa; attain a length of about 2 ft.

Suborder Pleuronectoidei

No spines in fins (one spine in pelvic fin of Citharidae); dorsal fin extending forward onto head; usually no supplemental bone on maxillary (may be present or absent in Citharidae); vertebrae 27–70 (generally numbering 34 or more); preopercular margin free; lower jaw prominent; nostrils asymmetrical (that on blind side being near edge of head).

Family Citharidae

Eyes either dextral or sinistral; anus on ocular side; gill membranes widely separated; dorsal- and anal-fin rays not shortened posteriorly. Five monotypic genera found in the Indo-Pacific and Mediterranean and off Africa and Japan; attain a length to about 12 in.

Family Scophthalmidae

Eyes sinistral; anus on blind side; gill membrane widely separated; dorsal- and anal-fin rays shortened posteriorly; pelvic-fin bases long (both extending forward onto the urohyal). Four genera and 10 species found in North Atlantic and Mediterranean Sea; attain lengths to about 3 ft. and weights to about 50 lb.

Family Bothidae (left-eyed flounders)

Eyes sinistral; anus generally far up on blind side; gill membranes connected; dorsal- and anal-fin rays shortened posteriorly; pelvic-fin bases on ocular side short, on blind side may be short or long, 6-fin rays in all but 1 species. Thirty genera with about 200 species, found primarily in the tropical and temperate seas of the world; generally small.

Family Pleuronectidae (right-eyed flounders)

Eyes dextral; anus on blind side, commonly on or near midline; gill membranes connected; dorsal- and anal-fin rays shortened posteriorly; pelvic-fin bases of ocular side short or long, on blind side short, 3–13 pelvic-fin rays. Forty-three genera with about 100 species, found primarily in northern and Arctic seas, but some occur in tropical and temperate seas.

Suborder Soleoidei

Same as Pleuronectoidei, but preopercular margin not free (covered by skin and scales); lower jaw not prominent; nostrils nearly symmetrical, that on blind side not near edge of head.

Family Soleidae (soles)

Eyes small, dextral; mouth curved downward; sensory or tactile papilla on head (primarily on blind side); caudal fin with numerous rays. Over 100 species found in tropic and temperate seas; many true soles are small and found along the shore, somes species inhabiting freshwater.

Family Cynoglossidae (tongue soles)

Eyes small, sinistral; head large; mouth curved downward, head lacking sensory or tactile papilla; dorsal and anal fins confluent with caudal fin; caudal fin with only about 12 fin rays; generally small, slender fishes, found in tropical seas.

BIBLIOGRAPHY. J.T. CUNNINGHAM, *A Treatise on the Common Sole,* Solea vulgaris, *Considered Both as an Organism and as a Commodity* (1890), a monograph on the morphology, development, life history, and economics of this species,

including a taxonomic review; C.L. HUBBS, "Phylogenetic Position of the Citharidae, a Family of Flatfish," *Misc. Publs. Mus. Zool. Univ. Mich.*, no. 63, pp. 5–38 (1945), presents a definition of the family Citharidae and phylogenetic relationships of the flatfish families; H.M. KYLE, "Flat-fishes (Heterosomata)," *Rep. Danish Exped. Medit.* 2:1–150 (1913), a major treatise on the development of some flatfishes in the families Bothidae, Soleidae, and Cynoglossidae (illustrations of larvae included when known); and "The Asymmetry, Metamorphosis and Origin of Flatfishes," *Phil. Trans. R. Soc.*, Series B, 211:75–129 (1921), a major paper on the origins of pleuronectiform metamorphosis; J.R. NORMAN, *A Systematic Monograph of the Flatfishes (Heterosomata)*, vol. 1, *Psettodidae, Bothidae, Pleuronectidae* (1934), a classic on the taxonomy of flatfishes, the only paper dealing with these families on a worldwide basis.

(E.J.G.)

Pliny the Elder

Gaius Plinius Secundus is usually called Pliny the Elder to distinguish him from his nephew and biographer, Gaius Plinius Caecilius Secundus. In his energy, public spirit, and culture, the elder Pliny typified upper class Roman society. He influenced the intellectual development of western Europe no less than did Cicero, Horace, and Virgil; unlike Virgil, however, Pliny's influence was not stylistic and literary but scientific and historical. Because of the enormous amount of material he collected, edited, and arranged under the then uncommon title *Historia naturalis* (*Natural History*), Pliny became one of the chief authorities on scientific matters in the Middle Ages. His encyclopaedic coverage and picturesque descriptions far outweigh his lack of critical discrimination and knowledge of technical details. By preserving the best of Greek science, the *Natural History* long remained a source for both correct and incorrect scientific information. The history of early science cannot be understood properly without recognizing Pliny's role, both in establishing a new type of scientific literature—the encyclopaedia—and in supplying the content that would be accepted for over 1,500 years.

Pliny was born in Novum Comum (modern Como), in northern Italy, in AD 23. Descended from a prosperous family, he was enabled to complete his studies in Rome. At the age of 23, he began a military career by serving in Germany, rising to the rank of cavalry commander. He returned to Rome, where he possibly studied law. Until near the end of Nero's reign, when he became procurator in Spain, Pliny lived in semi-retirement, studying and writing. His devotion to his studies and his research technique were described by his nephew. Upon the accession in AD 69 of Vespasian, with whom Pliny had served in Germany, he returned to Rome and assumed various official positions.

Seven writings are ascribed to him, of which only the *Natural History* is extant. There survive, however, a few fragments of his earlier writings on grammar, a biography of Pomponius Secundus, a history of Rome, a study of the Roman campaigns in Germany, and a book on hurling the lance. These writings probably were lost in antiquity and have played no role in perpetuating Pliny's fame, which rests solely on the *Natural History*.

The *Natural History*, divided into 37 *libri*, or "books," was completed, except for finishing touches, in AD 77. In the preface, dedicated to Titus (who became emperor shortly before Pliny's death), Pliny justified the title and explained his purpose on utilitarian grounds as the study of "the nature of things, that is, life" ("Preface," 13). Heretofore, he continued, no one had attempted to bring together the older, scattered material that belonged to "encyclic culture" (*enkyklios paideia*, the origin of the word encyclopaedia). Disdaining high literary style and political mythology, Pliny adopted a plain style—but one with an unusually rich vocabulary—as best suited to his purpose. A novel feature of the *Natural History* is the care taken by Pliny in naming his sources, over 100 of which are mentioned. Book I, in fact, is a summary of the remaining 36 books, listing the authors and sometimes the titles of the books (many of which are now lost) from which Pliny derived his material.

The *Natural History*

The *Natural History* properly begins with Book II, which is devoted to cosmology and astronomy. Here, as elsewhere, Pliny demonstrated the extent of his reading, especially Greek texts, which he read fluently. By the same token, however, he was sometimes careless in translating details, with the result that he distorted the meaning of many technical and mathematical passages. In Books III through VI, on the physical and historical geography of the ancient world, he gave much attention to major cities, some of which no longer exist.

Books VII through XI treat zoology, beginning with man (VII), then mammals and reptiles (VIII); fishes and other marine animals (IX); birds (X); and insects (XI). Pliny derived most of the biological data from Aristotle, while his own contributions were concerned with legendary animals and unsupported folklore.

In Books XII through XIX, on botany, Pliny came closest to making a genuine contribution to science. Although he drew heavily upon Theophrastus, he reported some independent observations, particularly those made during his travels in Germany. Pliny is one of the chief sources of modern knowledge of Roman gardens, early botanical writings, and the introduction into Italy of new horticultural and agricultural species. Book XVIII, on agriculture, is especially important for agricultural techniques such as crop rotation, farm management, and the names of legumes and other crop plants. His description of an ox-driven grain harvester in Gaul, long regarded by scholars as imaginary, was confirmed by the discovery in southern Belgium in 1958 of a 2nd-century stone relief depicting such an implement. Moreover, by recording the Latin synonyms of Greek plant names, he made most of the plants mentioned in earlier Greek writings identifiable.

Books XX through XXXII are concerned with medicine and drugs. Like many Romans, Pliny criticized luxury on moral and medical grounds. His random comments on diet and on the commercial sources and prices of the ingredients of costly drugs provide valuable historical evidence relevant to contemporary Roman life. The subjects of Books XXXIII through XXXVII include minerals, precious stones, and metals, especially those used by Roman craftsmen. In describing their uses, he referred to famous artists and their creations and to Roman architectural styles and technology.

Pliny's last assignment was that of commander of the fleet in the Bay of Naples, where he was charged with the suppression of piracy. Learning of an unusual cloud formation—later found to have resulted from an eruption of Mt. Vesuvius—Pliny went ashore to ascertain the cause and to reassure the terrified citizens. He was overcome by the fumes resulting from the volcanic activity and died on August 24, 79, according to his nephew's report. Pliny was unmarried and was survived by his only sister.

Assessment

In retrospect, Pliny's influence is based on his ability to assemble in a methodical fashion a number of previously unrelated facts; his perceptiveness in recognizing details ignored by others; and his readable stories, with which he linked together both factual and fictional data. Along with unsupported claims, fables, and exaggerations, Pliny's belief in magic and superstition helped to shape scientific and medical theory in subsequent centuries. Perhaps the most important of the pseudoscientific methods advocated by him was the doctrine of signatures: a resemblance between the external appearance of a plant, animal, or mineral and the outward symptoms of a disease was thought to indicate the therapeutic usefulness of the plant. With the decline of the ancient world and the loss of the Greek texts on which Pliny had so heavily depended, the *Natural History* became a substitute for a general education. In the Middle Ages, many of the larger monastic libraries possessed copies of the work; these and many abridged versions ensured Pliny's place in European literature. His authority was unchallenged, partly because of a lack of more reliable information and partly because his assertions were not and, in many cases, could not be tested.

The first attack on Pliny's work—Niccolò Leoniceno's *De erroribus Plinii* ("Concerning the Errors of Pliny")—was published in Ferrara in 1492. Thereafter, Pliny's influence diminished, as more writers dared to question

his statements. By the end of the 17th century, the *Natural History* had been rejected by the leading scientists. Up to that time, however, Pliny's influence, especially on nonscientific writers, was undiminished; he was, for example, almost certainly known to Shakespeare and Milton. Although Pliny's work was never again accepted as an authority in science, 19th-century Latin scholars conclusively demonstrated the historical importance of the *Natural History* as one of the greatest literary monuments of classical antiquity. It is still of value to those who wish an honest résumé of 1st-century Rome.

BIBLIOGRAPHY. *Natural History*, 10 vol., ed. and trans. by H. RACKHAM *et al.* ("Loeb Classical Library," 1938–63), Latin text has facing English translation, which is accurate, readable, with helpful notes and indexes. KONRAT ZIEGLER, WILHELM KROLL *et al.* "C. Plinius Secundus der Ältere," in *Pauly-Wissowa Realencyklopädie*, vol. 21, col. 271–439 (1951), is the best overall treatment. H. LE BONNIEC, *Bibliographie de l' "Histoire naturelle" de Pline l'Ancien* (1946), contains over 900 items, arranged according to the books of the *Natural History*. See also JERRY STANNARD, "Pliny and Roman Botany," *Isis*, 56:420–425 (1965).

(Je.St.)

Plotinus

Plotinus, the great philosopher and religious genius of the 3rd century AD, who transformed the revival of Platonism in the Roman Empire into what modern scholars call Neoplatonism, exercised a wide-ranging direct and indirect influence on European thought that lasted until the late 17th century. Neoplatonism has been enthusiastically studied and has provided inspiration for the thought of particular individuals and groups down to the present time. Neoplatonic philosophy has also deeply influenced the thought of the Islāmic world (see PLATONISM AND NEOPLATONISM).

Porphyry's "Life of Plotinus." The only important source for the life of Plotinus is the biography that his disciple and editor Porphyry wrote as a preface to his edition of the writings of his master, the *Enneads*. Other ancient sources add almost no reliable information to what Porphyry relates. This must be mentioned because, though Porphyry's "Life of Plotinus" is the best source available for the life of any ancient philosopher, it has some important deficiencies that must necessarily be reflected in any modern account of the life of Plotinus that does not use a great deal of creative imagination to fill in the gaps. The "Life" is the work of an honest, accurate, hero-worshipping, and serious-minded friend and admirer. Apart from a few fascinating scraps of information about the earlier parts of the life of Plotinus, Porphyry concentrates on the last six years, when he was with his master in Rome. Thus, a fairly complete picture is available only of the last six years of a man who died at the age of 65. It is the elderly Plotinus, as it is the elderly Socrates, who alone is known.

Early life. Plotinus' own writings contain no autobiographical information, and they can give no unintentional glimpses of his mind or character when he was young; they were all written in the last 15 years of his life. Nothing is known about his intellectual and spiritual development.

Origin. Plotinus was born in the year AD 205. Porphyry states that he never spoke about his parents, his race, or his country. Eunapius, a late-4th-century writer, and later authors wrote that his birthplace was Lyco or Lycopolis in Egypt, either the modern Asyūt in Upper Egypt or a small town in the Nile Delta. Though this may be true, there is no real evidence in the "Life" or in his own writings to suggest that Plotinus had any special knowledge of or affinity with Egypt; the fact that he later studied philosophy in the great cosmopolitan city of Alexandria is not necessarily evidence that he was an Egyptian. His name is Latin in form, but, in the 3rd century AD, this gives no clue to his ethnic origins. All that can be said with reasonable certainty is that Greek was his normal language and that he had a Greek education. For all his originality, he remains Hellenic in his way of thinking and his intellectual and religious loyalties.

Study under Ammonius Saccas. In his 28th year—he seems to have been rather a late developer—Plotinus felt an impulse to study philosophy and thus went to Alexandria. He attended the lectures of the most eminent professors in Alexandria at the time, which reduced him to a state of complete depression. In the end, a friend who understood what he wanted took him to hear the self-taught philosopher Ammonius "Saccas." When he had heard Ammonius speak, Plotinus said, "This is the man I was looking for," and stayed with him for 11 years.

Study of philosophy at Alexandria

Ammonius is the most mysterious figure in the history of ancient philosophy. He was, it seems, a lapsed Christian (yet even this is not quite certain), and the one or two extant remarks about his thought suggest a fairly commonplace sort of pre-Neoplatonic Platonism. A man who could attract such devotion from Plotinus, and who may also have been the philosophical master of the great Christian theologian Origen, must have had something more to offer his pupils; but what it was is not known. That Plotinus stayed with him for 11 years is in no way surprising. One did not enter an ancient philosophical school to take courses and obtain a degree, but rather to join in what might well be a lifelong cooperative following of the way to truth, goodness, and the ultimate liberation of the spirit.

Expedition to the East. At the end of his time with Ammonius, Plotinus joined the expedition of the Roman emperor Gordian III against Persia (242–243), with the intention of trying to learn something at first hand about the philosophies of the Persians and Indians. But the expedition came to a disastrous end in Mesopotamia, when Gordian was murdered by the soldiers and Philip the Arabian was proclaimed emperor. Plotinus escaped with difficulty and made his way back to Antioch. From there he went to Rome, where he settled at the age of 40. That a Greek philosopher, especially at this period, should be interested in Oriental thought is not extraordinary. Plotinus' own thought shows some striking similarities to Indian religious philosophy, but he never actually made contact with Eastern sages because of the failure of the expedition. Though direct or indirect contact with Indians educated in their own religious-philosophical traditions may not have been impossible in 3rd-century Alexandria, the resemblances of the philosophy of Plotinus to Indian thought were more likely a natural development of the Greek tradition that he inherited. That Plotinus was able to join the expedition of the senatorial emperor Gordian, that he went to Rome (an unusual place for a philosopher to settle), and that Porphyry found him, 19 years later, at the centre of a circle of friends and disciples—many of whom were members of the senatorial aristocracy—has been interpreted (probably erroneously) as meaning that he or his family had strong personal connections with Roman senators.

Life as an established philosopher in Rome. Whatever may have been the circumstances of Plotinus when he first came to Rome, by the time Porphyry made his acquaintance in AD 263 he was living in dignified and comfortable conditions, though maintaining a considerable degree of personal austerity. His reputation in society was excellent and earned by practical activity as well as by teaching. He acted as an arbitrator in disputes, without ever being known to make an enemy. And many of his aristocratic friends, when they were approaching death, appointed him guardian of their children. "His house," Porphyry says, "was full of young lads and maidens," and he most conscientiously fulfilled his obligations under Roman law as their guardian, taking care of their education and their property. Like other great contemplatives, he had plenty of time for other people and could attend to their worries (sometimes quite trivial) without losing his inward concentration. He heard a boy's lessons, found who had stolen a lady friend's necklace, or noticed that Porphyry was in a state of depression and contemplating suicide, and so sent him away for a change of scenery and companionship. "Present at once to himself and others" and "gentle and at the disposal of all who had any sort of acquaintance with him" are ways in which Porphyry described him. He was, it seems, a man who gave the im-

Practical and scholarly activities

pression of being in touch with the eternal without losing awareness of the earthly needs of his many friends.

His circle of friends was cosmopolitan, including men from the eastern half of the empire as well as Roman senators, their wives, and widows. Among those who venerated Plotinus, according to Porphyry, were the emperor Gallienus (reigned 253–268) and his wife Salonina, and this led Plotinus on one occasion to attempt practical activity on a larger scale. He asked the Emperor to restore a ruined city in Campania and endow it with the surrounding land; the restored city was to be called Platonopolis, and its citizens were to live according to the laws and customs of Plato's ideal states. Plotinus promised that he would go and live there himself with his friends. That a philosopher who shows in his writings such a total lack of interest in the political side of Plato's thought, and who preached withdrawal from public life, should have made such a proposal is interesting. He may well have thought it his duty as a Platonic philosopher to attempt the foundation of a Platonic city, if opportunity offered —however personally disinclined he might have been to such activity. The Emperor refused his request, and in the political circumstances of the time there was no chance of its being granted. Gallienus and the Senate were not on good terms. He had excluded members of the senatorial order from all military commands, and they took their revenge by successfully blackening his memory after his death. However much he might have respected Plotinus personally, the Emperor would inevitably have regarded Platonopolis as a most undesirable senatorial strongpoint and a centre of intrigue against his authority.

Teaching and writing for his school

The main activity of Plotinus, to which he devoted most of his time and energy, was his teaching and, after his first ten years in Rome, his writing. There was nothing academic or highly organized about his "school," though his method of teaching was rather scholastic. He would have passages read from commentaries on Plato or Aristotle by earlier philosophers and then expound his own views. The meetings, however, were friendly and informal, and Plotinus encouraged unlimited discussion. Difficulties, once raised, had to be discussed until they were solved. The school was a loose circle of friends and admirers with no corporate organization. It was for these friends that he wrote the treatises that Porphyry collected and arranged as the *Enneads*. Some, it seems from their complexity, were destined for an inner circle of his closest friends and philosophical collaborators, such as Porphyry, Amelius Gentilianus from Tuscany (the senior member of the school), and Eustochius, who was Plotinus' physician and who may have produced another edition of his works, now lost.

Some stories in the "Life," and some passages in the *Enneads*, give an idea of Plotinus' attitude to the religions and superstitions of his intensely religious and superstitious age, an attitude that seems to have been unusually detached. Like all men of his time, he believed in magic and in the possibility of foretelling the future by the stars, though he attacked the more bizarre and immoral beliefs of the astrologers. His interest in the occult was philosophical rather than practical, and there is no definite evidence that he practiced magic. A person called Olympius is reported to have once tried to use magic against Plotinus, but he supposedly found that the malignant forces he had evoked were bouncing back from Plotinus to himself. Plotinus was once taken to the Temple of Isis (a Greco-Egyptian mystery goddess) for a conjuration of his guardian spirit; a god, Porphyry stated, appeared instead of an ordinary guardian angel but could not be questioned because of a mishandling of the conjuring process which broke the spell. What Plotinus himself thought of the proceedings is not known, but apparently he was not deeply interested.

Attitude toward pagan cults and Gnostics

His attitude toward the traditional pagan cults was one of respectful indifference. Amelius, his closest friend and co-worker in philosophy, was a pious man, addicted to attendance at sacrifices. Plotinus refused to join him in his devotions but seems to have thought none the worse of him. Despite his rather aggressive piety, Amelius remained Plotinus' friend and collaborator. Some members of his circle of friends were Gnostics (heretical Christian dualists who emphasized esoteric salvatory knowledge), and they provoked him not only to write a vigorous attack on their beliefs but to organize a polemic campaign against them through the activities of Porphyry and Amelius. The reasons of Plotinus for detesting Gnosticism also would have applied, to some extent, to orthodox Christianity—though there is no evidence that he knew anything about it or that he had any contact with the church in Rome. Gnosticism appeared to him to be a barbarous, melodramatic, irrational, immoral, un-Greek, and insanely arrogant superstition. Plotinus' own religion, which he practiced and taught with calm intensity, was the quest for mystical union with the Good through the exercise of the pure intelligence.

Last years. In his last years Plotinus, whose health had never been very good, suffered from a painful and repulsive sickness that Porphyry describes so imprecisely that one modern scholar has identified it as tuberculosis and another as a form of leprosy. This made his friends, as he noticed, avoid his company, and he retired to a country estate belonging to one of them in Campania and within a year died there (270). The circle of friends had already broken up. Plotinus himself had sent Porphyry away to Sicily to recover from his depression. Amelius was in Syria. Only his physician Eustochius arrived in time to be with Plotinus at the end. His last words were either "Try to bring back the god in you to the divine in the All" or "I am trying to bring back the divine in us to the divine in the All." In either case, they express very simply the faith that he shared with all religious philosophers of late antiquity.

BIBLIOGRAPHY. The text of PORPHYRY'S *Life of Plotinus* may be found prefixed to most of the editions of the *Enneads* of Plotinus. *Plotins Schriften*, vol. 5c, ed. and trans. by R. HARDER (1958), is a critical edition of the Greek text with German translation and notes. A Greek text with English translation and short notes is in *Plotinus*, vol. 1, trans. by A.H. ARMSTRONG (1966). An English translation is in *The Enneads*, trans. by STEPHEN MACKENNA, 4th ed. rev. by B.S. PAGE, pp. 1–20 (1969). The best character study of Plotinus, and also an excellent introduction to his thought, is PIERRE HADOT, *Plotin, ou la simplicité du regard* (1963). See also A.H. ARMSTRONG in the *Cambridge History of Later Greek and Early Mediaeval Philosophy*, pt. 3, "Plotinus," ch. 12, "Life," and 13, "Teaching and Writing" (1967), with bibliography.

(A.H.A.)

Plumbing

In general, plumbing refers to the system of pipes, fixtures, and other apparatus used inside a building for supplying water and removing liquid and waterborne wastes. In practice, the term includes storm water or roof drainage and exterior system components connecting to a source such as a public water system or a point of disposal such as a public sewer system or a domestic septic tank or cesspool.

The purpose of plumbing systems is, basically, to bring into, and distribute within, a building a supply of safe water to be used for drinking purposes and to collect and dispose of polluted and contaminated wastewater from the various receptacles on the premises without hazard to the health of occupants. Codes, regulations, and trade practices are designed to keep the water system separated from drainage systems; to prevent the introduction of harmful material such as chemicals, micro-organisms, and dirt; and to keep the water system safe under all operating conditions. These protective codes also are designed to prevent flooding of drainage lines, provide venting of dangerous gases, and eliminate opportunities for backflow of dangerous waste water into the water system. It is essential that disease-producing organisms and harmful chemicals be confined to the drainage system.

Since the time of Moses man has been cautioned to dispose of his wastes safely, and cleanliness has been related to the availability of water and associated with social custom. Early man often lived near a water source that served as his water supply and drainage system in one. It was also his bath. Latrine-like receptacles with

Water system of King Minos

crude drains have been found in excavations in the Orkney Islands of Neolithic stone huts at least 10,000 years old. Both a water system and piping used as drainage fashioned of terra-cotta pipe were part of the royal palace of Minos in Crete, about 2000 BC. The palace also had a latrine with water-flushing reservoir and drainage. Nothing comparable to it was developed in Europe until the 18th century. Even the equipment of the modern bathroom, though much improved with hot and cold water under pressure and less crude provisions for drainage, is in concept little different from the Minoan version. It was not until the end of the 19th century that advances in plumbing practice were given serious attention as an integral part of housing.

A building plumbing system includes two components, the piping that brings potable water into the building and distributes it to all fixtures and water outlets and the piping that collects the water after use and drains it to a point of safe disposal.

Water systems. When a building is served by a public water system, the plumbing begins at the service connection with the public supply. It includes all meters, pumps, valves, piping, storage tanks, and connections required to make water available at outlets serving the fixtures or equipment within the building.

Many premises in rural areas are not served by public water supply. These may include private dwellings, apartment houses, hotels, commercial centres, hospitals, institutions, factories, roadside stands, and restaurants.

Public water supplies have surface water or groundwater as their source. Large water systems are almost entirely supplied with surface water (see WATER-SUPPLY SYSTEMS). In smaller communities and in certain areas groundwater is obtained from wells or springs. Independent semipublic, industrial, and private-premise water systems frequently take water from wells on the premises but may, under certain conditions, draw water from a spring, lake, or stream.

Public water systems supply treated water meeting public water-supply drinking-water standards. Private-premise systems are expected to provide water of equal quality, and to do so the private system requires a water-treatment plant including chlorination as a minimum and possibly sedimentation (settling out of solid particles) chemical treatment, primarily for softening, and filtration.

Pressure pumps

Water is supplied to fixtures and outlets under pressure provided by pumps or elevated storage tanks or both. In some installations a pump controlled by a pressure-activated switch on a pressurized storage tank takes water from a well and pumps until the upper limit of pressure for the system has been reached. If water is being used at the rate it is being pumped, the pump operates continuously. Elevated storage tanks are usually equipped with high- and low-level-float control switches to activate the pump. When the tank gets low the pump starts and continues pumping until the tank is full. A storage tank may be constructed as illustrated in Figure 1, or it may be located

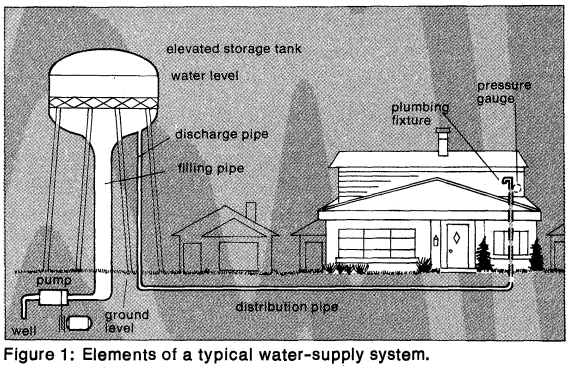


Figure 1: Elements of a typical water-supply system.

on the roof of a high building. Water from the tank feeds the distribution system by gravity. Water flowing through pipes causes a loss of head due to friction. Since building piping systems are designed to deliver water at the required outlet pressure, pipe size is a critical variable.

Plumbing codes have tables and graphs to show typical water demands of fixtures and outlets in a building. If the required water demand is not met because of undersized piping or underpowered pumping, the pressure drops and some outlets may have little or no water flow. Plumbing codes usually specify pressures and rates of flow for the fixtures in a building. The total amount of water that may be needed to supply the demand can be calculated from tables of fixture water demand. Minimum pipe sizes for different fixtures in a building are specified in plumbing codes.

Since it is uneconomical to design a water piping system that would provide flow with all outlets open simultaneously, judgment and experience are used to determine the probable maximum simultaneous demand.

Average daily water requirements vary according to the type of premises being served. A single-family dwelling unit averages from 20 to 100 gallons (80 to 400 litres) per capita per day. Apartment house occupants use less. Special users such as hospitals and industries usually require far greater allowances.

Fixtures and equipment. Most bathrooms of dwelling units comprise lavatory, water closet, and tub or shower (or both). These may be varied by including two lavatory fixtures and a stall shower. In some countries a bidet is also included. Washing and bathing units are normally supplied with both hot and cold water taps. Designs vary; some have separate faucets, while others have mixer valves or common outlets to regulate water temperature.

Kitchen plumbing

Normal kitchen plumbing provides for a single or double sink. Dishwashers, permanently installed or mobile, are finding greater use. In many communities where the drainage is connected to a separated, or sanitary, sewer system, waste food grinders may be installed under the kitchen sink. In small apartments provision may also be made for a mobile laundry facility with an automatic washing machine. Water is taken from sink taps, and wastewater is discharged to the sink through flexible hoses.

The laundry or utility room is usually equipped with a double laundry tub with hot and cold water taps. Clothes washing machines and vented dryers are also part of the laundry equipment. Flexible drains discharge water from the washer to the sink. Metal venting to the outside is usually required on dryers, particularly gas-fired models, but is desirable for all to avoid polluting room air with minute pieces of lint not caught on dryer screens.

Larger apartment houses, hospitals, commercial buildings, and institutions may be equipped with utility sinks in custodian's closets.

Water-system piping and fittings are made of a variety of materials, including copper, brass, bronze, plastic, galvanized iron, wrought iron, malleable iron, and cast iron. Exterior pipe of larger capacity may be constructed with asbestos cement, steel, or plastic materials. In all plumbing it is necessary to have a variety of fittings such as couplings, elbows (angular pipe fittings), wyes (pipe fittings shaped like the letter Y), tees (pipe fittings shaped like the letter T), and caps in order to transport the liquids within piping that must be concealed in walls and floors and provide service to fixtures. Pressure relief valves are required on all hot water tanks as a safety measure. Fittings are made with screw thread, flange (fitted rims), bell and spigot (formed by the insertion of the spigot or male end of one pipe into the bell or female end of another), and slip joint (a telescopic joint). Connections are required to be gastight and watertight for pressures specified for testing purposes by plumbing codes.

Special equipment and safety problems. Hospitals are designed with special fixtures for washing, scrubbing, surgical cleaning, instrument sterilization, housekeeping, disposal of dangerous liquids and solids, and special laundry facilities; kitchen equipment such as vegetable scrubbers, potato peelers, waste food grinders; and the various pieces of water-using equipment found in laboratories such as aspirators, chemical sinks, and autoclaves (pressure cookers).

Industrial plants have individual requirements for plumbing that are contingent on the sanitary facilities for

employees and the process needs of the industry. The sanitary needs are related to ordinary washing facilities and special protection requirements such as eyewash fountains and deluge showers for personnel, clothes, and equipment. Although processing operations involve plumbing, these are generally treated separately. The building process waste piping is, however, part of the plumbing system.

Water-softening systems

In regions where the water is hard and in special quality requirement situations in laboratories and industrial plants, water-softening systems may be installed. Such systems remove calcium and magnesium ions from the water. The typical unit of the past has been a chemical exchange unit in which sodium ions replace calcium ions (base exchange process). Other units employ resins and a few use filter membranes. Softening units require maintenance and either replacement or recharge of the chemical. Care must be taken to avoid contamination of the water system while maintaining the unit.

Sanitary hazards. These are of three kinds: those that endanger the water system, those that may pollute or contaminate the closed space (room) air environment, and those that expose premises to unclean water through leakage or flooding.

The most general and overall danger is related to the water supply, and the most common means of affecting the potability of the water system is by backflow or back siphonage. Liquids under pressure flow from a zone of higher pressure to a zone of lower pressure. Dangerous backflow is produced when a contaminated or polluted liquid under a higher pressure is connected to a potable water system under lower pressure. Back siphonage merely extends this principle to a condition in which the pressure is negative (*i.e.*, a vacuum condition) on the low-pressure side of the connection. Backflow connections endanger not only the building system but may affect the public water supply. If an industrial-fire water system of nonpotable quality pumping at 200 pounds per square inch (14 kilograms per square centimetre) is connected to a public water system operating at 100 pounds per square inch (seven kilograms per square centimetre), for example, the nonpotable fire water supply will flow to the public waterlines. Another type of backflow can be created in the event of a fire if fire engine pumpers draw more water from the mains than can be supplied to them, thus putting a negative pressure on the mains. In this case, water will flow from buildings adjoining the pumping centre and drain the building system by placing a negative pressure on lines in the building. Any connection within this building such as a water-softening unit, corrosion-control unit, laboratory sink, custodial sink, or toilet with submerged water connections will then flow to the building system and possibly to the public system, thereby contaminating both the building water supply and the public water supply.

Various devices are placed in plumbing systems to prevent backflow. Air gaps with a free air opening distance of two to three times the diameter of the fixture or pipe outlet are the only positive means of separating a nonpotable liquid from the water system. Other means of preventing backflow include manually supervised devices consisting of gate valves, check valves, and pressure indicators designed to prevent flow from a zone of higher pressure to one of lower pressure; vacuum breakers, providing a direct opening to air when vacuum or negative pressure occurs in the waterline; and swing connections that must be manually connected to a line to permit water flow.

Gases. In drainage systems, gases result from biological action on organic matter contained in wastes. Evolution gases, generated by chemical reactions, may be found in laboratories of industry, hospitals, and institutions. Sewer gases are present in the house drainage system as well as in the public sewer. Gases from the public sewer are prevented from entering the building drainage system by traps installed at or near the lower end of the building drainage system. House drainage gases are vented through pipes or vent stacks leading from the various fixture drainage lines to the outside air as in Figure 2.

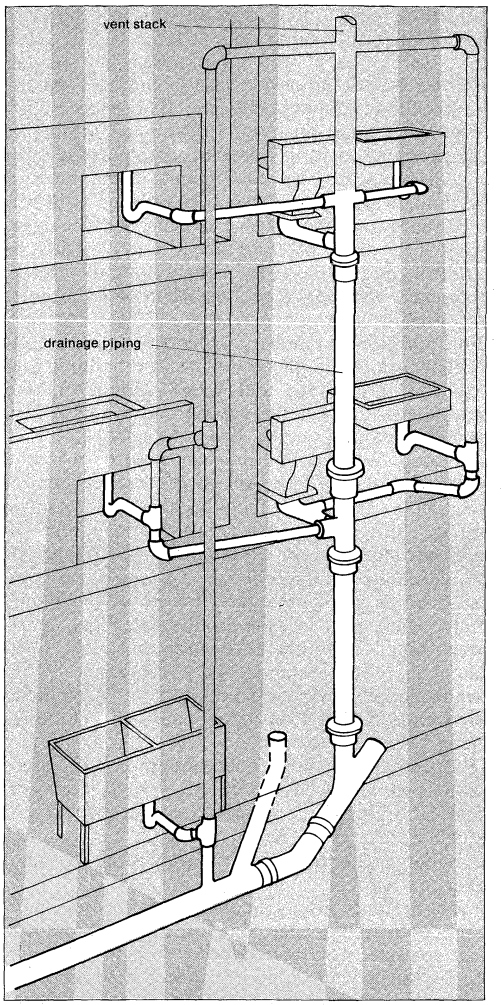


Figure 2: Building drainage and venting system.

Venting gases

Plumbing codes provide for uniform venting of residence fixtures through stacks ranging from two to four inches (five to ten centimetres) in diameter. In complicated and extensive plumbing systems, relief vents are provided to insure air circulation and keep the drainage system at atmospheric pressure. Vents are never less than one-half the diameter of the drain line being vented and at least 1.25 inches (3.3 centimetres) in diameter for individual fixture venting. Vent sizes vary with the length of vent and the fixture units connected. Plumbing codes contain tables from which the size of vent relative to the number of fixtures, size of waste drain, and length of vent may be

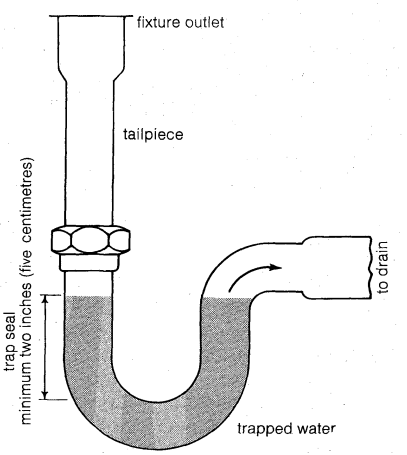


Figure 3: Sanitary fixture trap showing water column that seals the fixture from sewer gases, bacteria, and vermin.

determined. All fixtures require traps of the type shown in Figure 3. The trap places a water seal between the fixture and the drain line so that gases cannot escape into room air. The venting system (Figure 2) prevents a siphon action that would remove the seal. Though stoppage in a drain line at some lower point in the system causes liquids to back up into a fixture, the liquid seal is not broken. Refrigerator drains, commercial food-preparation equipment, and any other equipment in which backup of waste water could cause serious contamination are connected to the drainage system through an indirect waste pipe with adequate air-gap protection.

Under unusual conditions, laboratories may discharge flammable or toxic materials to the drainage system. In this case, precautions are taken to detect leaks or the presence of gases with special instruments. Under normal circumstances regular checks for leakage are performed as required by codes. Occasionally the gases discharged by vents may pollute the outside or ambient air unless special air-cleaning devices are installed on vents (see also POLLUTION CONTROL).

Plumber's equipment. Plumber's equipment is designed to handle the construction and repair of both water and drainage piping and installation of fixtures. A complement of wrenches, including adjustable crescent, monkey, chain, and Stillson, or pipe, wrenches is required. The plumber also needs gasoline or propane fuel torches, hacksaws, tubing and pipe cutters, reamers, chisels, files, pipe threaders, force cups for clearing clogged pipes or facilities ("plumber's helper"), soldering equipment, caulking tools, caulking materials (oakum, lead, and special compounds), fluxes, sealers, valve packing, washers, gaskets, and pressure and vacuum gauges (see also HAND TOOLS). For drain maintenance the plumber must have flexible wire or rods called "snakes" to go through traps and bends. Small snakes are used by hand while large ones are operated by electric or gasoline-powered equipment that rotates the flexible rod and forces it into the pipe obstruction. Special cutters may be attached to this equipment to cut plugging materials such as roots from house sewers.

Drainage systems. Drainage of residential buildings includes the collection of sanitary wastes and roof drainage (see also SEWAGE SYSTEMS; POLLUTION CONTROL). The sanitary wastes are collected in soil pipes and stacks usually made of cast iron, although certain portions of connecting pipe may be galvanized iron. When corrosive liquids are collected, as in laboratories and industrial plants, the system may include plastic or glass pipes or coated piping. The joints, bends, tees, elbows, wyes, and many other special fittings are designed to carry away wastes without having obstructions in the pipe or creating conditions that will cause clogging if some large object is dropped into a fixture or a receptacle. Cleanouts and receptacle outlets are provided with screens or gratings to prevent the entry of clogging materials. The entire piping system is sized so that the smallest size is at the fixture outlet. Plumbing codes specify the minimum sizes for drainage connections and the standards applicable to all pipe and fitting materials.

Normally all building drainage is constructed so that waste water flows by gravity to the main house drain. The house drain is usually connected outside the building to a gravity flow house sewer that leads to a public sewer or a point of treatment. In large buildings such as apartment houses, commercial buildings, hospitals, and industrial plants, the house drain may be lower than the point of discharge. Wastes are then discharged to a sump, or storage pit, and all of the building drainage is lifted by pumps or pneumatic ejectors to a point of discharge to the exterior sewer system. These lifting devices empty water from the storage sump on a cycling basis activated by float controls that prevent flooding of the storage area. Such units are usually installed in duplicate and often have alternate sources of power such as a diesel- or gasoline-powered generator for emergency pumping.

Roof drainage is collected in gutters and leaders and taken by appropriate piping to a point of discharge permitted by law. Isolated dwellings may drain to surrounding ground, while larger buildings have a drainage system similar to the sanitary system that connects into a public storm-water sewer system.

Home disposal systems

Home disposal systems are used in rural areas. The house drain is connected to a septic tank with a tile drainage field or to a cesspool. The septic tank removes heavy solid materials from the waste, and the effluent or treated water is allowed to percolate into the soil through buried, specially constructed, rock-filled trenches over which tiles with open joints are laid. Enough trench must be constructed to allow percolation without flooding the surface of the ground. The effluent from septic tanks contains disease-causing bacteria and cannot be allowed to flow directly into streams or underground waters. Health codes and regulations specify the sizes of home disposal units and control the discharge of effluent.

Premises often have other water uses including swimming pools (both outdoor and indoor), ornamental pools, fish pools, and fountains. These require water and are part of the plumbing system. Since swimming-pool water is easily contaminated by bathers, it must either be replaced frequently or filtered, chlorinated, and recirculated.

Pools constructed for private family use are subject to regulation to prevent backflow of pool water into the building water system. Hotels and motels frequently have swimming pools; provisions for handling the water and waste water are made in designing the plumbing system.

The practice of installing permanent lawn-sprinkling systems has grown. A piping system of plastic or metal is placed beneath the lawn and spray headers that pop up when water pressure is applied are placed according to the desired sprinkling pattern. As the sod forms, pockets around these headers hold water and provide a submerged connection that might, under backflow conditions, carry contaminated water back into the building supply system. Such systems need a vacuum breaker or other means of backflow prevention.

A mobile home—a vehicular, portable structure built on a chassis and therefore without a permanent foundation—must be provided with and connected to suitable utility services when in use as a dwelling. Conditions of transport and use are such that plumbing standards for permanent buildings are not always applicable. The same principles of plumbing practice for safety and protection must be preserved, however. Special codes have been developed for such homes.

Plumbing modules

Another variant in buildings is the prefabricated structure or prefabricated plumbing module. In such structures units such as wall panels are made at a factory and shipped to a site for erection. The plumbing (both water and drainage) is systematized, and modules with appropriate matched connection points in panels and floor sections and the like can be connected quickly on site. Prefabricated plumbing systems must comply with building and plumbing codes and must be tested after erection to insure against leakage.

BIBLIOGRAPHY. L. WRIGHT, *Clean and Decent* (1960), traces the bathroom and its development from Neolithic times to the 20th century. A.J. MATTHIAS, JR., *How to Design and Install Plumbing*, 3rd ed. rev. by E. SMITH, provides a practical treatise on the parts of a plumbing system and how they are put together. V.M. EHLERS and E.W. STEEL, "Plumbing," in *Municipal and Rural Sanitation*, 6th ed., ch. 11 (1965), offers a general exposition on water supply and wastewater systems and on public health hazards for both rural and urban conditions. AMERICAN NATIONAL STANDARDS INSTITUTE, *National Plumbing Code, ASA-40.8* (1955), is the standard reference on plumbing regulations and standards and has been adopted by many agencies for control purposes. PUBLIC HEALTH SERVICE TECHNICAL COMMITTEE ON PLUMBING STANDARDS, *Proposed Revision of ASA-40.8-1955* (1962), is often used as a supplement to the ANSI code. AMERICAN PUBLIC HEALTH ASSOCIATION, AMERICAN SOCIETY CIVIL ENGINEERS, WATER POLLUTION CONTROL FEDERATION JOINT EDITORIAL BOARD, *Glossary, Water and Wastewater Control Engineering* (1969), is a standard reference and a compendium of terms used in the field including terms related to plumbing, house drainage, water systems, and wastewater systems.

(W.T.I.)

Plutarch

Between the 16th century and the early 19th one of the most popular of classical authors in Europe was Plutarch, a 1st-century Greek biographer and miscellaneous writer. Admired for their wisdom as well as for their information, his writings were long used as source books for anecdotes and moral *exempla;* they influenced the origins and development of the essay, the biography, and the writing of history; and it was from his *Parallel Lives* that the generally accepted images of the great historical figures of Greece and Rome were derived.

Plutarch was born *c.* AD 46 at Chaeronea in Boeotia, an ancient district in east central Greece. He was the son of Aristobulus, himself a biographer and philosopher. In 66–67, Plutarch was studying mathematics and philosophy at Athens under the Peripatetic philosopher Ammonius. Public duties later took him several times to Rome, where he lectured on philosophy, made many friends, and perhaps enjoyed the acquaintance of emperors Trajan and Hadrian. According to the Suda lexicon (a Greek dictionary dating *c.* AD 1000), Trajan bestowed the high rank of an ex-consul upon him. Although this may be true, a report of a 4th-century church historian, Eusebius, that Hadrian made Plutarch governor of Greece is probably apocryphal. A Delphic inscription reveals that he possessed Roman citizenship; his *nomen* or family name, Mestrius, was no doubt adopted from his friend Lucius Mestrius Florus, a Roman consul.

Plutarch travelled widely, visiting central Greece, Sparta, Corinth, Patrae (Patras), Sardis, and Alexandria, but he made his normal residence at Chaeronea, where he held the chief magistracy and other municipal posts and directed a school with a wide curriculum in which philosophy, especially ethics, occupied the central place. He maintained close links with the Academy at Athens (he possessed Athenian citizenship) and with Delphi, where, from *c.* 95, he held a priesthood for life; he may have won Trajan's interest and support for the then renewed vogue of the oracle. The size of Plutarch's family is uncertain. In the *Consolatio* to his wife Timoxena on the death of their infant daughter he mentions four sons; of these at least two survived childhood, and he may have had other children. He died sometime after 119.

Plutarch's literary output was immense. The 227 titles in the so-called catalog of Lamprias, a list of Plutarch's works supposedly made by his son, are not all authentic, but neither do they include all he wrote. The order of composition cannot be determined.

Lives. Plutarch's popularity rests primarily on his *Bioi parallēloi (Parallel Lives)* of Greek and Roman soldiers, legislators, orators, and statesmen. These, dedicated to Trajan's friend, Sosius Senecio, who is mentioned in the lives "Demosthenes," "Theseus," and "Dion," were designed to encourage mutual respect between Greeks and Romans. By exhibiting noble deeds and characters they were also to provide model patterns of behaviour.

Plan of the *Lives*

The first pair, "Epaminondas and Scipio," and perhaps an introduction and formal dedication, are lost. But Plutarch's plan was clearly to publish in successive books biographies of Greek and Roman heroes in pairs, chosen as far as possible for their similarity of character or career, and each followed by a formal comparison. Internal evidence suggests that the *Lives* were composed in Plutarch's later years, but the order of composition can be only partially determined; the present order is a later rearrangement based largely on the chronology of the Greek subjects, who are placed first in each pair. In all, 22 pairs survive (one pair being a double group of "Agis and Cleomenes" and the "Gracchi") and four single biographies, of Artaxerxes, Aratus, Galba, and Otho.

The *Lives* display impressive learning and research. Many sources are quoted, and, though Plutarch probably had not consulted all these at first hand, his investigations were clearly extensive, and compilation must have occupied many years. For the Roman *Lives* he was handicapped by an imperfect knowledge of Latin, which he had learned late in life; for, as he explains in "Demosthenes," political tasks and the teaching of philosophy fully engaged him during his stay in Rome and Italy. The form of the *Lives* represented a new achievement, not closely linked with either previous biography or Hellenistic history. The general scheme was to give the birth, youth and character, achievements, and circumstances of death, interspersed with frequent ethical reflections; but the details varied with both the subject and the available sources, which include anecdote mongers and writers of memoirs as well as historians. Plutarch never claimed to be writing history, which he distinguished from biography. His aim was to delight and edify the reader, and he did not conceal his own sympathies, which were especially evident in his warm admiration for the words and deeds of Spartan kings and generals; his virulent and unfair attack on Herodotus, the great Greek historian of the 5th century BC, probably sprang from his feeling that he had done Athens more and Boeotia less than justice.

Attack on Herodotus

Moralia. Plutarch's surviving writings on ethical, religious, physical, political, and literary topics are collectively known as *Ethica* (Latin *Moralia*), and amount to more than 60 essays cast mainly in the form of dialogues or diatribes. The former vary from a collection of set speeches to informal conversation pieces set among members of Plutarch's family circle; the date and dramatic occasion are rarely indicated. The diatribes, which often show the influence of seriocomic writings of the 3rd-century BC satirist Menippus, are simple and vigorous. The literary value of both is enhanced by the frequent quotation of Greek poems, especially verses of Euripides and other dramatists.

Two educational works, "How a young man ought to hear poetry" (which moderated an old Platonic objection to poetry) and "How to listen," together with a pseudo-Plutarchean treatise, "On the education of children," were popular and influential during the Renaissance. In the treatise on moral virtue Plutarch considered how virtue must subordinate unreason to reason within the soul, a theme developed in many other of the works dealing with popular ethical problems; these adduced examples from the lives of famous men, and contain sound but unoriginal moralizing. Among them are "Vice and Virtue," "How to recognize progress in virtue" (dedicated, like the *Lives*, to Sosius Senecio), "How to distinguish a flatterer from a friend," "On having many friends," and "On Fortune." Another group includes the highly rhetorical historical essays "On the fortune of Alexander," "On the fortune of the Romans," and "Whether the Athenians were more famous in war or in wisdom"; these resemble the traditional topics of declamation, and parallel to them are "Whether water or fire is more useful," "Whether virtue can be taught," and "Whether mental or bodily afflictions are the worse."

Plutarch's interest in animals and their minds emerged in two books "On eating flesh," in "Whether land or sea animals are more intelligent" (a question which receives no clear answer), and "Gryllus" or "Do animals reason?" an entertaining dialogue set on Circe's island in which a pig, one of Odysseus' transformed companions, attacks the Stoic argument denying reason to animals and convinces Odysseus of the moral superiority of many animals over man. The tenets of the philosophical schools are the subject of several essays, including "Platonic questions," "On the creation of the soul in the Timaeus" (expounding Plutarch's views about Plato's teaching on the soul), "Against Colotes" (attacking Epicurean views), and "On the impossibility of living pleasurably according to Epicurus' teaching"; several other essays criticize Stoic doctrines. Physical and medical problems are discussed in "Precepts on health," "On the face of the moon's disk," and "On primary cold" (which argues that cold is something real, not the mere absence of warmth).

The treatises dealing with political issues are of especial interest. "Political precepts" is an enlightening account of political life in contemporary Greece; in "Whether a man should engage in politics when old" Plutarch urged his friend Euphanes to continue in public life at Athens; Stoic ideas appear in the short work "To the unlearned ruler" and the fragmentary argument that "The philoso-

pher should converse especially with princes"; in "The one, the many, and the few in government" the author (who may not be Plutarch) favours monarchy. The virtues of family life are treated in "On brotherly love," "On the love of one's offspring," and "Conjugal precepts"; with the *Consolatio* to his wife goes the fine essay "On exile"; the *Amatorius* is a consideration of love, which favours normal heterosexual relationships.

Religious interests

Plutarch's interest in religious history and antiquarian problems can be seen in a group of striking essays, the early "Daemon of Socrates," and three later works concerning Delphi, "On the failure of the oracles," in which the decline of the oracle is linked with the decline in the population, "On the E at Delphi," interpreting the word EI at the temple entrance, and "On the Pythian responses," seeking to re-establish belief in the oracle. Contemporary with these is "On Isis and Osiris," with its mystical tones; the dialogue "On the late vengeance of God," staged in Delphi, argued that the postponement of divine vengeance is to give time for repentance. "Convivial questions" (nine books) and "Greek and Roman questions" assembled a vast collection of antiquarian lore; "On the malignity of Herodotus" displayed the local patriotism of a Boeotian; and the "Comparison of Aristophanes and Menander" prefers Menander for moral reasons.

Among the more important works no longer accepted as authentic are the *Consolatio* to Apollonius for his son, the "Lives of the Ten Orators," "On Fate," the "Short sayings of kings and commanders," the "Short sayings of Spartans," and "Proverbs of the Alexandrines."

Assessment. Plutarch's perennial charm and popularity arise in part from his treatment of specific human problems in which he avoids raising disquieting solutions. He wrote easily and superficially, with a wealth of anecdote. His style is predominantly Attic, though influenced by the contemporary Greek that he spoke; he followed rhetorical theory in avoiding hiatus between words and was careful in his use of prose rhythms. He is clear, but rather diffuse. Plutarch's philosophy was eclectic, with borrowings from the Stoics, Pythagoreans, and Peripatetics (but not the Epicureans) grouped around a core of Platonism. His main interest was in ethics, though he developed a mystical side, especially in his later years; he reveals that he had been initiated into the mysteries of the cult of Dionysus, and both as a Platonist and as an initiate he believed in the immortality of the soul. He believed too in the superiority of Greek culture and in the meritoriousness and providential character of the Roman Empire. Personally, he preferred a quiet and humane civic life as a citizen of a small Boeotian town, where his writing and teaching did much to illuminate the darkness of provincial life in 1st-century Greece.

Reputation and influence. Plutarch's later influence has been profound. He was loved and respected in his own time and in later antiquity; his *Lives* inspired a rhetorician, Aristides, and a historian, Arrian, to similar comparisons, and a copy accompanied the emperor Marcus Aurelius when he took the field against the Marcomanni. Gradually, Plutarch's reputation faded in the Latin West, but he continued to influence philosophers and scholars in the Greek East, where his works came to constitute a school book. Proclus, Porphyry, and the emperor Julian all quote him, and the Greek Church fathers, Clement of Alexandria and Basil the Great, imitate him without acknowledgment. His works were familiar to all cultivated Byzantines, who set no barrier between the pagan past and the Christian present. It was mainly the *Moralia* which appealed to them; but in the 9th century the Byzantine scholar and patriarch Photius read the *Lives* with his friends.

Plutarch's works were introduced to Italy by Byzantine scholars along with the revival of classical learning in the 15th century, and Italian Humanists had already translated them into Latin and Italian before 1509, when the *Moralia*, the first of his works to be printed in the original Greek, appeared at Venice published by the celebrated Aldine Press. The first original Greek text of the *Lives* was printed at Florence in 1517, and by the Aldine Press in 1519. The *Lives* were translated into French in 1559 by Jacques Amyot, a French bishop and classical scholar, who also translated the *Moralia* (1572). The first complete edition of the Greek texts by a French Humanist in 1572 marked a great improvement in the text.

That Rabelais knew Plutarch well is proved by the frequency with which he quotes from both the *Lives* and the *Moralia*. It was Montaigne, however, who read Plutarch in Amyot's version, who first made his influence widely felt. The style of Montaigne's *Essays* (1580–88) owed much to the *Moralia*, and from the *Lives* he adopted Plutarch's method of revealing character by illustrative anecdote and comment, which he applied to self-revelation. Moreover, the *Essays* made known the ideal, derived from Plutarch's presentation of character and openly expressed opinion, of "high antique virtue and the heroically moral man" which became the Humanist ideal of the Renaissance period.

The *Lives* were translated into English, from Amyot's version, by Sir Thomas North in 1579. His vigorous idiomatic style made his *Lives of the Noble Grecians and Romans* an English classic, and it remained the standard translation for more than a century. Even when superseded by more accurate translations, it continued to be read as an example of Elizabethan prose style. North's Plutarch was Shakespeare's source for his Roman plays and influenced the development of his conception of the tragic hero. The literary quality of North's version may be judged from the fact that Shakespeare lifted whole passages from it with only minor changes.

Influence on Shakespeare

In 1603 the complete *Moralia* was first translated into English directly from the Greek. Its influence can be seen in the 1612 edition of Bacon's *Essays*, which contain counsels of public morality and private virtue, recognizably derived from Plutarch. Francis Bacon was more attracted by Plutarch the moralist than by Plutarch the teller of stories or painter of character, but to the Renaissance mind it was the blend of these elements that gave him his particular appeal. His liking for historical gossip, for the anecdote, and the moral tale, his portrayal of characters as patterns of virtue or vice (in the manner of the morality play and the "character"), his emphasis on the turn of fortune's wheel in causing the downfall of the great, all suited the mood of the age, and from him was derived the Renaissance conception of the heroic and of the "rational" moral philosophy of the ancients.

Historians and biographers in the 16th and 17th centuries followed Plutarch in treating character on ethical principles. The 17th-century English biographer Izaak Walton knew Plutarch well, and his own *Lives* (collected 1670, 1675) imitated him in dwelling on the strength, rather than the weakness, of his subjects' characters.

Influence in 17th and 18th centuries

Plutarch was read throughout the 17th and 18th centuries. The English poet and dramatist John Dryden edited a new translation of the *Lives* first published in 1683–86, and abridged editions appeared in 1710, 1713, and 1718. The *Moralia* was retranslated in 1683–90 and also frequently reprinted. In France, Amyot's translations were still being reprinted in the early 19th century, and their influence on the development of French classical tragedy equalled that of North's version on Shakespeare. Admiration for those heroes of Plutarch who overthrew tyrants, and respect for his moral values, inspired the leaders of the French Revolution; Charlotte Corday, who assassinated the revolutionary leader Jean Paul Marat, spent the day before that event in reading Plutarch.

In Germany, the first collected edition of Plutarch's works was published in 1774–82. The *Moralia* was edited by Daniel Wyttenbach in 1796–1834, and was first translated in 1783–1800. The *Lives*, first edited in 1873–75, had already been translated in 1799–1806. The German classical poets—Goethe, Schiller, and Jean Paul (Johann Paul Richter) especially—were influenced by his works, and he was read also by Beethoven and Nietzsche.

In the 19th century, Plutarch's direct influence began to decline, in part as a result of the reaction against the French Revolution, in part because the rise of the Romantic movement introduced new values and emphasized the free play of passions rather than their control, in part because the more critical attitude of scholars to historical

accuracy drew attention to the bias of his presentation of fact. He was still admired, however, notably by the U.S. poet, philosopher and essayist Ralph Waldo Emerson, and although in the 20th century his direct influence was small, the popular ideas of Greek and Roman history continued to be those derived from his pages.

BIBLIOGRAPHY. R.H. BARROW, *Plutarch and His Times* (1967), is a good general introduction that lists further reading. An excellent study (in German), with comprehensive bibliography, is K. ZIEGLER in *Pauly-Wissowa's Real-encyclopädie der classischen Altertumswissenschaft*, vol. 41, col. 636–962 (1951). The dating of the works is discussed by C.P. JONES in "Towards a Chronology of Plutarch's Works," *Journal of Roman Studies*, 56:61–74 (1966). T.J.B. SPENCER, *Shakespeare's Plutarch* (1964), prints North's *Lives* of Caesar, Brutus, Antony, and Coriolanus, with the relevant parts of Shakespeare.

Editions and translations: (*Lives*): *Vitae Parallelae* by C. LINDSKOG and K. ZIEGLER in the "Teubner Series" (1914–39), is being revised (1957–); *Plutarch's Lives* by B. PERRIN in the Loeb edition, 11 vol. (1914–26), gives a text and translation. (*Moralia*): *Moralia* by G. BERNADAKIS in the "Teubner Series," 7 vol. (1888–96), is being revised by various authors and is nearly complete (1925–). *Plutarch's Moralia* in the Loeb edition, with text and translation, is also being revised by several authors, 16 vol. (1928–).

(F.W.W.)

Pluto

Pluto, symbol ♇ in astronomy, is the ninth planet in the solar system. Long considered the solar system's most distant planet, Pluto is now believed to have an even more distant neighbour, known as Planet X. The principal orbit and physical data are shown in the accompanying table.

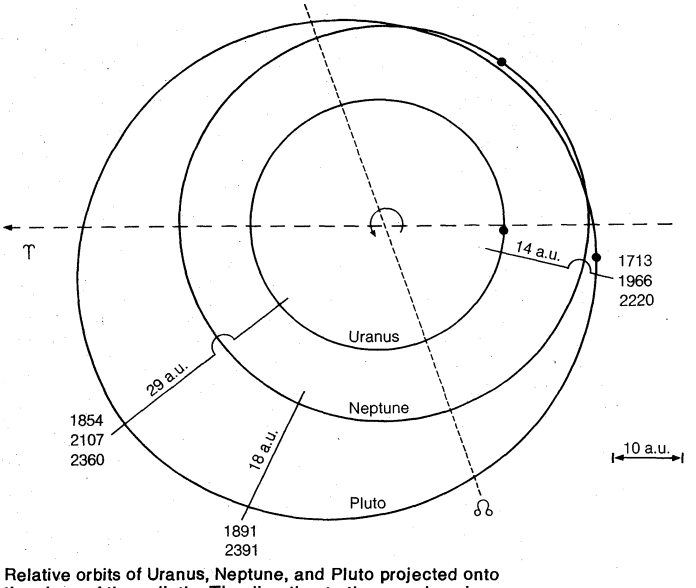
Discovery. Before Pluto's discovery in 1930 several astronomers, making calculations similar to those that led John Couch Adams and Urbain Leverrier to predict correctly the position of Neptune before its discovery, had interpreted the motions of Uranus and Neptune as revealing gravitational perturbations beyond those accountable for by all the planets known at that time. Among those who thus predicted a planet beyond Neptune was Percival Lowell, to whom credit for a successful prediction was generally given. Not only were his predicted orbital elements close to those of Pluto but he also initiated the effort that led ultimately to its discovery by his successors at the Lowell Observatory, Flagstaff, Arizona. Although Lowell's famous *Memoir on a Trans-Neptunian Planet* (*Memoirs of the Lowell Observatory*, vol. 1) was not published until 1915, the year before his death, he and his colleagues carried out intermittent searches between 1905 and 1916 with a variety of telescopes of limited capability. On the basis of experience gained in these early efforts, a fast wide-field telescope was designed specifically for the trans-Neptunian planet search.

After the telescope was completed in 1929, the project was placed in the hands of Clyde William Tombaugh. He painstakingly compared pairs of photographic plates of selected sky regions, taken several days apart, looking for images of moving objects. The project was carefully planned so that, in examining millions of stellar and starlike images, other moving bodies such as the numerous asteroids could be easily recognized and not be misidentified as the sought-for planet. On Feb. 18, 1930, Tombaugh discovered the trans-Neptunian planet on plates that he had taken on Jan. 21, 23, and 29, 1930. Announcement of the successful discovery was not made until March 13, 1930, the anniversary of Lowell's birth date, but in the interval the Lowell Observatory astronomers were able to assemble further supporting evidence and confirm the discovery. After a long deliberation over many suggested names for the new planet, the name Pluto was chosen, as it could be denoted by the symbol ♇, made up of Percival Lowell's initials.

Even while the announcement of this significant discovery was drawing the attention of the entire astronomical world to the new planet, certain grave doubts were being expressed as to whether it indeed was Lowell's trans-Neptunian planet. The difficulty was that Pluto was much fainter than had been expected (about one-tenth as bright) and it had no observable disk. It was therefore difficult to resist the conclusion that it was altogether too small and, hence, of insufficient mass to have caused the large perturbations in Uranus' and Neptune's motions that had led Lowell and others to predict its existence and orbital motion. Soon after Pluto's discovery, the noted celestial mechanician E.W. Brown showed that Lowell's prediction was, in fact, dynamically invalid. Then, E.C. Bower of the Lick Observatory, Mt. Hamilton, California, among others, convincingly demonstrated that Pluto's mass was probably less than one-half that of the Earth and that the resulting perturbations to the motions of Uranus and Neptune must have been significantly smaller than the probable errors of the earliest observations of these bodies. Modern data have since confirmed this conclusion. Thus Pluto, at first considered as the second planet, after Neptune, to have been discovered through prediction, is actually the second planet, after Uranus, to have been discovered by accident. Rather it is the first planet to have been discovered through a systematic and comprehensive search.

The orbit. Pluto moves in a markedly eccentric orbit about the Sun with a period of about 248 years. The orbit is more eccentric than that of any other planet and has a very high inclination, about 17°, to the mean plane of the solar system. The ellipse designated as Pluto's orbit in the figure is the projection of the true orbit onto the ecliptic plane; a proper representation of the true orbit would intersect this plane in the line of nodes marked ☊ and lie below it to the left, above it to the right.

The near commensurability of the periods of revolution of the three outermost planets brings the planets into successive conjunctions at about the same places in space. At intervals of approximately 500 years, during which Pluto makes two revolutions and Neptune three, these two planets come closest to each other and are then separated by only 18 astronomical units (1 a.u. = 1,496 × 10^5 km); this last occurred in 1891 and will occur again in 2391, as indicated in the figure. Similarly, Pluto and Uranus have close conjunctions although in two series. For example, in the years 1854, 2107, and 2360, Pluto and Uranus are approximately 29 a.u. apart, and in the years 1713, 1966, and 2220 only 14 a.u. apart.



Relative orbits of Uranus, Neptune, and Pluto projected onto the plane of the ecliptic. The direction to the vernal equinox is indicated by ♈. The line of nodes between Pluto's orbital plane and the ecliptic plane is indicated by ☊. The orbital motion is counterclockwise and the positions of the three planets in 1967 are indicated by dots. Approximate locations of conjunctions at the dates shown are indicated by the lines joining the orbits.

The surface of the planet. Pluto has no observable disk and is so faint that it has not been profitable to look for surface detail or to employ the more customary visual and spectroscopic techniques used in studying other planets. As early as 1933, however, W. Baade, in determining

Orbital and Physical Data for Pluto
(osculating elements for Epoch JD 2,430,000.5)

Mean distance from Sun	39.53 a.u.
Least distance from Sun	29.81 a.u.
Greatest distance from Sun	49.26 a.u.
Eccentricity	0.246
Inclination to ecliptic	17.123°
Period of revolution	248.5 years
Time of perihelion	1989.8
Visual magnitude at mean opposition	15.0
Period of axial rotation	6.3867 days
Diameter	probably between 5,000 and 6,000 kilometres
Mass	very uncertain; estimated to lie between 0.03 and 0.1 terrestrial mass
Satellites	none known
Atmosphere	unknown; traces of helium and neon probably present

Pluto's brightness at the Mount Wilson Observatory, Pasadena, California, found some evidence that it might be variable.

About 20 years later the postwar development of photoelectric techniques in astronomical photometry made possible much greater sensitivity and accuracy. Photoelectric observations of Pluto's brightness made between 1952 and 1965 by G.P. Kuiper, M.F. Walker, and R.H. Hardie at the McDonald, Mount Wilson, Lowell, and Dyer observatories confirmed Baade's suspicion. Pluto varies by about 15 percent in brightness with a period of 6.39 days. The periodic variation in reflected light is interpreted as due to the combined effects of axial rotation and of surface features of differing brightness. As yet there are insufficient data for determining the direction of the spin axis in space; but ultimately, as Pluto is observed over a larger portion of its orbit, this determination should become possible as the observed light curve goes through gradual changes in amplitude or shape with time.

Mass, diameter, and density. Because Pluto is not known to have a satellite of its own, the only direct means of determining Pluto's mass is through its perturbations on other bodies. To date the only other bodies available have been Uranus and Neptune, although in the future artificial space vehicles may provide a much more sensitive means for determining the mass of Pluto. Very slight adjustments among the orbital elements and masses of Saturn, Uranus, and Neptune can have a substantial influence on the determination of Pluto's mass, so that even the most reliable direct determination carried out by applying perturbation theory to Uranus' and Neptune's motions must be regarded as only approximate. Such a determination made at the U.S. Naval Observatory in 1970 by R.L. Duncombe, W.J. Klepczynski, and P.K. Seidelmann resulted in a mass for Pluto of 0.17 that of the Earth. Other considerations, noted below, lead astronomers to believe that this is overestimated by at least a factor of 2.

Bearing on the problem of Pluto's mass is that of its diameter. A direct observation made in 1950 by G.P. Kuiper with the 200-inch telescope at the Palomar Observatory revealed a diameter of close to 6,000 km, although a value of about 5,000 km was not ruled out. In later astrometric observations of an appulse, or near-occultation, of a faint star by Pluto, it was determined by O.G. Franz, J.B. Priser, and L.W. Sanders at the U.S. Naval and Lick observatories, that Pluto's diameter could not exceed 5,800 km for it would then have occulted the star's light, a circumstance ruled out by numerous observers. A lower limit to the diameter of about 2,000 km can be fixed from the knowledge of Pluto's brightness and distance and the assumption of perfect reflection, but on the basis of a more reasonable and typical estimate of the albedo, a diameter between 4,000 and 6,000 km is indicated. It would thus appear that taking Pluto's diameter to be about 5,500 km cannot be far wrong.

The dynamically determined mass, 0.17 of the Earth's mass, taken at face value, together with a diameter of 5,500 km, implies a mean density for Pluto of about 11.7 grams per cubic centimetre, a highly unlikely value. On the other hand, if the mean density were only as high as that of Earth (that is, 5.5 grams per cubic centimetre) a diameter of 7,000 km would be required to account for a total mass of 0.17 of the Earth's mass. As this value of diameter lies outside of the recognized possible range, one must conclude that either the mean density is higher than that of the Earth or the dynamically determined mass is too great. The present opinion of authorities in planetary structure is that the mean density is probably no higher than that of the Earth or Mercury; accordingly the mass most probably is less than 0.1 and could be as little as 0.03 of the Earth's mass.

Possible atmosphere. Whether Pluto has an atmosphere is not known with any degree of assurance, but certain conditions relating to a possible atmosphere can be stated in general terms. It can be estimated that the surface temperature on the sunlit portion of Pluto's surface is about 60° K when the planet is at perihelion and 47° K at aphelion. The dark side of the planet probably has a temperature of about 20° K. At such temperatures, virtually all materials are frozen, or at least liquefied, except a few gases such as hydrogen, helium, and neon, so that any atmosphere would be composed of these gases. While hydrogen would be expected to be the most abundant of these due to its high cosmic abundance, it is also the most likely to escape. Until more exact knowledge is available regarding Pluto's mass, size, and surface temperature, it is not possible to deduce whether a hydrogen atmosphere exists. On the other hand, any traces of helium or neon are probably retained in atmospheric form, the neon probably freezing out on the dark side and evaporating on the sunlit face. None of these atmospheric gases would produce absorption lines in the visual portion of the spectrum and hence are undetectable by customary spectroscopic means; this has been confirmed by Tobias Owen and R.H. Hardie, who find Pluto's spectral features identical with those of the Sun. Since such an atmosphere would be essentially transparent, the slightly redder quality of the observed reflected light compared with sunlight is no doubt due to the character of the surface.

Comparison with other planets. Pluto differs greatly from the major planets in its general physical properties —the other outer planets have great mass, large size, extensive atmosphere, and very rapid axial rotation—and has much higher orbital eccentricity and inclination. Thus Pluto probably had an origin and evolution different from that of the major planets. Because Pluto's physical properties appear more akin to those of the satellites, especially Triton, and because Pluto's orbit nearly intersects Neptune's, it has been proposed by R.A. Lyttleton and Kuiper that Pluto may in fact be a runaway satellite originally bound to Neptune.

BIBLIOGRAPHY

Discovery: C.W. TOMBAUGH, "The Trans-Neptunian Planet Search," in G.P. KUIPER and B.M. MIDDLEHURST (eds.), *Planets and Satellites*, vol. 3, ch. 2 (1961).

Light variation: W. BAADE, "The Photographic Magnitude and Color Index of Pluto," *Publs. Astr. Soc. Pacif.*, 46:218–221 (1934); M.F. WALKER and R.H. HARDIE, "A Photometric Determination of the Rotational Period of Pluto," *Publs. Astr. Soc. Pacif.*, 67:224–231 (1955).

Diameter: I. HALLIDAY *et al.*, "An Upper Limit for the Diameter of Pluto," *Publs. Astr. Soc. Pacif.*, 78:113–129 (1966).

Mass: R.L. DUNCOMBE, W.J. KLEPCZYNSKI, and P.K. SEIDELMANN, "Orbit of Neptune and the Mass of Pluto," *Astr. J.*, 73:830–835 (1968).

(R.H.H.)

Pneumatic Devices

Pneumatic devices generate and utilize compressed air in a wide variety of applications. Examples include the inflation of automobile tires, the operation of pneumatic construction tools (rock drills, pavement breakers, rivetters), and machine tools for metal processing (forging presses and grinders), for air cushions (to absorb shocks), paint sprayers, blast cleaners, atomizers and conveying materials by pneumatic tube.

Advantages of compressed-air power

Compressed-air power is flexible, economic, and safe. An air device creates no spark hazard in an explosive atmosphere and can be used under wet conditions without electric shock hazard. For these reasons compressed-air power is the only type of power used in some mining and construction operations. A relatively small compressor suffices to fill a storage tank for intermittent use, and no return lines are needed.

Several characteristics of a compressed-air system are important in meeting special service requirements. It is relatively easy to connect one device (such as a valve or a cylinder and piston) to another by pipe, tubing, or flexible hose. Many actions can be controlled by a simple manipulation of valves. The motion of an actuating piston in a cylinder can be changed quickly and in small steps with practically no shock. An air system can provide great flexibility in speed and motion control. Relief valves are easily arranged to protect a system and avoid damage. Control of operations is simple, efficient, and centralized. In general, air systems have relatively few moving parts, contributing to high reliability and low maintenance costs.

Early devices. The ordinary hand bellows, used by early smelters and blacksmiths for working iron and other metals, was a simple type of air compressor. The air intake consisted of several holes in a piece of wood, covered with flaps that served as valves. A simple check valve in the discharge prevented air from being drawn back into the bellows during the suction stroke. In the time of Hero (probably 1st century AD) a simple jet-type compressor was used to provide air for smelting and forging.

In the 17th century Otto von Guericke, of Magdeburg, Germany, carried out notable experiments and made significant improvements in compressors. In 1829 a stage or compound compressor was patented, which involved compressing air in successive cylinders. Cooling during compression by jets of water sprayed into the cylinder was introduced about 1872; later, a better system of cooling by the use of water-jacketed cylinders was developed. In the United States the first compressor used in large-scale work was a four-cylinder unit for the Hoosac Tunnel, at North Adams, Massachusetts, in 1866.

The 20th century witnessed a large increase in the use of compressed air and compressed-air devices. The introduction of jet engines for military and passenger aircraft stimulated the use and improvement of centrifugal and axial-flow compressors. The further development of automatic machinery, labour-saving devices, and automatic-control systems led to an increase in the use of pneumatics. In the late 1960s there began a significant development of a new class of compressed-air devices: digital-logic pneumatic-control components, which can be used in various power and control systems.

AIR COMPRESSORS

An air compressor is a power-driven machine for compressing air from some initial intake pressure (usually atmospheric) to a higher pressure. The familiar hand tire pump is an example of a simple machine of this type. For air pressures below eight ounces per square inch a *fan* is the most economical device, a *blower* is most efficient for pressures up to about ten pounds per square inch, and for higher pressures a *compressor* must be used. Compressors are used on aircraft engines to compress the incoming air before fuel combustion.

An air compressor may be in a fixed position, as in a factory, or portable as in building and road construction work.

Compressors, as well as other fluid machines, can be classified in two main types, depending on the air or fluid action: (1) the positive-displacement type and (2) the velocity or dynamic type. (Further information is contained in the article PUMP.)

In the positive-displacement or static-pressure type, the characteristic action is a volumetric change or displacement action. Successive volumes of air are confined within a closed space and the pressure increased by reducing the volume of the space. In the simple hand tire pump pressure is developed by moving a piston in a cylinder. The positive-displacement type may be subdivided into reciprocating (back-and-forth straightline motion) and rotary (motion in a circular path) compressors. In a positive-displacement machine, neglecting leakage, the volume rate of flow (cubic feet per second) through the compressor is essentially constant over a wide range of discharge pressures.

The common propeller type of desk fan is an example of a dynamic machine. The dynamic type may be subdivided into centrifugal (flow through a rotating runner or rotor primarily in a radial direction), axial-flow (flow through a runner primarily in a direction parallel to the axis of rotation), and fluid jet compressors.

Reciprocating compressors. A positive-displacement reciprocating compressor consists of a cylinder with a reciprocating piston and check valves for directing the air to and from the cylinder. The check valves permit flow in one direction only. The crank can be driven by an electric motor, an internal combustion engine, or a steam turbine. If it be assumed that air at atmospheric pressure exists outside the inlet valve and that the piston is at the right end of the stroke, then as the piston moves to the left the outlet valve will be closed, the pressure inside the cylinder will drop below atmospheric and air will flow into the cylinder through the inlet valve. When the piston is at the left end of the stroke the cylinder is full of air. As the piston starts to move from left to right, both valves are closed and the air is compressed. Assume that it is desired to develop an air pressure of 100 psig (pounds per square inch gage, or above atmospheric pressure). The discharge valve is set to operate at this pressure. Then at some position of the piston (between the left and right ends of the stroke) the air pressure reaches 100 psig and, as the piston continues to move, the outlet valve opens and compressed air leaves the cylinder. In a single-acting compressor the air is compressed only on one side of the piston. In a double-acting piston arrangement, the air is compressed on both sides of the piston.

Air discharged from a compressor is frequently piped into a tank or receiver for storage, to smooth out pulsations, and to remove some of the heat generated by compression. In some compressors cooling fins are arranged on the outside of the cylinder. The compression may be done in two or more stages; that is, the air is compressed only part way to its final pressure in the first stage, some heat is extracted by an intercooler, and the air is raised to its final pressure in an additional stage or stages.

In positive-displacement reciprocating compressors the number of stages is normally one for 0 to 150 psig, two for 80 to 500 psig, and three for 500–2500 psig. The reciprocating compressor is useful for supplying small amounts of air at relatively high pressures.

Rotary compressors. Positive-displacement rotary compressors are designated according to the type of mechanical rotor element. A rotary lobe compressor or blower consists of a pair of meshed identical lobes in a housing or casing. The two lobes are usually kept in proper relative rotation by a pair of external meshed gears. The incoming air is trapped between the lobe surfaces and the casing, compressed, and pushed to the outlet. The construction is simple and no separate valves are needed. Longitudinally the lobes can have helical or straight surfaces. Compressors of this type can develop 10 psig in one stage. In a rotary sliding-vane compressor the rotating member, with its sliding vanes, is eccentric (off-centre) in the casing by an amount *e*. The entering air is trapped between the vanes (which ride on the inside of the case) and is carried to the outlet. Sliding-vane compressors are used for pressures of 50 psig in one

Rotary lobe compressor

stage and 125 psig in two stages, with motor drives of up to 700 horsepower.

Centrifugal compressors. A centrifugal compressor consists of a rotating member or "impeller" with fixed vanes and a casing surrounding it. In a low-pressure centrifugal compressor the air flows axially through the inlet pipe to the centre of the rotating impeller and then outward radially, perpendicular to the axis of rotation. In the impeller the centrifugal action serves to build up velocity and pressure. At the impeller exit the air has a high velocity and flows into the volute, where it is led through the conical diffuser to the exit. The function of the volute and conical diffuser is to slow down the fluid and build up the pressure.

Centrifugal compressors are built with many different arrangements of impellers and with many variations in other details of construction. In the single suction compressor the fluid enters the impeller from one side only. In the double-suction compressor air enters both sides. A compressor may consist of several impellers on one shaft; that is, several single-stage compressors may be arranged in series.

The impeller can be driven by an electric motor, a gas turbine, or a steam turbine. The speed of rotation may range from 3,000 to 10,000 revolutions per minute and pressure up to 100 psig, depending on the number of impellers used. Flow rates may range up to as much as 150,000 cubic feet per minute. Centrifugal compressors are employed to handle large volumes of flow at moderate pressures.

Axial-flow compressors. Figure 1 shows a diagrammatic sketch of a multistage axial-flow compressor. Such

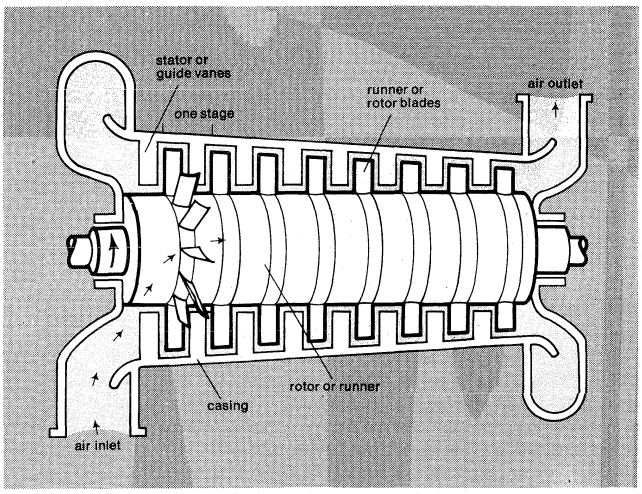


Figure 1: Multistage axial-flow compressor.

a machine may include a number of runner or rotor blades and a number of fixed, stationary guide or stator blades. The term stage designates a pair of blade rows, one row on the runner and the other on the adjacent fixed guide vanes. In an axial-flow compressor the main flow is in an annular space parallel to the axis of rotation of the runner. In a compressor stage, the curved runner blade does work on the air to increase the pressure; the fixed guide vane serves to change the velocity and increase the pressure. Like the centrifugal compressor, the axial-flow compressor is usually operated at high rotative speeds and the pressure rise is relatively low.

The centrifugal compressor is normally employed where relatively high pressures and low flow rates are desired, whereas the axial-flow compressor is more useful for relatively low pressures and high flow rates.

Jet compressors. In the jet compressor the pressure of the air is increased as it is forced through an arrangement of fixed channels by a motive fluid. These devices are sometimes called injectors, ejectors, or evacuators.

Injectors, ejectors, and evacuators

In operation, some motive fluid enters the nozzle at a high pressure. This motive fluid is accelerated in the nozzle as the pressure drops. The air to be compressed is then entrained by the high-velocity motive fluid. The entrained air is accelerated while the motive fluid is decelerated. The mixture enters the diffuser whose function is to slow the fluid down and to increase the pressure. The pressure at the exit is higher than the pressure at inlet. The motive fluid can be separated from the compressed air and reused.

In a steam-jet ejector, steam is the motive fluid. In many applications a steam-jet ejector functions as a vacuum pump; it removes air from a region initially at a pressure less than atmospheric and exhausts this air at atmospheric pressure.

PNEUMATIC TOOLS

Pneumatic tools can be divided into two groups depending on the driving method: rotor, or reciprocating piston. Both types are known as air motors. A rotating type of compressor, operating in reverse, serves as one type of motor. Compressed air enters the housing, pushes on the vanes, and rotates a central shaft or spindle. A drill, grinding wheel, or other device is fastened to the spindle. A reciprocating compressor, operating in reverse, also functions as a motor. Compressed air enters the cylinder, expands, and forces the piston to move. The return stroke may be actuated by compressed air on the other side of the piston, or by spring action. A tool, such as a rivetting hammer, may be connected to the reciprocating piston. Pneumatic tools are normally supplied with compressed air at about 90 psig.

Principal types of pneumatic tools

With compressed air as the power source, tools have been designed that are relatively light in weight, compact, portable, easy to operate, and free from electrical shock and spark hazards. In underwater operations, compressed air prevents water from entering the air motor.

Pneumatic tools can also be divided into two groups according to the type of tools: portable tools and rock drills.

Portable pneumatic tools. Portable pneumatic tools include abrasive devices (grinders, buffers, and sanders), drills, reamers, tappers, stud setters, screwdrivers, nutsetters, shears, wrenches, and impact tools. They are normally powered by a rotary-vane type of air motor. Operating speeds can be varied by throttling the air to the motor. Air motors do not become hot when overloaded; they will stand repeated stalling and rapid reversals without damage. Figure 2 illustrates a grinder; the air motor is typical for this class of device.

Portable tools also include percussion tools, chipping hammers, rivetters, scaling hammers, sand rammers, compression and yoke rivetters, air hoists, concrete vibrators, paint mixers, valve grinders, and saws.

Reprinted from the *CAGI Fundamentals of Compressed Air Power* with permission of Compressed Air and Gas Institute, 122 East 42nd Street, New York, New York 10017

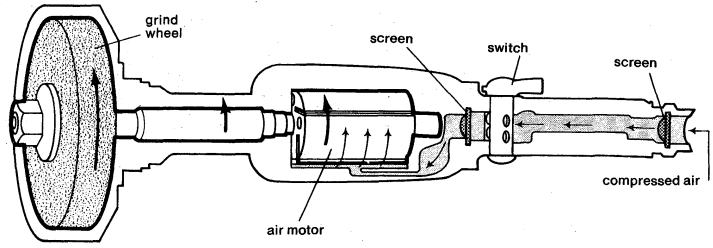


Figure 2: Pneumatic grinder.

Pneumatic chipping or scaling hammers contain an air-operated piston that delivers successive blows to a chisel or forming tool at the end of the hammer. The valve type of tool has a separate mechanism to control the air flow to the piston, thus allowing the operator to control the speed and force of the blows. This type is used for chipping and rivetting, for which control is important.

There are two designs of rivetting hammers. In a compression rivetter the compression or squeezing action on the rivet is obtained from an air piston connected to a cam, wedge, or toggle. A yoke rivetter has an air-operated clamp or vise that holds the work in place; the yoke absorbs the hammering action and thus reduces operator fatigue.

Hoists operated by compressed air are used in operations requiring accurate control of lifting or lowering speeds. Air hoists are used outdoors and under conditions in which corrosive fumes, explosive gases, or inflammable fluids are present. Safety devices prevent dropping of the load when the air pressure falls or the brakes fail.

Specialty tools

There are also various portable specialty tools, such as concrete vibrators, countersinking tools, spikedrivers, paint mixers, air cranking motors, railway roadbed tampers, valve grinders, reciprocating filing machines, and shank grinders.

Rock drills. For mining and rock excavation the usual air-operated tool is the hammer drill or percussion hammer, composed of a piston and a drill made of high-carbon steel. The drill is held loosely in a chuck at the end of the cylinder and is struck by rapid blows from the freely moving piston.

For downward-sloping holes, some means must be provided for removing drill cuttings, dust, and sludge. A hollow drill is usually used, and water or air is passed through it to remove the cuttings and cool the drill bit. Also, the bit must be rotated to produce a round uniform hole; this may be done either manually by the operator, using a wrench attached to the drill, or automatically by a ratchet-and-pawl mechanism that turns the bit slightly after each blow from the piston.

The hand-held rock drill with a self-rotating hammer is sometimes called a sinker or jackhammer. Rock drills are frequently classified on the basis of weight. A drill of less than 25 pounds may be used for drilling shallow holes in brick or concrete. A 55-pound drill may be used for road construction.

Drifter drills

The drifter drill, with an air-cylinder diameter of 3, 3½, or 4 inches, is used for horizontal holes in mining operations and tunnel driving. It is mounted on some type of rig or frame and is mechanically fed into the work. Stoper drills are used primarily on up-hole or overhead drilling because of the automatic feed characteristics. The usual stoper is a hammer drill with a self-rotating drill bit and an automatic feed by means of an air piston.

Large air-operated earth drills, mounted on motor trucks on trailers, are used for digging water wells and blast holes for quarry operations. A high capacity compressor provides air not only to power the drill tool but also to raise the tools in the hole and to remove drill cuttings from the hole. Such machines are used to advantage in areas where surface water supplies are insufficient to provide the drilling fluid needed for standard rotary and cable-tool well drilling machines.

Paving breakers. Hand-operated pneumatic paving breakers usually use solid steel drills and are not equipped for automatic rotation. Figure 3 shows a portable paving breaker. One type of tool is valve actuated, another is valveless. Heavy machines of about 80 pounds are used to break concrete pavement, foundations, and boulders. Medium breakers, weighing about 50 to 70 pounds, are used to break light concrete floors, macadam, and frozen ground. Light tools, weighing less than 50 pounds, are used for breaking floors, paving, and masonry walls. Heavy and medium-weight breakers can be adapted for driving spikes.

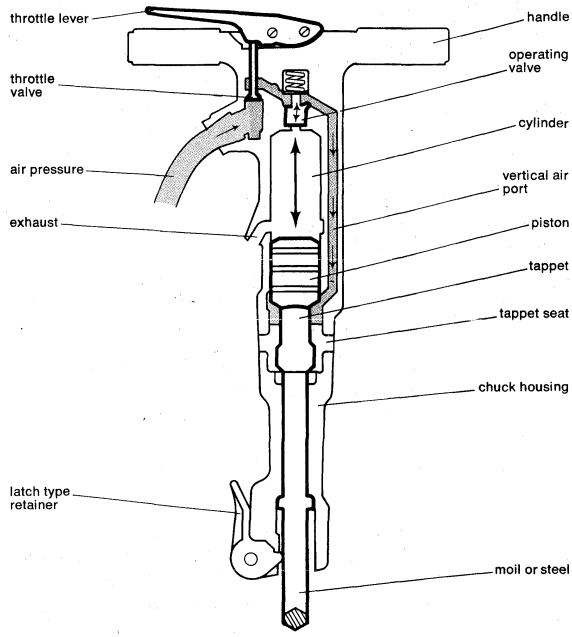


Figure 3: Pneumatic paving breaker.
Reprinted from the *CAGI Fundamentals of Compressed Air Power* with permission of Compressed Air and Gas Institute, 122 East 42nd Street, New York, New York 10017

OTHER PNEUMATIC DEVICES

Compressed air is a good vehicle for conveying a paint spray. In a spray gun, the paint (lacquer, enamel, or plastic coating) is atomized and mixed with compressed air. The principle of operation is similar to that of the jet compressor, with the compressed air serving as the motive fluid to draw the paint into the mixing area. Spray painting has various advantages, such as good coverage of the surface and low cost. Spray painting usually implies covering relatively large surfaces, such as a building. The term airbrush usually implies a device for developing a fine, small diameter spray of paint, protective coating, or liquid colour. The airbrush can be a pencil-shaped atomizer used for shading drawings and retouching photographs.

Pneumatic conveyers are used in many applications for handling materials. In a pressure system the outlet of the compressor leads into the inlet of the conveyer system. In a vacuum system the compressor inlet is at the end of the system. The air pressure difference across the system depends on the material to be handled.

Pneumatic mail delivery

In many places, mail is handled by pneumatic transport capsules in tubes—a technique based on an invention. All sorts of materials may be conveyed by pneumatic systems, from ashes and cement to frozen foods, minerals, nuts, and seeds. Pneumatic handling is safe, fast, clean, automatic, and flexible.

Various recently developed vehicles are supported by a cushion of air. In the planning of many city transit systems, consideration is being given to air-cushion vehicles capable of speeds up to 300 miles per hour.

Brakes on trains are operated by air pressure. A piston rod from an air cylinder exerts force on the braking device. On railroad cars the air-brake system includes a compressor, pneumatic valves, regulators, piping, reservoir, and other accessories. There are levers, cylinders, and other rigging to apply forces to the brake shoe, which bear directly on the rim of the wheel. Various automatic-control safety arrangements assure a definite braking action should some malfunction develop.

The city of Paris has a system of underground pipes, varying in diameter from 1½ inches to 20 inches, which carry compressed air to users of pneumatic devices in all sections of the city. The system was installed in the 1880s, before equipment for the generation and distribution of electric power was available.

BIBLIOGRAPHY. COMPRESSED AIR AND GAS INSTITUTE, *Compressed Air and Gas Handbook*, 3rd ed. rev. (1966), a reference book on all phases of industrial air and gas compressors, and compressed-air-powered portable tools and rock drills used by industry; W.E. WILSON, *Positive-Displacement Pumps and Fluid Motors* (1950), an engineering text and reference book covering the design, theory of operation, and characteristics of a class of pumps and fluid motors that includes pneumatic devices; B.W. ANDERSEN, *The Analysis and Design of Pneumatic Systems* (1967), a discussion of the basic mechanics of compressible flow, the detailed design of pneumatic components and systems, and the methods for analyzing pneumatic systems; *Hydraulics and Pneumatics* (1948–), a monthly trade magazine including articles on the design and application of hydraulic and pneumatic components and systems.

(R.C.B.)

Poales

Poales is an order of flowering plants consisting of a single family, the grass family, called the Gramineae, or Poaceae. The family Gramineae probably is the most

abundant family of vascular plants, considered in terms of numbers of individuals; it dominates great areas of the earth's surface, including prairies, steppes, savannas, pampas, and paramos (all types of grassland). The family is also fairly sizable in its number of species, with an estimated 600 to 700 genera containing 6,000 to 10,000 species. It ranks about fifth among the angiosperms (flowering plants) in total number of species. The family is of prime importance to the world's agriculture because it contains all the important cereal grains. Moreover, a large number of species furnish the principal sustenance of grazing livestock and wild herbivores. The group was intimately involved in the evolution of man's domestic animals and very likely of man himself.

A great diversity of products is derived from grasses, but grain is the most important single category. Scarcely less important but more difficult to evaluate are the cultivated and wild forages obtained from grasses. The bamboos, for instance, which are highly specialized woody grasses, provide building and handcraft material, paper pulp, and edible shoots.

GENERAL FEATURES

Structural diversity and habitat range. Grasses are found in all parts of the land surface of the earth, and their ecological diversity is tremendous. The greatest number of species is found in the tropics, as is true of most plant families. The family as a whole ranges from the tropics to the high Arctic, with two species occurring on the northern extremity of the Antarctic continent.

Habitats occupied by grasses

Although the greatest abundance of grasses occurs on open lands, such as the prairies, steppes, and paramos, some species are adapted to many specialized habitats, such as saltwater and freshwater marshes, running or still water, rain forests, deserts, and Alpine or Arctic tundra. Because of their manifold adaptations to varying climates and habitats, the grasses have become an extremely successful family, and, consequently, there are probably more individual grass plants than there are of all other vascular plants combined.

Grasses range from minute forms, such as the annual bluegrass (*Poa annua*) and the high Alpine *Phippsia algida*, which are only a few centimetres tall, to the giant timber bamboos, which may reach 100 feet (30 metres) in height. The stem width is also extremely variable. Small grasses have stems only a millimetre or two in diameter, while the larger bamboos have stems as much as ten inches (25 centimetres) thick. Such large stems are produced only by very old clumps that have a large amount of stored food in their underground parts.

Grass leaf blades are commonly thought of as being ribbon shaped and with parallel veins. While the great majority of grass leaf blades are of this shape, some of the tropical forest grasses have broadly oval blades. One remarkable grass genus, *Pharus*, has broad leaf blades with the veins diverging from the midrib. The leaf blade is attached to a petiole (leafstalk, not a true one in this case), which has 180° twist, so that the leaf blade is actually upside down. Certain broad-leaved grasses of the genus *Raddia* have short sprays of crowded leaf blades that mimic the appearance of the compound leaves of certain legumes (members of the pea family, Leguminosae or Fabaceae). In this case, the leaf blades even fold up at night as legume leaves do. The leaves of desert grasses, however, are stiff, harsh, and frequently inrolled. In some species they become bristly and are probably unattractive to grazing animals. Aquatic grasses are also known, and some of these are able to float on the surface of the water because of inflated leaf sheaths or waxy, nonwettable leaves.

Economic importance. The grasses are surpassed in numbers of species by the orchids, composites (sunflowers, asters, etc.), legumes, and members of the coffee family. No other plant family, with the exception of the legumes, however, remotely approaches the grasses in agricultural importance. Their principal areas of utility include exploitation for cereal grains, sugar, timber and paper pulp, hay, pasturage, native forage, and numerous other lesser known products.

Cereal grains. The grains, which form the staple foodstuffs for much of the less affluent majority of mankind, are all derived from grasses. The more important of these are rice, wheat, maize (corn), oats, barley, rye, sorghum, and millets.

Rice (*Oryza sativa*) is an Asiatic tropical grass and is the only semi-aquatic species among the cultivated cereals. Cultivation is widespread in all warm temperate and tropical climates where sufficient moisture is available. Rice is traditionally thought of as a crop grown under flooded conditions, but certain strains will grow under moist conditions without flooding. Such strains are called upland rices, and they generally yield much less grain than does flooded rice. Although statistics are inexact, rice is probably the world's leading grain crop. Recent introduction of semidwarf strains with high productivity has greatly increased rice yields and is helping to make former rice-importing regions self-sufficient.

Wheat (*Triticum aestivum* and other species) originated in the Middle East and has spread to all temperate and cool temperate areas of the earth, being cultivated widely in southern Europe, the U.S.S.R., the United States, Canada, Argentina, and parts of India, China, and Australia. Wheat grain, containing an abundance of the elastic protein gluten, is the only cereal capable of producing a satisfactory light bread. In quantity of production, it probably ranks second to rice.

Maize

Maize or corn (*Zea mays*) is the only cereal of American origin. Discovered in the primitive wild state by the early Indians of Mesoamerica and developed extensively by them, corn became the principal foodstuff of the great Indian civilizations of Mexico, Central America, and the Andean countries. Although corn is of tropical origin, man has selected strains that are adapted to temperate climates with warm summer seasons. Its culture is now extended to all the inhabited continents. The principal use of corn is as animal feed, but great amounts are used as human food or as a source of starch, syrup, and alcohol.

Oats (*Avena sativa*), like wheat, are native to Eurasia and probably originated in the Mediterranean Basin. They are widely cultivated in the cooler temperate portions of the Northern Hemisphere and, to a limited extent, in the Southern Hemisphere. The grain is much used for animal feed, with lesser amounts going into cereals for human consumption. Cultivation of oats has decreased in recent years, however, probably because of farm mechanization and the decline in numbers of draft animals. Although the grain lacks gluten and cannot be used to make bread, the protein quality is excellent, and the quantity is susceptible to being increased by selective breeding. In Latin America, pulverized oatmeal is frequently mixed with water and used as a cold drink, called *avena*.

Barley (*Hordeum vulgare*) is another small grain of Middle Eastern origin. It is adapted to about the same areas and climatic conditions as is wheat, but barley has notable tolerance to drought and cold. Much of the grain is used in animal feed except in less developed countries. One of the principal uses of barley is the production of malt for use in brewing beer.

Rye (*Secale cereale*) is a comparative newcomer among the cereals, and its cultivation is largely confined to areas of cold climates. As a breadstuff, it scarcely competes with wheat. The grain is also used in the preparation of alcoholic liquors, and the straw is used in hats. Rye is the host plant for the fungus that produces medicinal ergot. In the United States, rye is much used to stabilize the shoulders of new roads, because of its ability to grow on poor subsoil.

The sorghums (*Sorghum bicolor*, *S. sudanense*, and others) are the most diverse of the grains, in both aspect and usages. They are of African and Middle Eastern origin and thrive best in warm, often subhumid climates. The various types furnish feed grains, silage, temporary and perennial pasture, syrups, and broomcorn.

Lesser grains

The term millet generally applies to various minor and usually inferior grains, mostly grown in less advanced agricultural areas or as emergency crops. The most common types are foxtail millet (*Setaria italica*), proso millet

(*Panicum miliaceum*), pearl millet (*Pennisetum glaucum*), and ragi, or coracan millet (*Eleusine coracana*). Millet grain is used in animal feeds, birdseed, and for human food.

Sugar. Sugar derived from grasses is mostly the product of sugarcane (*Saccharum officinarum*). Smaller amounts, nearly all in the form of syrup or molasses, are produced from certain types of sorghum.

Timber and paper pulp. These are the principal products of the bamboos (subfamily Bambusoideae). Certain types of soft construction board, such as Celotex, are often produced from sugarcane refuse (bagasse).

Hay, pasturage, and native range. The total production of these vegetational types is difficult to estimate, but land devoted to them aggregates more total area than all cultivated cropland. Many of these areas can be economically productive only through harvest of forage by grazing animals. The species prevalent in a given area differ greatly in various climatic regions. In the temperate zones some of the principal hay and pasture grasses include timothy (*Phleum pratense*), smooth brome (*Bromus inermis*), orchard grass (*Dactylis glomerata*), redtop (*Agrostis alba*), alta fescue (*Festuca arundinacea*), and Bermuda grass (*Cynodon dactylon*). The widespread Kentucky bluegrass (*Poa pratensis*), a Eurasian species, has assumed the aspect of a native in North America, furnishing large amounts of pasture, as well as being the principal lawn grass in the cooler regions.

Principal hay and pasture grasses

In the western Great Plains and the Southwest of the United States, great areas are still populated by native forage grasses, including buffalo grass (*Buchloë dactyloides*) and species of *Bouteloua* (grama grasses) and *Hilaria* (tobosa and galleta grasses). Species of *Panicum, Andropogon, Spartina,* and *Sorghastrum* are important in tallgrass areas.

In subtropical and tropical areas of the world, most of the wild and cultivated grasses belong to the subfamilies Eragrostoideae and Panicoideae. Species of *Eragrostis* (love grasses), *Cynodon* (Bermuda grass), *Chloris* (finger grass), *Panicum, Paspalum, Digitaria* (crabgrass), *Axonopus, Andropogon* (bluestem grasses), and their allies are important in cultivation.

Unusual products. A number of tropical Asiatic grasses produce sufficient quantities of aromatic compounds to merit use in perfumes. The roots of the vetiver (*Vetiveria zizanioides*) of India yield an excellent perfume after drying. The lemon-oil grass, *Cymbopogon citratus,* has a strong lemon scent, and its foliage is widely used in the tropics for tea or medicine. The grass is grown on a commercial scale for production of lemon oil, which is distilled from the leaves. Citronella oil is similarly distilled from the foliage of *Cymbopogon nardus* and is used as an insect repellent, as a perfume in soaps, and in the synthesis of menthol. The stems of sugarcane are covered with quantities of finely powdered wax that gives them a bluish hue. This wax may be recovered in the processing of cane sugar and is used in the manufacture of coatings for carbon paper.

Paper is made from a variety of grasses, including many bamboos and the North African esparto (*Stipa tenacissima*) and alfa (*Lygeum spartum*). The straw of wheat and rye is used in the manufacture of straw hats. In many parts of the world, hay or straw is used for the thatching of roofs, which last for many years when properly laid.

Special uses of bamboo

The tropical peoples of the world use the bamboos in a great variety of ways, in addition to their common usage as a timber substitute. Fishing poles and props for banana plants are commonly made from bamboo stems. Larger canes are reamed to remove the partitions at the joints and then serve as water pipes. Thin peelings from the exterior of the stems are used in basketry and mat weaving. Elegant screens of fine bamboo fibres are used to filter the suspended pulp used in the making of paper, and the pulp itself is frequently made from bamboos. Young shoots of certain bamboos are used for food and are commonly found in Oriental cookery. Tabashir, a fine, white powder (colloidal silica) found in the hollow stems of some bamboos, has been used in Oriental folk medicine. Other bamboos are grown as ornamentals for their graceful, wandlike manner of growth or as clipped hedges.

One of the most widespread tropical grasses is Job's tears (*Coix lacryma-jobi*). The hard, bony sheaths that surround the female flowers are excellent natural beads and are extensively used for necklaces and rosary beads.

Brooms are fabricated of many materials, but one of the most common materials is the stiff straw from the seed heads of the broomcorn, a variety of sorghum. The immature seed heads are dried and threshed to provide the common broomstraw of commerce.

Deleterious grasses. A considerable number of grasses are regarded as pests by man. Most of these are weeds that invade clean-tilled croplands, although there are various species that invade ranges and pastures; they are disliked because they are mechanically injurious or of lower feed value than the species they displace. Some species are invaders of lawns, where they are objectionable because of colour, texture, or growth habits.

The majority of grass weeds are annuals of rapid growth, not easily suppressed by the usual tillage methods or by the herbicide 2,4-D. Chemicals, however, are increasingly being used to control such grasses. The following list includes some of the more important annual and short-lived perennial grass weeds of temperate North America, listed with the crops or habitats in which they are important: species of *Aristida* (needlegrasses, in deteriorated ranges and wasteland); species of *Avena* (wild oats, in spring wheat or on dry rangelands of the Southwest); species of *Bromus* (ripgut grasses, cheatgrasses, and bromegrasses, in grain crops or on deteriorated Southwestern ranges); species of *Cenchrus* (sandburs, in disturbed light soils); species of *Digitaria* (crabgrasses, in lawns, gardens, and crops); species of *Echinochloa* (barnyard grasses, in crops, especially in moist areas); species of *Hordeum* (wild barley, along roadsides, in pastures and native ranges); and species of *Setaria* (foxtail grasses, in crops).

Important weedy grasses

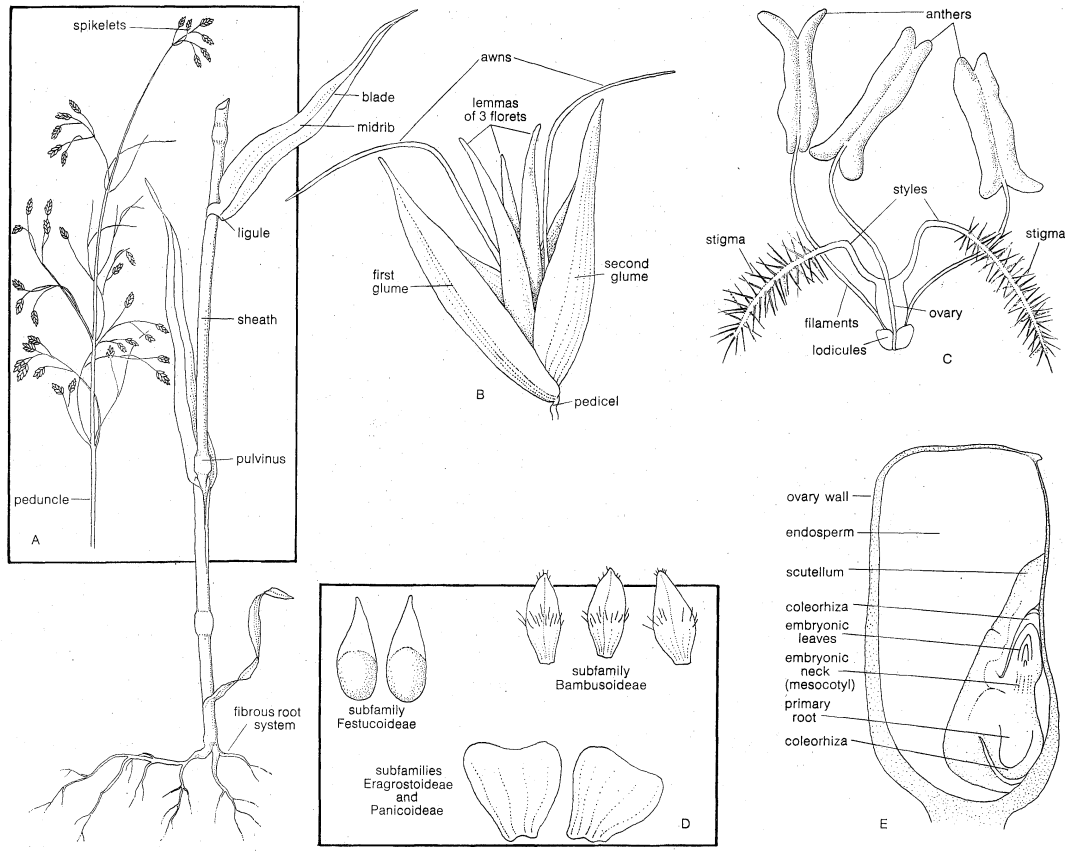
Long-lived perennial grass weeds (those that live through more than one growing season) of importance are far fewer. In the northern United States and Canada, quackgrass (*Agropyron repens*), a persistent European weed that spreads extensively by rhizomes (underground leafless creeping stems) is the most important species. In the south of the United States, Johnson grass (*Sorghum halepense*), a tall perennial with thick, widely spreading rhizomes, is a major competitor with cotton agriculture. Bermuda grass (*Cynodon dactylon*), while extensively planted for lawns and pasturage, is also a severe pest when it invades fields and gardens. Several species of *Andropogon,* notably *A. virginicus* (broomsedge), are invaders of abandoned fields and have low apparent forage value.

NATURAL HISTORY

Life cycle. The fruit of the grasses is a single-seeded dry grain—the caryopsis, as it is known botanically; it is characteristic of the family. Grains of many grasses are capable of immediate germination, but other species have dormant seeds that require afterripening or chilling to permit germination. Such dormancy has adaptive value in preventing germination in the fall or during dry periods, when the seedling would have little chance of establishment. Cultivated types of oats (*Avena sativa*), for example, are capable of germination immediately after the grain is harvested. Probably because of this trait, the plants are virtually unable to maintain themselves in nature. The wild oat (*A. fatua*), which is a probable parent of many cultivated types, has seed that is dormant when mature. Thus the grain is able to survive the winter without injury and then germinate in the spring. As a consequence, *A. fatua* is a widespread weed. Similar examples of dormancy are widespread among grasses.

Value of seed dormancy

In germination, the embryo ruptures the thin ovary wall and any enclosing structures. The first visible signs of germination are the emergence of the tubular sheath (coleoptile) surrounding the young shoot and a similar sheath (coleorhiza) surrounding the primary root. The primary root breaks through the coleorhiza, and, shortly



Grass structures.
(A) Left, grass inflorescence (panicle); right, culm bearing leaves. (B) A common type of grass spikelet (from *Avena*). (C) A grass flower (*Rattraya* species). (D) Grass lodicules (rudimentary perianth parts) from the four largest grass subfamilies. (E) A mature caryopsis, or grain (longitudinal section).
Drawing by M. Moran

thereafter, one or more roots (seminal roots) may arise near the attachment of the embryo to a shield-shaped mass of tissue, the scutellum. The coleoptile and the structures it conceals are pushed through the soil by the elongation of the embryonic neck (mesocotyl) until the tip of the coleoptile is exposed to air and light. It is then burst by the emergence of the young leaves. The shape of the first leaf differs in the various subfamilies. In the common grass subfamily Festucoideae of the temperate zone, for example, it is very slender and elongated and tends to be vertical. In the subfamily Panicoideae, whose members are common in warmer climates, the first leaf is short, broad, and horizontal. The young seedling may continue to produce basal leaves for a time, without signs of a visible stem, or it may rapidly produce a flowering stem (culm). Except in rare instances, some form of vegetative multiplication soon becomes evident. Some species, especially perennials, may produce stolons (creeping leafy stems) or rhizomes. Grasses lacking these usually produce new stems by basal branching (tillering), to form larger and larger clumps. Old clumps may decay in the centre, and the segments of the growing periphery become isolated, forming new, small clumps.

Initiation of flowering

The sexual reproductive cycle in grasses is initiated with the development of the embryonic rudiments of a flower cluster hidden within the sheathing leaves at the tip of a culm or one of its branches. The first evidence of incipient flowering may be the swelling of the upper sheath, or "boot." The young inflorescence (flower cluster), with tightly appressed erect branches, begins to emerge, and the branches spread by means of swellings, the pulvini, at their bases (axils). Flowering usually begins at the tip of the inflorescence and proceeds downward and from the branch tips toward their bases. Reversals or alterations of this flowering sequence also occur. The spikelets of the inflorescence mature and begin to flower after emergence from the boot. Some species bloom early in the morning, others at various times of the day.

Pollination. Practically all grasses that have open flowers are exclusively wind pollinated, although a few instances of insect visits to grass flowers are known. The great prevalence of grasses in the world is a strong indication of their evolutionary success in adaptation to wind pollination. Structural specializations contributing to this success include the clustering of the tiny, naked flowers into scaly spikelets and the aggregation of the spikelets into large compound inflorescences, the presence of large anthers (pollen containers) and abundant light, fluffy pollen, and the relatively enormous and extremely feathery stigmas, which are the female structures that receive the pollen. All these factors and the colonial habit increase the chances of pollination.

After pollination has occurred, the open florets (a floret is a flower plus the two modified leaves that enclose it) close once again, and rather rapid development of the caryopsis (grain) occurs. When the caryopsis is mature, the spikelets dry and the grains are dispersed from the parent plant. In many instances, the spikelets break up into individual florets, each enclosing a grain, but, in many other grasses, entire spikelets or spikelet groups fall from the parent plant and serve as means of distribution.

Many grasses have florets that do not open because the lodicules (swollen motor structures at the base of the flower) do not function or because the spikelets are hidden within the sheaths or even, in rare instances, are underground. Such flowers that fail to open are said to be cleistogamous. They are of necessity self-pollinated. There are both evolutionary advantages and hazards in this type of reproduction. If a grass is predominantly or nearly entirely self-pollinated, the population becomes increasingly inbred. At times, this inbreeding may bring out undesirable traits and cause the death or low fertility of the individuals exhibiting them. In other cases, the inbreeding may cause the expression of traits highly desirable in fitting the species into a specific limited type of habitat, and the species may become extremely success-

Cleistogamy

ful. The homozygous, or inbred, condition of such a species or the individual populations of which it is composed, however, may make it so inflexible that it cannot change rapidly enough to cope with alterations in the environment. Most cleistogamous species, however, have a small percentage of open-pollinated flowers, allowing some crossing. Cleistogamy is especially prevalent in annual grasses, which have to produce a seed crop every year to survive well. A number of species of the genera *Sporobolus* (dropseed grass), *Panicum, Amphicarpum,* and *Danthonia* (poverty oats), among others, have well-developed cleistogamous reproduction. In some species, highly specialized hidden spikelets (cleistogenes) occur. Usually, these are quite different in appearance from the normal spikelets of the same species and may be hidden within the sheaths. In the genera *Triplasis* and *Danthonia,* the cleistogenes are concealed in the lower sheaths. At maturity the culms disarticulate at the nodes (joints), and the segments of stems, carrying the cleistogenes concealed within the sheaths, serve as disseminules (means of dispersal). In a few extreme cases, grasses produce underground spikelets on the tips of slender rhizomes. This process occurs in the genera *Chloris* and *Amphicarpum.* Despite the apparently considerable advantage of having "preplanted" seeds, such grasses are rare and restricted in range.

Cleistogamy is the dominant mode of reproduction in wheat, oats, and barley. Since there is little open pollination and, hence, little opportunity for crossing, strains of these crops tend to remain pure, even though they are grown in close proximity to one another. Rye (*Secale cereale),* on the other hand, is open-pollinated and crosses so freely that few selected strains exist.

Asexual production of seed. In certain grasses, no sexual fusion of egg and sperm occurs in the production of the seed. The embryo arises from a vegetative cell of the ovule, and the offspring are genetically identical with the maternal parent. Pollination may be necessary in some species to initiate the development of endosperm (the starchy food stored outside of the embryo), even though no fertilization of the egg occurs, and in many cases no egg, as such, is produced. Such asexually produced seed is said to be apomictic, and this process of seed production is called apomixis. The process is of great importance in the grass family and occurs in many only distantly related grasses. There is a general tendency for apomixis to be more frequent in grasses with high chromosome numbers. It is particularly prevalent in the large temperate-zone genus *Poa* (the bluegrasses) but occurs as well in many tropical grasses, including species of *Panicum, Paspalum,* and the numerous relatives of the bluestems (*Andropogon* and related genera).

Like cleistogamy, apomixis has both evolutionary advantages and drawbacks. Since an apomictic plant produces offspring exactly like the maternal parent in genetic composition, it may generate a highly uniform stand, with all individuals adapted to the same conditions and all susceptible to the same diseases. If conditions of climate or growth change radically, the population may be unable to adapt rapidly enough to accommodate to the environmental changes and may be destroyed. It is rare, however, that a species is completely apomictic, and generally a small percentage of sexually produced seed is generated as well. Thus the apomictic mechanism may serve as a stabilizing procedure, serving well to adapt a species to a relatively constant environment, while the occasional sexually produced seed allows the introduction of new variation into the population. Along with cleistogamy and vegetative reproduction, apomixis tends to preserve particularly advantageous genetic combinations but to slow down the rate of evolutionary change.

Relationship between grass evolution and grazing animals

Grasses are predominantly plants that tend to form large, treeless communities, usually composed of only a few species. Such open areas were undoubtedly responsible for the evolutionary development of the grazing herbivores, and the grasses themselves bear the evidences of evolutionary specialization to accommodate themselves to being grazed. Among these specializations are the great predominance of vegetative reproduction, by stolons, rhizomes, or tillering. Frequently, heavily grazed grasses may survive for years by vegetative spread, where other types of plants that reproduce only by seed are practically eliminated. Certain savanna grasses, growing in areas where fires occur frequently, have survived because they have evolved deep-seated corms (hard, bulblike stem bases) that are immune to fire damage. Many grassland communities occur in areas that are dry at some period of the year. Perennial grasses of such areas are well adapted to surviving seasonal drought by mechanisms that cause folding or rolling of the leaves or by the dormancy of the plants during drought periods.

The slender, erect culms of grasses are often subject to being knocked down or "lodged" by the wind. Stems that have been lodged are often able to erect themselves by the operation of a swollen region of the stem or sheath at the nodes. This swollen area, the pulvinus, is sensitive to position changes, and, when a normally erect stem is laid in a horizontal position, the pulvini begin to grow more rapidly on the lowermost side and cause the stems to erect themselves.

The grass leaf blade has a growing zone (meristem) near its base, which permits regrowth when the leaf tip is cropped or mowed. This feature of the blades, plus the abundance of basal buds possessed by grasses, serves to permit their survival under intensive grazing or mowing.

Dispersal mechanisms. A great variety of mechanisms has been exploited by the grasses in providing for the dispersal of the plants. Vegetative reproduction is often extremely efficacious in promoting dispersal by means of rhizomes, stolons, or the gradual spread of old clumps. Circumstantial evidence suggests that many colonies of rhizomatous perennial grasses may be hundreds of years old. Grasses of severe habitats, such as the sand binders of coastal dunes (*Ammophila, Elymus, Uniola, Jouvea,* and others) and of continental sand dunes (*Calamovilfa, Redfieldia, Swallenia*) may reproduce almost entirely by rhizomes or stolons in an environment hostile to the establishment of seedlings. Of all the grasses, the bamboos have most fully exploited vegetative reproduction. Many large colonies of bamboos appear to have originated as single individual plants from one seed (such colonies are called clones), as is evidenced by the fact that they remain vegetative for many years, and, when flowering finally occurs, the entire colony blooms simultaneously over great areas. The bamboos may produce short, thick rhizomes and hence produce dense clumps, or they may produce very long, slender rhizomes that spread the plants over considerable distances.

Vegetative reproduction

Another dispersal mechanism that is essentially vegetative is the production and release of vegetative bulbils in the inflorescence. These are modified spikelets that function in place of seed. Bulbous bluegrass (*Poa bulbosa*), a European species occasionally found in the United States, possesses such bulbils. Most grasses, however, produce normal seed on a regular basis, and the dispersal of the grain, either naked or with accessory structures, is the common method of spread.

Some grasses, such as the cereals wheat and rye, release naked caryopses (grains) from the spikelets. Grains as large as these have great appeal to birds and rodents and apparently are not very efficient disseminules. Many fine-seeded grasses, such as the love grasses (*Eragrostis*), however, produce an abundance of fine grains that have no special means of dispersal. The success of these plants indicates that specialized means of transportation are not always necessary to accomplish dissemination.

In a few cases, especially in the genus *Sporobolus* (dropseed grass), the ovary wall of the mature grain becomes gelatinous and swollen when wet, and such grains often adhere to passing animals. *Sporobolus indicus,* a highly successful pantropical weed, uses this dispersal method.

In most grasses, dispersal is effected by the movement of florets, entire spikelets, or groups of spikelets. In many species, individual florets are released from the plant and may be light or hairy enough to be dispersed easily by the wind. The tremendous range of the common aquatic reed (*Phragmites australis*) is testimony to the effectiveness of this type of wind dispersal of the hairy florets. In the

Wind and water dispersal

manna grasses (species of *Glyceria*), which usually grow in or around water, the light florets float upon the water surface. Wild rice (*Zizania aquatica*), which grows in water, is similarly dispersed.

A small number of grasses have spikelets enclosed in hard, bony structures that may be disseminated in various ways, including ingestion by animals that cannot digest them. The hard beads of Job's tears (*Coix lacryma-jobi*) and the hardened fruit cases of *Manisuris* and *Tripsacum* are examples of these structures. Similarly, the female spikelets of buffalo grass (*Buchloë dactyloides*) are enclosed in a bony husk that not only retards germination but also protects the grain. Ants may harvest these structures and carry them away. Certain grasses have spikelets that either form or are enclosed in spiny burs that adhere to animals and clothing. The sandburs (*Cenchrus*) are the most common examples of this. In the genus *Tragus*, the glumes (outer bracts of the spikelet) are covered with hooks that adhere to the wool of sheep, so that this genus is a commensal of the animals. In the bristly foxtail (*Setaria verticillata*), the sterile branches of the spikelike flower clusters (panicles) have minute backwardly directed barbs, so that the whole inflorescence may be carried away by passersby.

In open areas, some grasses have become tumbleweeds —their entire panicles break loose from the mature plant and roll with the wind, dispersing seed as they go. Witchgrass (*Panicum capillare*) is a common example.

Self-planting adaptations

Elaborate self-planting devices have been evolved in certain grasses. Sometimes these work all too well, causing puncture wounds in grazing animals. In the wild barleys (species of *Hordeum*) the mature spike breaks up into individual joints, or internodes, each bearing a triad of spikelets at its apex. The internode has a sharp chisel edge, and the awns, or beards (needlelike projections), of the spikelets have minute, upwardly directed barbs. This combination ensures that spikelet groups placed upon a rough surface tend to move only in one direction, forcing the chisel edge into crevices of the soil or sometimes into animal tissues. Even more elaborate is the ratchet drill device of the genus *Stipa* (porcupine grass, needle and thread). Its lower end, or callus, is extremely sharp pointed and is provided with upwardly directed bristles. When this pointed callus has penetrated any object, be it soil, clothing, or animal fur, it is extremely difficult to remove. The upper end of the floret bears a long awn with a highly hygroscopic (water-absorbing), twisted lower segment. As moisture changes occur, this twisted segment winds up (when drying) and unwinds (when moistened). These changes cause the bent awn to flail about until it lodges against an obstacle and then cause the floret to screw itself into the soil or into the unfortunate animal. Similar devices, although fabricated of different anatomical structures, occur in distantly related genera, such as *Heteropogon* and *Trachypogon*.

Finally, some grasses have adapted themselves as parasites of man's agriculture. Two of these are especially notable—chess (*Bromus secalinus*) and goat grass (*Aegilops*). These annual grasses tend to grow with wheat. They mature at the same time as wheat does, are harvested along with the grain and, unless great care is taken with cleaning the crop, will be replanted along with the next year's seed.

Ecology. Great grasslands exist on all the continents with the exception of Antarctica. In such widely separated areas, climatic, physiographic, and soil differences are great, but in general it may be said that grasses occupy areas where there is at least one period of the year with deficient moisture (usually the winter). With higher and more evenly distributed rainfall, forests hold dominance over grasslands, although man, fire, and grazing animals have apparently caused the replacement of forest by savanna or grassland in many areas, especially in the tropics. Grassland verges into desert as rainfall becomes increasingly inadequate, while forests penetrate the grasslands in especially favourable situations, as along major streams. Within grassland regions, areas of high rainfall tend to be occupied by sod-forming, often rhizomatous grasses, while scattered bunch grasses live in drier parts.

Climatic regulation of grassland occurrence

Even in deserts, scattered grasses occur. The so-called six-weeks grasses, including various short-lived annual species of *Festuca* (*Vulpia*), *Bouteloua*, and *Aristida*, may appear in profusion after heavy rains in semidesert areas.

Other grassland types occur above timberline on mountains. The Alpine tundra of high mountains in temperate North America contains many dwarf species of *Poa, Festuca, Trisetum, Calamagrostis, Agrostis,* and *Phleum.* There is a gradual transition between these Alpine grasslands and the Arctic tundra of low elevations, which includes many of the same genera. The high-elevation grasslands of the American tropics are often designated as paramo. Such areas contain many grass genera that are typically temperate-zone inhabitants and frequently dwarf bamboos as well.

In addition to the grassland species, there are many grasses that live as secondary or primary components of other vegetational types. In the temperate zone, grasses form a characteristic component of the herbaceous understory of deciduous forests. In temperate North America, various species of the following genera may be found in such forests; *Agrostis, Brachyelytrum, Bromus, Cinna, Elymus, Festuca, Hystrix, Melica, Muhlenbergia, Oryzopsis, Poa, Panicum,* and *Schizachne*. In tropical climates, a number of primitive grasses are found in the rain forests. Some of these have broad leaves, very unlike those of typical grasses. Included in this group are such genera as *Streptochaeta, Pharus, Leptaspis, Olyra, Raddia,* and others. In open areas of tropical-forest margins or where forests have been destroyed, the bamboos, which are merely giant woody grasses, may form dense, nearly impenetrable stands. In coastal areas, salt marshes are often populated by dense stands of various species of *Spartina* or *Distichlis.* Coastal or inland dunes are stabilized by species of *Uniola, Ammophila, Jouvea, Sporobolus, Elymus, Redfieldia, Calamovilfa,* and others.

The grasses are primarily a group of plants adapted to land habitats, and they are replaced in wet areas by sedges (family Cyperaceae) or rushes (family Juncaceae). A few specialized grass species, however, occur in or around water. The most widespread is the common reed, *Phragmites australis.* The genera *Leersia* and *Glyceria* are also usually confined to wet habitats. Rice (*Oryza sativa*) and wild rice (*Zizania aquatica*) have previously been noted as aquatic grasses. In many of the major groups of grasses, a few species adapted to moist or wet habitats are to be found. The most truly aquatic of all grasses are *Hydrochloa caroliniensis* and *Luziola fragilis.* Both of these species have bases rooted in the mud of ponds or streams. The weak, slender culms allow the foliage leaves and inflorescences to float on the water surface.

Aquatic and moist-habitat grasses

At the other ecological extreme, there are a wide variety of grasses adapted to extremely dry habitats. The genera *Aristida, Bouteloua, Muhlenbergia, Hilaria, Schismus,* and others occur commonly in desert areas of the American Southwest. The desert grasses are predominantly of two contrasting types. Some, such as the six-weeks grasses, are small, shallow-rooted, short-lived annuals, surviving unfavourable periods in the seed stage. The remainder are harsh perennials with narrow, fibrous, often rolled leaves, propagating largely by rhizomatous spread or by the basal growth of clumps, and are capable of remaining dormant for long periods during drought.

At least 30 percent of the land surface of the world was once grassland, although much of this area has now been put into cultivation. Soils that developed under temperate grasslands are uniquely fertile and form some of the most productive croplands. Grassland areas of great extent are found on all the continents of the world except Antarctica, and, even there, two of the three species of flowering plants that occur are grasses of the genus *Deschampsia.* (For descriptions, maps, and other material relating to these areas see GRASSLANDS.)

FORM AND FUNCTION

Vegetative characteristics. The adult grass plant consists of a jointed axis or stem, usually with solid nodes

(joints) and hollow internodes (stem regions between nodes). At each node a single leaf, with a sheathing base, is borne. The leaves are arranged distichously (in two vertical rows, 180° apart). Fibrous roots (called adventitious roots) may be borne at the basal nodes of the stems, forming the majority of the root system. Leafy, rooting portions of the stem (stolons) may creep along the surface of the earth. Beneath the soil surface, scaly, leafless stems (rhizomes) occur. Both stem types (stolons and rhizomes) serve for vegetative reproduction and food storage. Aerial stems, or culms, bear foliage leaves, each of which consists of a sheath (tubular sheathing portion), usually with overlapping edges, that surrounds the stem, and an elongated blade with parallel veins. A hairy or membranous collar, the ligule, arises at the juncture of the sheath and blade. At the base of the internode, inside of the midrib of the sheath, a bud is borne that may give rise to a branch. Between the branch and the main stem, a peculiar two-keeled modified leaf, the prophyllum, occurs. All other vegetative or floral structures of the grass plant may be homologized with one or another of the foregoing structures.

Horizontal creeping stems

Roots of grasses are usually of small diameter and are much branched, forming the so-called fibrous root system. Except in small, short-lived annual grasses, the primary root and other roots arising directly from the embryo are soon lost, and the remainder of the root system is entirely adventitious (growing from other than the usual place, such as on stems) from the lower stem nodes. In perennials, much of the root system is replaced each growing season.

Rhizomes are extremely important in the economy of many grasses, providing the principal means of reproduction, in addition to storing much food. Creeping leafy stems lying above the soil surface, the stolons, function similarly. Short, leafy basal shoots, without visible stems, are called innovations. Their leaf area may be important in furnishing forage or adding to the quality of lawns. Kentucky bluegrass (*Poa pratensis*), for example, has many such innovations, which are responsible for the density of the grass sward. Erect stems of grasses, which ultimately bear inflorescences (flower clusters), are called culms. The culm terminates in an inflorescence borne on a more or less exserted (extended beyond the rest of the plant) stalk or peduncle.

Inflorescences. Inflorescences are of many types, but panicles (much-branched flower clusters) are by far the most common. Grass flowers are extremely reduced in size and structure and are always borne in small, scaly-bracted spikes called spikelets. The spikelet, not the individual flower, is the unit of the grass inflorescence. The spikelet, which is basically homologous with a leafy stem bearing branches (the flowers), may be specialized in many diverse ways. Practical identification depends largely on characteristics of the spikelets and other parts of the inflorescence, although these characteristics have proved inadequate for the production of an evolutionary system of classification for the family Gramineae.

Importance of spikelet to grass identification

A simple type of spikelet is depicted in the accompanying Figure. The spikelet is usually supported on a stalk conventionally called a pedicel. The central axis, or rachilla, of the spikelet bears at its base two empty bracts (scalelike leaves), the glumes. These, like all the subsequent bracts of the spikelet, are distichous (arranged in two rows), as are the foliage leaves. Above the glumes each node of the rachilla bears a bract, the lemma, hidden inside of which a flower arises. At the base of the flower and actually on the flower branch is another bract, the palea, homologous with the prophyllum at the base of a vegetative branch. The palea is a two-keeled structure, somewhat H-shaped in cross section, having one vascular bundle near each keel but no midrib. The flower is enclosed between the lemma and the palea.

Flowers. Grass flowers are extremely small and simple, usually only a few millimetres long at the time of flowering. The perianth (surrounding flower parts, calyx and corolla) consists of a single cycle of small flattened or fleshy bodies, the lodicules. The number and structure of these vary in the different subfamilies of the grasses. In the bamboos and the subfamily Olyroideae, for example, three lodicules are present, two being placed between the ovary and the lemma and one between ovary and palea. The lodicules of bamboos are thin and flat, narrow and pointed, and they have several vascular bundles. In other grasses, two lodicules ordinarily are present, and these are located at the base of the lemma. In grasses of the subfamily Festucoideae they are pointed, nonvascular, and capable of swelling near the base. In the subfamilies Eragrostoideae and Panicoideae, the lodicules are square-tipped and fleshy and contain several vascular bundles. The sole surviving function of lodicules in grass flowers is to become engorged with water at the time of maturity of the flower, thereby forcing the lemma outward and allowing the anthers (pollen sacs) and stigmas (pollen receptors) to emerge.

In the great majority of grasses, three stamens (male reproductive structures) are present. Bamboos, however, characteristically have six stamens, and one genus (*Ochlandra*) has up to 120. The stamens show high evolutionary specialization toward wind pollination. The filaments (stalks of the anthers) are long and lax, permitting the anthers to dangle well out, exposed to air currents. The centrally attached anthers are proportionally very large and liberate enormous amounts of pollen. A field of rye (*Secale cereale*) during flowering, for example, can be seen to liberate clouds of yellow pollen. Plants of maize (*Zea mays*) often become noticeably dusted with pollen. Because of its abundance, grass pollen is a frequent cause of allergies, often but wrongly called rose fever. This vast excess of pollen is characteristic of wind-pollinated plants in general. The pollen, which is light and fluffy, consists of individual grains that are simple spheres with a single pore and no surface sculpturing. Each pollen mother cell produces four independent pollen grains.

Stamens and pollen specialized to wind pollination

The pistil (female reproductive organ) of grasses is composed of three united segments, or carpels, but produces only a single ovule and one seed. In bamboos and related grasses, three stigmas are usually present but in other grasses only two. In maize the two stigmas have become united, forming the silk. The stigmas are extremely feathery, serving to "comb" pollen out of the air. Despite the great waste of pollen, the combination of large anthers and extremely large stigmas has proved efficient enough to make the grasses one of the most successful and numerous families of plants.

The flower, the enveloping lemma and palea, and the usually adhering rachilla segment form a unit called a floret. Frequently the floret becomes the disseminule of the plant when the grain (caryopsis) matures.

Grass spikelets have multitudinous forms and structures, and no single description can encompass all of the variations known. Authorities generally agree that the spikelets of the ancestral grasses, typified by those of such present-day types as the bamboos, had bisexual (perfect) flowers that were borne in spikelets of several to many flowers. One common evolutionary tendency is for the loss of some or all of the upper flowers from an ancestrally many-flowered spikelet type, resulting in few-flowered or more commonly single-flowered spikelets. In the subfamily Panicoideae and in a few other forms, the lower flowers are reduced or absent, and only one perfect flower remains at the apex of the spikelet. In certain grasses, notably in the rice subfamily (Oryzoideae), the glumes have become reduced to mere ridges or a minute cup at the tip of the pedicel. In the crabgrasses (species of *Digitaria*) and related forms, the first glume has been lost or reduced to a minute rudiment.

Relatively few grasses have flowers that are unisexual. Frequently, the rudiments of the organs of the other sex may be detected in flowers of this type. Grasses with unisexual flowers may be monoecious (having both sexes on the same plant) or very rarely dioecious (the two sexes on different plants). Maize (*Zea mays*) is a prime example of monoecious grass. In maize, the male inflorescence (tassel) is situated at the tip of the stem, so that pollen may easily drift down upon the silks (stigmas) of the female spikelets, which are borne on the ear shoot below. The opposite situation occurs in wild rice (*Zizania*

Variations in sexuality

aquatica), in which the female spikelets are borne in a bunch at the tip of the stem, and the male spikelets are on lower branches of the same inflorescence. This situation encourages cross-pollination, since most of the pollen tends to drop downward. Grass species that are completely unisexual (dioecious) are much rarer.

Since the grasses are wind pollinated, there is a great wastage of pollen at any great distance from the parent plant. The difficulty in obtaining good seed set in unisexual grasses has apparently restricted the evolution of such types. The genus *Distichlis*, which occurs in maritime or saline habitats, and the common buffalo grass (*Buchloë dactyloides*) of the American Great Plains are examples of dioecious grasses. In both of these genera and other similar ones, the plants rely largely on vegetative reproduction by stolons or rhizomes, and highly effective sexual reproduction is not essential for the survival of the populations.

Shedding of spikelet parts

Spikelets vary as to the manner in which they or their parts are shed from the parent plant. In the more primitive grasses, the glumes remain on the plant, while the florets (or the grains) are shed individually or, more rarely, in a group. The shedding (disarticulation) of individual grains or florets has the obvious advantage of securing more even dispersal of the seed. In grasses with only one flower per spikelet, the entire spikelet may be shed as a unit, as in the rice subfamily (Oryzoideae) or the subfamily Panicoideae. Certain grasses that have evolved special means of seed dispersal may have clustered spikelets that disarticulate as a group from the plant, often surrounded by hooks or spines, as in the sandburs (species of *Cenchrus*).

In more primitive grasses, spikelets are laterally compressed—that is, they are flattened from the sides—so that the individual bracts are more or less boat shaped. Many common grasses, such as bluegrass (*Poa pratensis*), orchard grass (*Dactylis*), and the species of brome grass (*Bromus*), illustrate this type of structure. In the subfamily Panicoideae, however, the spikelets are flattened from the backs of the bracts (dorsally compressed), as in the genus *Panicum* or crabgrasses (species of *Digitaria*).

Fruit and seed. The fruit of the grasses is ordinarily a grain, or caryopsis. Technically, this differs from other small indehiscent (nonopening) fruits, such as the achene, by having the pericarp, or ovary wall, completely adherent to the seed. The seed coats themselves are usually crushed by the growth of the seed. In a few cases, the ovary wall may be shed, as in *Sporobolus* (dropseed grass) and *Eleusine* (goose grass), and in rare instances, especially in the bamboos, other types of fruits may occur.

The seed itself consists of a flattened, more or less boat-shaped embryo lying along one side of a large mass of starchy endosperm (stored food outside of the embryo). The bulk of the embryo consists of a fleshy body, the scutellum, which lies in contact with the endosperm and serves as a digestive and absorptive structure during germination. The true nature of the scutellum is debatable, but it has conventionally been called a cotyledon (seed leaf). The remainder of the embryo is attached on the external surface of the scutellum by its medial portion. The portion above the point of attachment is enclosed by a cylindrical sheath, the coleoptile, which protects the young shoot during germination. Within the coleoptile several embryonic leaves and the apical meristem (growing tip) of the stem are borne. The coleoptile and its contents may be attached directly to the face of the scutellum, or a short internode (embryonic neck or mesocotyl) may intervene. The lower portion of the embryo, which in some grasses is fused to the lower part of the scutellum, is a tubular sheath containing the primordium (the initial tissues) of the primary root. In certain subfamilies of the grasses a small flap of tissue, the epiblast, occurs on the external face of the embryo opposite the point of attachment. Certain features of the embryo have great value in determining the relationship of the grasses.

Leaves. The anatomy of grass leaves has been shown to be indicative of relationships within the order. The leaf epidermis, or outer "skin," in grasses is composed fundamentally of longitudinally elongated cells, usually with rippled walls. Interspersed with these are various specialized short cells, which are either suberized (corky) or impregnated with silica. The simplest of these are found in the subfamily Festucoideae, in which most of the isolated circular silicified cells are accompanied by crescent-shaped suberized cells. The other subfamilies have silicified cells in the form of double-bitted axheads, X's, collapsed quadrilaterals, or other complicated shapes. Frequently, such silicified cells occur in long rows over the vascular bundles. Many grass leaves bear elongated hairs, but these have little importance in determining relationships. Short, stiff, clawlike hairs also are common on leaf margins and may cause cuts when the leaves are handled. Plants of most of the grass subfamilies also bear specialized two-celled microhairs on the leaves. These may be linear and composed of two equal cells, the upper pointed, or they may be blunt or obtuse, with the upper cell smaller than the basal one. The type of microhair is specific to particular subfamilies. Such bicellular hairs are entirely lacking in the subfamily Festucoideae but are present in most of the other subfamilies. The shape of the complex of cells surrounding the stomates (air pores) in the epidermis is of importance in determining the subfamily to which a grass belongs. This group of cells consists, in grasses, of the two guard cells that surround the opening and two additional cells, the subsidiary cells, alongside each guard cell. In the subfamily Festucoideae, the general outline of this group of cells is rectangular, but in the other subfamilies, it varies from broadly oval to rhombic (diamond-shaped).

Internal leaf structure

The leaf cross section also reveals many features of importance in the classification of the family. Transverse sections of the leaf blade reveal a series of parallel vascular bundles, at intervals joined by small, transverse connecting bundles. Ordinarily each two large bundles are interspersed with one to several smaller and simpler ones. A typical bundle is surrounded by a thick-walled, woody sheath and often by an external layer of thin-walled, living cells (parenchyma). Internally the bundle contains a patch of phloem (food-conducting vascular tissue), oriented toward the bottom of the leaf, and several xylem vessels (water-conducting tubes). The number and nature of the bundle sheaths often indicate the subfamily to which the grass belongs. In the subfamily Eragrostoideae, for example, the outer sheath consists of large parenchyma cells that store starch, while in the subfamily Arundinoideae it consists of apparently empty cells. The green parenchyma cells filling the leaf are variously arranged. In grasses of the subfamily Festucoideae, the cells are loosely packed and without apparent order. In those of the subfamily Eragrostoideae, however, the cells are tightly packed and radially oriented in a ring around each vascular bundle. In several subfamilies, especially the Bambusoideae, Olyroideae, and Arundinoideae, the mesophyll cells (middle layers of green parenchyma cells of the leaf) often have folded-in walls or armlike processes (rachymorphous cells). Both the subfamilies Bambusoideae and Olyroideae and a few species of the subfamily Oryzoideae have large empty cells running crosswise in the leaf, between each two vascular bundles. These spindle-shaped, or fusoid, cells appear in the longitudinal sections of the leaf as I-shaped structures with large openings between them. Bulliform cells are enlarged epidermal cells, usually present between the vascular bundles on the upper leaf surface. Their function has been debated, but many authorities have claimed that they aid in folding or rolling the leaf blades at times of water stress, thereby reducing water loss.

Some grass leaf blades, especially those of the subfamilies Bambusoideae and Olyroideae, possess a constriction at the base of the blade that has been called a petiole (leafstalk) or pseudopetiole. Anatomically, it is merely a narrow portion of the leaf blade. Bamboo leaf blades are deciduous (*i.e.*, they fall) from the sheath at this point. Ordinarily, the petioles are only a few millimetres long but may reach several centimetres in *Zeugites* and as much as 13 centimetres (about five inches) in *Rattraya petiolata*.

Chromosome numbers. Since the first major compilation of grass cytology in 1931, interest in the correlation of grass chromosomes to the possible phylogenetic relationships within the family has been high, and probably the cytology of the family is better known than that of any other large family. Members of the subfamily Festucoideae, which includes the common grasses of temperate and cold climates, possess relatively large chromosomes, with the basic number of $x = 7$. All other subfamilies of the grasses possess much smaller chromosomes with basic number of $x = 5$, 6, 8, 9, 10, and 12. The bamboos, which are regarded as primitive from the standpoint of floral and spikelet structure, have $x = 10$ or 12. The evolutionary significance of the differences in chromosome number and size can only be guessed, and trying to assign one subfamily as parental to the others on the basis of cytology or morphology of living forms seems an exercise in futility.

Polyploidy among grasses

A high percentage of grasses are polyploid; that is, their somatic (body) cells possess more than the two sets of chromosomes typical of ancestral forms. Numerous species with 4, 6, 8, 10, 12, or more basic sets of chromosomes are known. This increase of chromosome number has occurred by natural means, probably after the accidental formation of nearly sterile interspecific hybrids (crosses between species). When the chromosome number of such a hybrid becomes doubled, the offspring (known as an allopolyploid) may be fertile and sometimes highly successful in nature. This spontaneous increase of chromosome numbers has been a major evolutionary method in the grasses and has undoubtedly aided greatly in fitting the members of the family to the exceedingly diverse climatic and soil conditions under which they live. Many of the most valuable agronomic grasses have had this type of origin. Examples of grass allopolyploids include timothy (*Phleum pratense*), orchard grass (*Dactylis glomerata*), wheat (*Triticum aestivum*), oats (*Avena sativa*), and many bluegrasses (*Poa* species).

Biochemical and physiological features. The grasses are notably free of alkaloids and similar compounds that abound in other monocot families, such as the Liliaceae and Amaryllidaceae. In a few instances, notably in the genus *Sorghum,* young or injured plants may contain amounts of cyanogenetic glycosides, which are organic compounds containing sugar and capable of liberating lethal amounts of hydrocyanic acid when the plants are grazed by livestock. Many grasses also may become infected with various species of ergot fungus (species of *Claviceps*), which renders the seed heads toxic.

A limited number of attempts have been made to determine relationships within the grass family by means of immune reactions similar to those used in blood typing. The results indicate that various genera formerly placed in the subfamily Festucoideae are more properly attached to the subfamily Eragrostoideae, as has been indicated also by numerous anatomical and cytological data.

EVOLUTION

Since the grasses are primarily herbaceous plants, they have not left an impressive fossil record. While claims have been made for the presence of grasses from the Cretaceous Period of some 65,000,000 to 136,000,000 years ago, all reasonably well identifiable fossil grass remains are from the Tertiary Period (beginning 65,000,000 years ago) or later. Vegetative remains of rhizomes and stem bases ascribed to the genera *Arundo* or *Phragmites* have been described from Europe. An excellent series of fossil grasses from the Great Plains of central North America was described in 1942. These fossils consist mostly of hardened fertile florets of species of grasses related to the living genera *Stipa, Oryzopsis, Piptochaetium, Nassella,* and *Setaria* or *Panicum*. They span the period from the early Miocene Epoch through the Pliocene (beginning about 26,000,000 years ago to about 2,500,000 years ago). Their nature, relationships, and associated plant and animal fossils indicate that they were inhabitants of open, dry plains. It is significant that the first appearance of these fossils of steppe-type grasses coincides closely with the evolution of ancestral horses having high-crowned teeth and strong jaws, suited to eating siliceous dry-land grasses. Their elongated legbones and sharp hooves enabled rapid running. These adaptations indicate that the animals evolved and lived in extensive grasslands. The presence of the grass fossils over a wide area indicates that rather modern type grasses and grasslands were well developed by the Miocene time and that the origin of the family must have occurred much earlier. It is likely that the ancestral grasses were delicate forest inhabitants and were incapable of fossilization.

Fossil record

Primitive grasses

Tropical-rain-forest areas of the world presently contain various peculiar, frequently broad-leaved grasses that are thought to be primitive in certain respects. The genus *Streptochaeta* of the American tropics is frequently mentioned as the most primitive living grass, yet its spikelets are highly atypical of those of other grasses and may themselves represent reduction from still earlier forms.

The relationships of the grasses to other monocotyledons has been the subject of much speculation. Earlier and undoubtedly erroneous systems of classification developed in the late 19th century were based upon the premise that the grasses are primitively simple; that is, that their small, simple, wind-pollinated flowers represent an early stage in floral evolution. A radically different system, based upon later anatomical studies of flowering plants, was presented early in the 20th century by the famous American botanist C.E. Bessey. His system, which is basic to most later arrangements, is based upon the idea that the primitive flowering plants are dicotyledons, such as the buttercups and their relatives (order Ranales in older systems, Ranunculales here) and the magnolias and their kin (order Magnoliales). The monocotyledons are regarded as having originated from the older dicotyledon orders. The lily relatives (order Liliales) represent a basic group of monocotyledons from which several lines of specialization arose. The grasses were regarded by Bessey as the apex of the wind-pollinated line of the monocotyledons, derived from the lily order by the loss of sepals and petals and various adaptations to wind pollination. Most recent authorities on plant phylogeny have followed somewhat similar arrangements of the monocotyledon families.

Recent concepts concerning relationships

A recent French writer states that the grasses arose in the Cretaceous Period from a lilylike ancestry. The rushes (family Juncaceae), which are wiry, grasslike plants with minute, six-parted green flowers, and the sedges (family Cyperaceae) are thought by this writer to have diverged from the same ancestry at about the same time. As treated here, a similar derivation of the order Poales from the order Liliales is suggested, but the sedges and rushes are separated on a different line of development. A recent system developed in the United States places both the family Cyperaceae (sedges) and the grass family (Poaceae or Gramineae) in one order, called the Cyperales. Both of these systems propose that the grasses are derived from a line based on the order Commelinales. This order contains insect-pollinated plants whose flowers have both calyx and corolla but lack nectar production. The absence of nectar is regarded as a first step in the derivation of wind-pollinated flowering plants from insect-pollinated ones, since the flowers have lost one of the structures (nectar glands) making flowers attractive to insects. Certain members of the order Commelinales are grasslike in aspect (*e.g.,* the family Xyridaceae), and all have sheathing leaf bases, somewhat similar to those of the grasses. The fruit in these groups, however, is a capsule, and the overall resemblances of these plants to the grasses are rather remote. More directly on the line of descent of the order Poales is the order Restionales. This order includes plants with wind-pollinated, reduced flowers and ovaries that have single ovules in their cavities. They are similar to the grasses in their pollen and in having sheathing leaf bases.

In the absence of a meaningful fossil record for the monocots, all of these varying systems are based upon interpretations of structure of living plants. While the details of the newer systems differ, all are alike in regarding the grasses and their kin as reduced, highly specialized plants rather than primitively simple ones.

CLASSIFICATION

Distinguishing taxonomic features. There is no single distinguishing trait of the order Poales. The presence of reduced flowers in scaly spikelets, two-ranked leaves with sheathing bases that have overlapping margins, hollow internodes, two bracts per flower, and a caryopsis as the fruit are the usual features used, but all of these are subject to variation. While the stems are ordinarily annual and herbaceous, they are long-lived and woody in the bamboos.

Plants similar to grass in appearance

Plants of three common families, the grasses (Gramineae), the sedges (Cyperaceae), and the rushes (Juncaceae), have a similar aspect and are frequently confused by casual observers. Species of the family Juncaceae are at once separable from the other two families by having flowers with a well-developed, six-parted perianth and a capsular, dehiscent fruit containing three to many seeds. Differences between the Gramineae and the Cyperaceae are summarized in the Table.

Comparison of Two Easily Confused Families, the Grasses and Sedges

character	grass family (Gramineae)	sedge family (Cyperaceae)
Leaf orientation	two rows	three rows
Leaf sheath	usually with overlapping edges	always with united edges
Stem structure	usually hollow, cylindrical	usually solid, triangular
Bracts per flower	two: lemma and palea	one: glume
Type of fruit	caryopsis	achene
Habitats	mostly dry land	mostly moist or wet areas

The features most used in modern classifications to subdivide the order Poales include the basic chromosome number and relative chromosome size, nature of specialized epidermal cells, internal leaf anatomy, root hair position on the epidermal cells, embryo structure, structure of the flower (including the number and nature of the lodicules), biochemistry and immunology, ecology, and distribution. The systems of classification produced by use of these and other characteristics must be supplemented with keys based upon easily observed superficial characteristics. Such aids to identification are now beginning to appear in recent works on grasses that use the modern evolutionary classifications as a basis for subdividing the family.

The classification used below represents a consensus of modern opinions regarding grass classification that has grown up in recent years. Since there is no modern worldwide treatment of the family, some segments of the classification stated here are adapted from different authors.

Annotated classification.

ORDER POALES

Anemophilous (wind-pollinated) herbs, rarely woody or of giant structure, with jointed, cylindrical, usually hollow stems; leaves 2-ranked (distichous), alternate, each composed of a sheathing base, a ligule, and a usually linear blade; root system fibrous, mostly adventitious; rhizomes or stolons frequent; inflorescences various (panicles, spikes, rames, or racemes) but always bearing distinctive subunits, the spikelets. Spikelets diverse but typically consisting of a central axis, the rachilla, bearing at its base a pair of empty bracts, the glumes, and above these 1 or more florets. The floret consists of an outer bract, the lemma, an inner, or secondary, bract, the palea, and the reduced flower included between them. Flowers small, typically consisting of 2 reduced perianth segments (the lodicules), 3 stamens with anthers attached at the centre, and a pistil fundamentally made up of 3 united carpels or segments but usually with 2 styles bearing feathery stigmas; ovule 1; fruit a caryopsis. Number of genera estimated at 600 to 700; species 6,000 to 10,000.

Family Poaceae or Gramineae

The only family of the order; it has the characteristics of the order.

Subfamily Bambusoideae (the bamboos)

Tropical and warm temperate woody perennial grasses; culms from diminutive to as much as 45 m, or 150 ft (rarely 90 m, or 300 ft), long, usually with hollow internodes and solid nodes, from less than 1 cm (0.5 in.) to more than 25 cm (10 in.) in diameter; propagation mostly by rhizomes; main culms ordinarily lacking foliage leaves but bearing deciduous sheaths; foliage leaves produced, after the cessation of culm growth, on small lateral branches; leaves with a short pseudopetiole between the sheath and blade; blades often deciduous at this point; leaf anatomy and epidermis "bambusoid" (*i.e.*, the long cells of the epidermis bear rows of warts; silica-bearing cells of various complex forms abundant; stomates, including the subsidiary cells, of rhombic form; bicellular microhairs pointed, the 2 cells about equal; cross-sectional anatomy of blades characterized by presence of large oval fusoid cells in the mesophyll, these appearing as large empty spaces in cross section); flowering usually rare, although some species (*e.g.*, *Chusquea subtessellata*) seem always to be in bloom. Flowering is often gregarious and periodic at long intervals of up to 60 years. Plants frequently monocarpic (flowering only once and then dying). Inflorescences terminal or axillary, paniculate, spicate, or racemose, sometimes of dense heads. The units of the inflorescence may be typical grass spikelets, or they may be "pseudospikelets," which are bracted structures concealing further lateral spikelets; spikelets with 1 to many florets; disarticulation usually between florets; lemmas usually awnless; flowers usually perfect, but sterile flowers frequent; lodicules 3, flat and with vascular bundles; stamens 3 or 6, rarely numerous; ovary of 3 carpels and with 3 stigmas; fruit a caryopsis, rarely berrylike or nutlike. Basic chromosome number $x = 10$ or 12. An expert on the subfamily states that there are 76 effectively published generic names and 1,000 species names but that many of the latter are to be regarded as synonyms. Taxonomically, the bamboos are the least known of all the grasses. Tropics and warm temperate regions of both hemispheres; a few species extending about as far north as the 40th parallel in North America.

Subfamily Olyroideae

Tropical herbaceous or semiwoody grasses having annual blooming; lacking the specialized woody perennial culms of the bamboos and their long-continued vegetative phase; leaves borne on the main culms, not on small lateral branches as in the bamboos; leaf blades often petiolate, often very broad and with conspicuous crossveins joining the longitudinal ones; leaf cross section "bambusoid," with large internal fusoid cells; spikelets 1-flowered, usually unisexual and the 2 kinds different, the pistillate (female) spikelets larger, often with a bony floret; disarticulation above or below the glumes; lodicules 3, like those of the bamboos; stamens rarely 2, usually 3 or 6; stigmas 3. Chromosomes small, with base numbers of $x = 11$ or 12. The members of this subfamily are a small group of tropical-forest grasses, apparently evolutionary relicts. About 12 genera, including 75 species, mostly in tropical America, a few in the Old World tropics.

Subfamily Festucoideae

Herbaceous grasses; spikelets from many- to single-flowered, the lemmas usually with 5 to many nerves (lines or ridges); flowers with 2 pointed, nonvascular lodicules with fleshy bases, 3 stamens, and ovary with 2 stigmas. Leaf anatomy simple, the vascular bundles with a single sclerenchyma sheath (made up of cells with thick walls), the mesophyll without special cellular arrangement; leaf epidermis simple, with few, simple specialized cells that are either solitary or paired, with 1 circular silicified cell and 1 lunate suberized (corky) cell together; stomatal complex (guard cells and subsidiary cells of stomates) rectangular in outline; bicellular hairs lacking; scutellum of the embryo lacking a separate basal limb; embryo neck (mesocotyl) lacking; epiblast present; 1st foliage leaf linear; chromosomes relatively very large, basic number $x = 7$. About 100 genera and 2,000 species. The group is widespread in both Eastern and Western hemispheres. The species are common in the temperate zones and constitute the entire grass flora of Arctic and high Alpine areas. In the tropics they are restricted to high mountains at upper elevations.

Subfamily Oryzoideae (the rice subfamily)

Spikelets 1-flowered, sometimes unisexual and then the 2 sexes borne in the same or separate inflorescences; spikelets reduced basally, with a single apical floret, the glumes reduced to mere ridges at the apex of the pedicel; basal sterile florets present in the genus *Oryza* and resembling glumes; spikelets laterally compressed or terete (cylindrical or slightly tapering); stamens 3, 6, or rarely 1; leaf epidermis complex with acuminate (sharp-pointed) bicellular hairs, rhombic or rounded stomates, complex silica bodies, often in rows; long cells often with rows of papillae (small nipple-shaped projections), as in the bamboos; leaf cross section with complex midrib, often with superposed vascular bundles; vas-

cular bundles with thin inner sclerenchyma sheath and a large-celled outer parenchyma sheath; mesophyll cells close-packed, with invaginated (inward-folded) walls; a few fusoid cells sometimes present near the major veins; embryos lacking basal limb of the scutellum and mesocotyl; epiblast present; plants mostly adapted to marshy or aquatic habitats. Chromosomes small, the base number $x = 12$ (15 in the genus *Zizania*). Genera 10, including about 35 species, found in warm temperate and tropical regions of both hemispheres.

Subfamily Arundinoideae (the reed subfamily)

Plants annual or perennial, often very tall; leaf epidermis sometimes with bicellular microhairs; silica cells of complicated forms; vascular bundles with a weak sclerenchyma sheath and outer parenchyma sheath lacking chloroplasts, mesophyll densely packed, the cell walls with invaginations; spikelets usually with several to many florets, disarticulating above the glumes; flowers usually bisexual, but some functionally dioecious types are known; lodicules 2; stamens 3; stigmas 2. Basic chromosome number $x = 11$ or 12, the chromosomes small. Genera 30, species about 450, distributed throughout the temperate and tropical regions of both hemispheres.

Subfamily Eragrostoideae

Mostly warm-climate grasses of small to moderate stature; leaf epidermis with rhombic stomates, bicellular microhairs short and obtuse, with a short upper cell; silica cells numerous, of complicated forms; leaf cross section complex, of nearly isolated units surrounding the veins and separated from the adjacent units by enlarged bulliform cells; bundle sheaths 2, the inner of small sclerenchyma cells, the outer of enlarged parenchyma cells, which store much starch and react to the iodine test; cells of the mesophyll small, mostly arranged compactly radially around the vascular bundles; spikelets usually disarticulating above the glumes, laterally compressed; florets 1 to many, sometimes the upper ones sterile; embryo with an embryonic neck and epiblast; chromosomes small, the base number $x = 6$, 9, or 10; inflorescences various, but often of 1-sided spikes or racemes. Genera about 80, including about 900 species. Distribution mostly in warm temperate and tropical regions of both hemispheres; plants especially prevalent on steppes and savannas.

Subfamily Panicoideae

Grasses of extremely diverse habits, primarily of warm climates and none Arctic or high Alpine. Culms herbaceous or rarely semiwoody, sometimes solid, as in *Zea* and *Sorghum*; leaf epidermis with acicular (needle-shaped) bicellular microhairs, rhombic stomates, and silica cells of complicated forms; leaf cross section with outer bundle sheath better developed than the inner and storing starch; mesophyll parenchyma cells somewhat radially arranged around the bundles; inflorescences are panicles or various aggregations of racemes, rames, or spikes; disarticulation below the glumes; spikelets dorsally compressed with a single perfect terminal floret and a sterile or staminate floret below it. Flower with 2 fleshy, truncate vasculated lodicules; embryo with free lower limb of scutellum and embryonic neck but no epiblast. Basal reduction of the spikelet is universal, and frequently the 1st glume is reduced or absent. Seedlings with broad, short, horizontally oriented 1st foliage leaf. Basic chromosome numbers $x = 9$ or 10, but numbers of 3, 4, 5, 7, 8 and 13 have been reported rarely; chromosomes small. The 2 tribes of the subfamily Panicoideae (Paniceae and Andropogoneae) contain 130 genera and about 2,500 species. The grasses of this subfamily are abundant and dominant in the steppe and savanna vegetation of the tropics and subtropics of both hemispheres. Numerous species extend well up into the temperate zones.

Critical appraisal. Ideas as to the classification of the grasses have undergone a veritable revolution in recent years, unparalleled in any other group of angiosperms.

Basis of early classification systems

The earlier arrangements were based primarily upon the nature of the spikelet, with secondary emphasis given to the form and structure of the inflorescence. Doubts arose as to the validity of this system when the first major compilation of chromosome numbers of the grasses was published in 1931, and the family was divided on the basis of chromosome number and size. The groupings thus defined ran counter to the system of classification current at that time, however, and were not taken up. Shortly thereafter, extensive studies were made on the grass epidermis. The wide variety of specialized cells of the epidermis was used as a basis for dividing the family into new groupings. By 1936 three subfamilies, the Bambusoideae, the Festucoideae, and the Panicoideae, were recognized. The grasses currently grouped in the subfamily Eragrostoideae were uncomfortably wedged into the subfamily Panicoideae. This system of classification, while no longer entirely accepted as to its taxonomic conclusions, was the first to incorporate such features as leaf epidermis, cross-sectional leaf anatomy, chromosome number and size, nature of the embryos and seedlings, starch grains, physiology, and ecology. Publications based upon such data led to a greatly increased interest in the construction of a more truly phylogenetic classification of the grasses. The status of grass classification was the subject of a symposium at the International Botanical Congress in Montreal in 1959. Numerous papers concerning the overall and partial classification of the group have been published since that time. More or less inclusive classifications have been published by grass evolutionists for temperate North America, for tropical Africa, and for the entire world. While these systems are not completely in accord with one another, they all show general similarities.

All of the new systems recognize the bamboos as a separate subfamily. This appears to be entirely justified, since all of this group have woody, perennial stems. In many respects, they are a highly specialized group, although their flower structure is primitive in the order. There is less general accord as to the subfamily Olyroideae, whose members are similar to the bamboos in many anatomical and cytological aspects but lack the woody stems.

The subfamily Festucoideae of the older standard system of classification has suffered the greatest remodelling. As it was originally construed, this group was united only by spikelet characteristics. It was shown by later investigators that the members differed greatly in numerous anatomical characteristics, in their basic chromosome numbers and size, and in their ecological and geographical relationships. The removal of certain genera to the subfamilies Bambusoideae, Eragrostoideae, Arundinoideae, and Oryzoideae has left the subfamily Festucoideae much more homogeneous.

The subfamily Panicoideae, which had long been recognized on the basis of spikelet structure, continues to be accepted in the original circumscription, although with some internal rearrangements.

The subfamily Arundinoideae is the least well characterized of the recently accepted divisions and will probably undergo further revision. A few isolated genera of grasses are so unusual that placement of them in any system of classification is problematical, and they can be only arbitrarily assigned to a subfamily.

Basis of recent classification systems

In general, the new systems of classification recognize about six or seven major groups (subfamilies or series), based upon numerous criteria, including microscopic anatomy and cytology. One major work on grass classification published in 1954 is at considerable odds with those previously discussed. This system uses nine subfamilies, but their constitution is very different from that proposed by later authors. The large subfamily Festucoideae of this system is even more inclusive than that of the previous standard system containing such aberrant genera as *Phyllorachis, Pariana, Thysanolaena, Arundinella, Pappophorum, Phragmites,* and others, which are placed in other subfamilies by more recent authors. The subfamily Arundinoideae, containing the giant reed grasses and other grasses, is not recognized, but, on the other hand, the bluestem grasses (*Andropogon* and its relatives) and the maize relatives are separated into a new subfamily, the Andropogonoideae. This radically different classification has not been favourably received by agrostologists, but the comprehensive nature of the work and its keys to genera make it useful.

The work of the last 40 years on grass cytology, anatomy, physiology, biochemistry, and ecology has succeeded in establishing the general outlines of a new and more informative system of classification. It seems reasonably clear that paleobotany, on the other hand, will do little or nothing to either confirm or contradict the newer classifications.

Future cytological and anatomical research is still need-

ed to fix the positions of many anomalous genera. Since the 1954 proposal of a complete general system, no attempt has been made to provide a worldwide treatment of the grasses at the generic level, using a modern system of classification. Such a treatment would be an excellent skeleton on which to place the accumulating body of data on grasses. There are no world monographs of the great majority of grass genera, even on the level of gross morphology. The availability of such monographs would be a great stimulus to the study of the family in the various parts of the world.

The wet forests of the tropics contain many rare or isolated primitive grasses, especially members of the subfamily Olyroideae. Study of these, which may reveal further clues as to grass relationships, is urgently needed before the surging population pressures of the tropical lands completely wipe out their habitats.

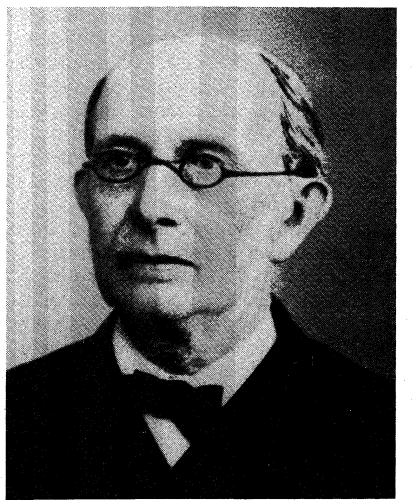
BIBLIOGRAPHY. AGNES ARBER, *The Gramineae: A Study of Cereal, Bamboo, and Grass* (1934), a comprehensive study of grass anatomy; J.W. BEWS, *The World's Grasses: Their Differentiation, Distribution, Economics and Ecology* (1929), a standard reference with keys but no descriptions, now obsolescent; W.E. BOOTH, *Agrostology* (1964), discussions of the history of agrostology, vegetative and reproductive morphology, cytology, evolution, classification, nomenclature, culture, diseases, and grass identification, with extensive citation of literature; E.G. CAMUS, *Les Bambusées*, 2 vol. (1913), the only taxonomic work on bamboos claiming to be complete, now obsolescent; AGNES CHASE, *First Book of Grasses: The Structure of Grasses Explained for Beginners*, 3rd ed. (1959); F.W. GOULD, *Grass Systematics* (1968), a modern text including discussions of structure, relationships, utility of grasses, reproduction and cytogenetics, modern grass classification and its bases, with keys to genera and descriptions and illustrations of genera of U.S. grasses; A.S. HITCHCOCK, *Manual of the Grasses of the United States, Misc. Publs. U.S. Dep. Agric. 200*, 2nd ed. rev. by AGNES CHASE (1951), a standard reference abundantly illustrated, with full descriptions, keys, and distribution maps; H. JACQUES-FELIX, *Les Graminées (Poaceae) d'Afrique Tropicale*, vol. 1, *Généralités, classification, description des genres* (1962), valuable for its modern approach to the tropical grasses, includes much anatomical and cytological information; F.A. MCCLURE, *The Bamboos: A Fresh Perspective* (1966), an extended treatment (not primarily taxonomic) of seedlings, mature vegetative plants, flowering and fruiting, classification of bamboos, bamboo culture and propagation; R. PILGER, "Das System der Gramineae," *Bot. Jb.*, 76:281–384 (1954), the most recent attempt at a complete classification of the Gramineae, but the system is regarded as containing many unacceptable features; R.W. POHL, "The Grasses of Iowa," *Iowa St. J. Sci.*, 40:341–566 (1966), keys and notes, illustrations, and distribution maps; *How to Know the Grasses*, rev. ed. (1968), a guide for collection and study with illustrated keys to common North American grasses, including distributions, notes, and descriptions; R.J. POOL, *Marching with the Grasses* (1948), a popular treatment of the grass family, largely from an agricultural standpoint; G.L. STEBBINS and B. CRAMPTON, "A Suggested Revision of the Grass Genera of Temperate North America," *Recent Advances in Botany*, 1:133–145 (1961), the first formulation of an entirely modern system of grass classification; UNITED STATES DEPARTMENT OF AGRICULTURE, *Grass: Yearbook of Agriculture for 1948* (1948), a sourcebook of much general information on grasses, legumes, and grassland agriculture.

(R.W.P.)

Pobedonostsev, Konstantin Petrovich

Russian civil servant, tutor and adviser to tsars Alexander III and his son Nicholas II, lay leader of the Russian Orthodox Church, and political philosopher, Konstantin P. Pobedonostsev was the dominant figure in the turn to reaction that followed the assassination of Alexander II in 1881. Nicknamed the "Grand Inquisitor," he came to be the symbol of Russian monarchal absolutism.

The youngest son of a Russian Orthodox priest who was also professor of Russian literature at Moscow University, Pobedonostsev was born in Moscow on May 21, 1827. He was educated at home and at the Oldenburg School of Law in St. Petersburg (now Leningrad), from 1841 to 1846. His adult life was devoted to service at the centre of the Russian state bureaucracy, beginning in the Moscow office of the senate. The publications he produced in his spare time there on the history of Russian civil law and institutions led to his being invited in 1859 also to lecture on civil law in Moscow University. His courses were so distinguished in organization, learning, and clarity that in 1861 Alexander II asked him to serve also as a tutor to his sons during the time they spent in Moscow each year. At the same time, he was an important contributor to the 1864 reform of the Russian judicial system. In 1865 he accepted the Tsar's invitation to leave Moscow University, the senate, and his beloved Moscow to serve as a tutor to the Tsar's sons and their families in St. Petersburg. He was appointed to the senate in 1868, to the council of state (a high advisory body) in 1872, and in 1880 to the director generalship, or chief administrative position, of the Most Holy Synod of the Russian Orthodox Church, a position he held until the fall of 1905 and that was the foundation of his political influence and reputation.



Pobedonostsev.
H. Roger-Viollet

Political philosophy

Pobedonostsev was valued as an adviser by the two tsars whom he had tutored, Alexander III and Nicholas II. He considered man by nature "weak, vicious, worthless, and rebellious," denounced the 18th-century Enlightenment view of the perfectibility of man and of society, and therefore strongly supported paternalistic and authoritarian government. He looked upon each national society as an independent organic growth, based on the land, the family, and the national church, and regarded the maintenance of stability as the principal purpose of government. He sought, therefore, to defend Russia and the Russian Orthodox Church against all rival religious groups, such as the Old Believers, who had separated from the main Russian Orthodox Church, Baptists, Catholics, and Jews, against the national minorities, and against political and intellectual challenges. As a Russian nationalist, he defended Russian rule over the various minority groups and supported their Russianization, especially that of the minorities living in the western border areas. As lay head of the church, he promoted rapid expansion of primary education in parish schools because he saw it, with its emphasis upon religion, the Russian language, and song, as a strong bulwark of the autocracy. He sought to keep each person in that station in life into which he had been born and to restrict higher education to the upper classes and exceptionally talented. He tried also to prohibit and to banish all foreign influences, especially ideas concerning constitutional and democratic government, except for the "aristocracy of intellect," who needed such information in order to govern well.

Influence

Pobedonostsev had great influence in 1881, immediately after the assassination of Alexander II, when he persuaded Alexander III to reject the so-called Loris-Melikov constitution that was designed to bridge the gap between the government and the leading elements of society. He also influenced the government's reactionary domestic policies during the 1880s but exercised little authority during the last 15 years of his life. He exerted almost no

influence on Russian foreign policy, largely because he believed Russia should isolate itself from world affairs, a belief greatly strengthened by the disappointment that followed his brief passionate advocacy of Pan-Slavist expansionism and the unsuccessful war against the Turks in 1877–78. His role, however, was exaggerated during his lifetime by critics of the regime and since then by historians, largely because his personality, appearance, and known views superbly qualified him as the symbol of a system of government deeply unpopular among many educated Russians and among all liberals and radicals.

Pobedonostsev was a childless, dry, reserved, and deeply pessimistic ascetic with almost no close friends, except for the novelist Dostoyevsky, who died in 1881. At the same time, he was a man of immense learning and scholarship, widely respected among foreign diplomats. He read and spoke most European languages, was deeply familiar with the great body of European and American literature and philosophy, and spent many of his vacations in western Europe—although he strongly supported censorship and tight controls for other Russians. Especially after 1890 he was convinced the regime would be overthrown by revolution. While he read Marx and other socialists, he saw constitutionalism and democracy as the enemies and ignored the various Russian Socialist groups. He also sought to introduce American techniques into Russian agriculture and supported rapid industrialization, which began in the early 1890s. He died in St. Petersburg on March 23 (March 10, old style), 1907.

BIBLIOGRAPHY. R.F. BYRNES, *Pobedonostsev: His Life and Thought* (1968), a full biography and critical analysis based in part on Russian archival materials; *Reflections of a Russian Statesman* (1898), an incomplete translation of Pobedonostsev's best known book, *Московский сборник* (1896); *Письма Победоносцева к Александру III*, 2 vol. (1925–26), a valuable collection of his letters to Alexander III and to Nicholas II as well; *L'Autocratie russe, Constantine Pobédonostsev: Mémoires politiques, correspondance officielle et documents inédits* (1927), an incomplete but still valuable translation of *К. П. Победоносцев и его корреспонденты* (1923), collection of Pobedonostsev's memoranda and of letters to him, especially from Alexander III and Nicholas II.

(R.F.B.)

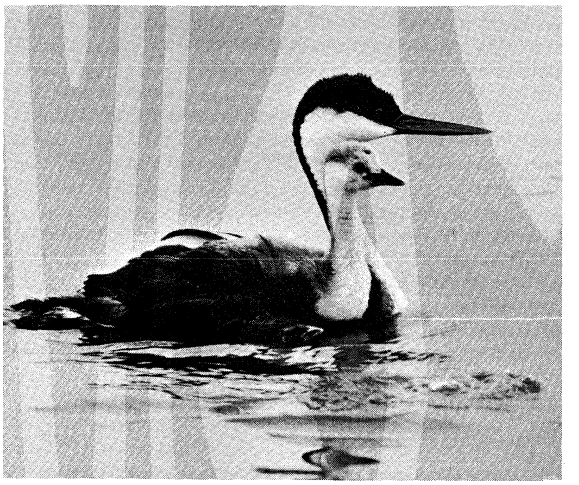
Podicipediformes

The order Podicipediformes includes only the grebes, foot-propelled diving birds best known for the striking courtship displays of some species and for the silky plumage of the underparts, which formerly was much used in millinery. The speed with which grebes can submerge has earned them such names as water-witch and helldiver, while the position of the feet near the tail is responsible for the early English name arsefoot, from which the family name, Podicipedidae, was derived.

Adult grebes range in weight from less than 150 grams (five ounces) to more than 1.4 kilograms (about three pounds) and in total length from 21 to 73 centimetres (eight to 29 inches). They vary largely in bill shape and ornamentation of the head. The group is found on all continents and many island groups, but it is best represented in temperate regions. Seven species each are found in North and South America, five in Eurasia, and three each in Africa and Australia. The species range from conspicuous and gregarious to solitary and skulking.

Natural history. Both parents share in nest building, incubation, and care of the young. The long pair bond that makes this possible is formed and strengthened by elaborate courtship displays, including ritualized preening, head shaking, diving, weed carrying, and rapid water treading with the body in a nearly vertical position. These displays may be combined into complex ceremonies such as the discovery ceremony of the great crested, rednecked, horned, eared, and related species or the rushing display of the western grebe. In nearly all courtship ceremonies the roles of the sexes are interchangeable. The same is true of the precopulatory displays, and reverse mounting has been reported for all species that have been thoroughly studied. Courtship feeding is known only in the western grebe (*Aechmophorus occidentalis*), the male of which feeds the female. Grebe vocalizations include advertising calls, copulation trills, "conversational" notes, and duetting trills. In the courtship of more secretive species, such as the pied-billed grebe (*Podilymbus podiceps*) and the dabchicks (a name given to several of the smaller grebes), vocalizations are relatively more important than visual displays.

Courtship displays



Western grebe (*Aechmophorus occidentalis*). The young are carried upon the parents' backs for protection and warmth.
Willis Peterson

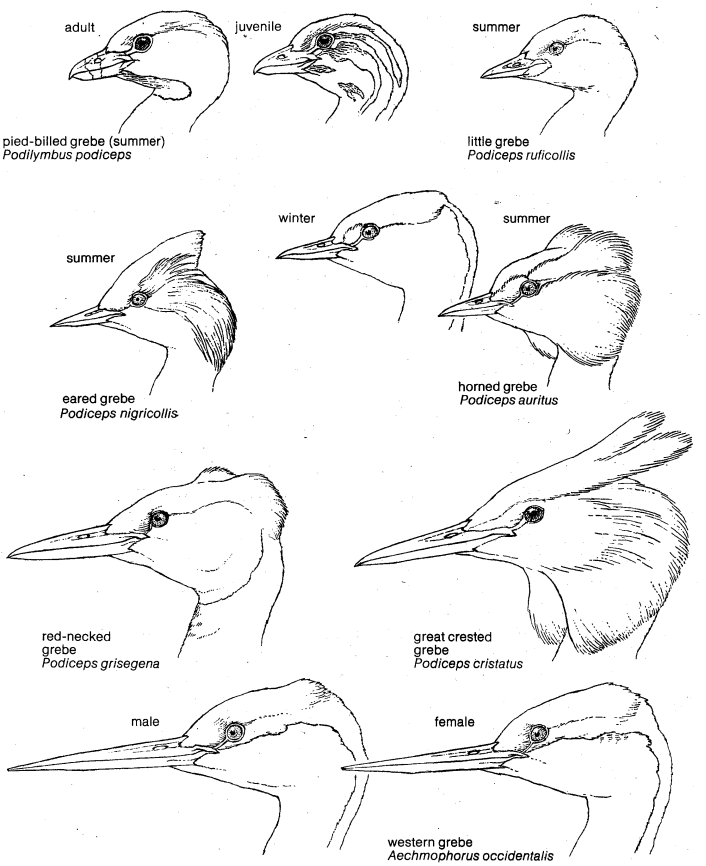
Following pair formation, grebes build one or more floating platforms of aquatic vegetation. On these mating takes place and the three to seven unmarked chalky, white eggs are laid. On leaving the nest, adults often cover the eggs with the soggy nest material, and, by the end of the incubation period (three to four weeks), the eggs are strongly stained. The nest is usually deserted after the young hatch, and the small young spend most of their time on the swimming parents' backs, which they reach by clambering up the adult's foot. On hatching, the downy young are marked with bold longitudinal stripes, though these are rarely visible on the gray young of the western grebe. Remnants of the head stripes are carried over into the juvenile plumage. In addition to the stripes, the young have either one or more patches of rufous down or a bare spot on the crown. The skin of the latter changes from pink to red when the young become excited.

In winter plumage most grebes are brown, black, or gray above and white or light brown below. In summer plumage, rufous, buff, black, or white markings or elongated plumes are found, chiefly on the head and neck. In this plumage, the males tend to be more brightly coloured and to have somewhat longer plumes than the females. The males also tend to be larger and larger billed than the females.

Grebes' wings are short and the wing bones small in diameter. Some species nesting in cold regions are migratory. Major flights are usually made at night. One species, the short-winged grebe (*Rollandia micropterum*) of Lake Titicaca, between Peru and Bolivia, is flightless; two others, Taczanowski's grebe (*Podiceps taczanowskii*) of Lake Junín, Peru, and the giant pied-billed grebe *Podylimbus gigas*) of Lake Atitlán, Guatemala, have somewhat reduced wings and fly little, if at all.

While on the surface, grebes usually swim with alternate strokes of the feet, whereas under water, the feet move simultaneously. Between recovery and power strokes, the feet are turned 90° so that the broad side can be used for propulsion; just before the recovery stroke, the feet are rotated in the opposite direction. While resting, grebes keep one or both feet shipped under the wings, and the head is kept forward with the bill under the neck (the "pork-pie" posture).

Grebes breed on still or slow-moving bodies of fresh water. The dabchicks and the pied-billed grebe are most numerous on small waters with much emergent vegetation, the western grebe on large bodies of open water. Some species winter on sheltered coastal waters. Grebes'



Heads of typical grebes.
Drawing by R. Keane

diets consist almost entirely of animal matter. The western grebe feeds largely on fish, the eared grebe (*Podiceps nigricollis*) on small invertebrates, and the pied-billed grebe takes many heavy-bodied crustaceans. Other species, such as the horned and red-necked grebes (*Podiceps auritus* and *P. grisegena*), have more varied diets. Grebes are noted for swallowing feathers, which form a plug in the pyloric pocket of the stomach and effectively filter material passing to the intestine. Adults feed feathers to the young, establishing this plug shortly after hatching. Grebes feeding on invertebrates containing much chitin regurgitate pellets rather frequently; fish eaters tend to maintain masses of feathers in the main portion of the stomach, presumably to hold fish bones until they are digested.

Feather-eating

Grebes can be distinguished by their lobed toes with flat nails, greatly reduced tail feathers, and satiny breast feathers. The last is due to the loose structure of the feathers and to spirally coiled barbules (the secondary branches of the feathers), which lie parallel to the barbs (the primary branches). As with most diving birds, the conspicuous features that distinguish species are found on the head and figure prominently in courtship.

Classification and paleontology. Grebes comprise a single family, Podicipedidae, of the order Podicipediformes. There are approximately 18 species, usually placed in five genera: *Aechmophorus* (the western grebe), *Podiceps* (most species), *Podilymbus* (the pied-billed grebes), *Rollandia* (Rolland's and short-winged grebes), and *Tachybaptus* (dabchicks). The order and family are defined on the basis of structural features, especially the form of the nostrils, the absence of certain processes in the skull, and the absence of the ambiens muscle in the leg. The knee process consists of a large pyramidal patella (kneecap) and a large projection on the tibia (the main bone of the lower leg). The wing is diastataxic (*i.e.*, there is a space where the fifth secondary feather is located in many birds), with 12 primaries, seven of which are attached to the metacarpus (midhand). The oil gland is tufted and has two openings. Only the left carotid artery is present. The intestinal caeca (blind pouches) are small.

Grebes are highly specialized diving birds with no obvious relatives, living or fossil. Statements to the effect that they are "primitive" or "reptilian" are erroneously based on a superficial resemblance to the toothed Cretaceous Hesperornithiformes, themselves very specialized forms. The earliest known fossil grebe, *Podiceps oligoceanus* from lower Miocene deposits (about 25,000,000 years old) in Oregon, is essentially like modern species.

The names *Colymbus*, Colymbidae, and Colymbiformes were formerly applied to the grebes in North America and to the loons (*Gavia*) in Europe.

BIBLIOGRAPHY. R.S. PALMER (ed.), *Handbook of North American Birds*, vol. 1 (1962), provides an up-to-date summary of information on North American species; K.E.L. SIMMONS, "Studies on Great Crested Grebes," *Avicult. Mag.*, 61:3–13, 93–102, 131–146, 181–201, 235–253, 294–316 (1955), is a classic study of grebe behaviour; R.W. STORER, "The Behavior of the Horned Grebe in Spring," *Condor*, 71:180–205 (1969), contains comparative material on several species, and "The Patterns of Downy Grebes," *Condor*, 69:469–478 (1967), is concerned with the colour patterns of grebe chicks. The same author's "Courtship and Mating Behavior and the Phylogeny of the Grebes," *Proc. 13th Int. Orn. Congr.*, pp. 562–569 (1963), and "Evolution in the Diving Birds," *Proc. 12th Int. Orn. Congr.*, pp. 694–707 (1960), provide insight into the evolution and classification of grebes.

(R.W.St.)

Poe, Edgar Allan

The creator of the American Gothic tale and of the whole detective fiction genre, Edgar Allan Poe cultivated the literature of mystery and the macabre to an extent unknown before. His tales of mystery and detection initiated the modern detective story, and the other-world atmosphere in his tales of horror is unrivalled in American literature.



Poe, oil portrait by S.S. Osgood, 1845.
By courtesy of The New-York Historical Society

Life. Edgar Allan Poe was born January 19, 1809, in Boston, the son of the English-born actress Elizabeth Arnold Poe and David Poe, Jr., an actor from Baltimore, Maryland. After his mother died in Richmond, Virginia, in 1811, he was taken into the home of John Allan, a Richmond merchant (presumably his godfather), and of his childless wife. He was taken to Scotland and England (1815–20), where he was given a classical education that was continued in Richmond. For 11 months in 1826 he attended the University of Virginia, where he studied Greek, Latin, French, Spanish, and Italian; but his gambling losses at the university so incensed his guardian that he refused to let him continue, and Poe returned to Richmond to find his sweetheart, (Sarah) Elmira Royster, engaged. He went to Boston where in 1827 he published a pamphlet of youthful Byronic poems, *Tamerlane and Other Poems*, some of which allude to Elmira. Poverty

forced him to join the Army under the name of Edgar A. Perry, but on the death of Poe's foster mother, John Allan purchased his release from the Army and helped in getting him an appointment to the U.S. Military Academy at West Point. Before going, Poe published a new volume at Baltimore, *Al Aaraaf, Tamerlane and Minor Poems* (1829), showing the influence of Milton and Thomas Moore, as well as of a local poet, E.C. Pinkney. He successfully sought expulsion from the academy, where he was absent from all drills and classes for a week. He proceeded to New York City and brought out a volume of *Poems*, containing several masterpieces, some showing the influence of Keats, Shelley, and Coleridge. He then returned to Baltimore, where he began to write stories. In 1833 his "MS. Found in a Bottle" won $50 from a Baltimore weekly, and by 1835 he was in Richmond as editor of the *Southern Literary Messenger*. There he made a name as a critical reviewer and married his young cousin Virginia Clemm, who was only 13. Poe seems to have been an affectionate husband and son-in-law. He wrote a tribute "To My Mother" to his aunt, Mrs. Clemm, in 1849.

Editor and critical reviewer

Poe was dismissed from his job in Richmond, apparently for drinking, and went to New York. Drinking was in fact to be the bane of his life. To talk well in a large company he needed a slight stimulant, but a glass of sherry might start him on a spree; and although he rarely succumbed to intoxication, he was often seen in public when he did. This gave rise to the conjecture that Poe was a drug addict, but according to medical testimony he had a brain lesion. While in New York in 1838 he published a long prose narrative, *The Narrative of Arthur Gordon Pym*, combining (as so often in his tales) much factual material with the wildest fancies. It is considered one inspiration of Melville's *Moby Dick*. In 1839 he became co-editor of *Burton's Gentleman's Magazine* in Philadelphia. There a contract for a monthly feature stimulated him to write "William Wilson" and "The Fall of the House of Usher"—stories of supernatural horror. The latter contains a study of a neurotic now known to have been an acquaintance of Poe, not Poe himself.

Later in 1839 his *Tales of the Grotesque and Arabesque* appeared (dated 1840). He resigned from *Burton's* about June 1840 but returned in 1841 to edit its successor, *Graham's Lady's and Gentleman's Magazine*, in which he printed the first detective story, "The Murders in the Rue Morgue." In 1843 his "The Gold Bug" won a prize of $100 from the Philadelphia *Dollar Newspaper*, which gave him great publicity. In 1844 he returned to New York, wrote the "Balloon Hoax" for the *Sun*, and became subeditor of the *New York Mirror* under N.P. Willis, thereafter another lifelong friend. In the *New York Mirror* of January 29, 1845, appeared, from advance sheets of the *American Review*, his most famous poem, "The Raven," which gave him national fame at once. Poe then became editor of the *Broadway Journal*, a short-lived weekly, in which he republished most of his short stories, in 1845. During this last year the now forgotten poet Frances Sargent Locke Osgood pursued Poe. Virginia did not object, but "Fanny's" indiscreet writings about her literary love caused great scandal. His *The Raven and Other Poems* and a selection of his *Tales* came out in 1845, and in 1846 Poe moved to a cottage at Fordham (now part of New York City), where he wrote for *Godey's Lady's Book* (May–October 1846) on the "Literati of New York"—gossipy sketches on personalities of the day, which led to a libel suit.

National fame

His wife, Virginia, died in January 1847. The following year Poe went to Providence, Rhode Island, to woo Sarah Helen Whitman, a poet. There was a brief engagement. Poe had close but platonic entanglements with Annie Richmond and with Sarah Anna Lewis, who helped him financially. He composed poetic tributes to all of them. In 1848 he also published the lecture "Eureka," a transcendental "explanation" of the universe, which has been hailed as a masterpiece by some critics and as nonsense by others. In 1849 he went south, had a wild spree in Philadelphia, but got safely to Richmond, where he finally became engaged to Elmira Royster, by then the widowed Mrs. Shelton, and spent a happy summer with only one or two relapses. He enjoyed the companionship of childhood friends and an unromantic friendship with a young poet, Susan Archer Talley.

Poe had some forebodings of death when he left Richmond for Baltimore late in September. There, after toasting a lady at her birthday party, he began to drink heavily. The indulgence proved fatal, for Poe had a weak heart. He died on October 7, 1849, and was buried in Westminister Presbyterian churchyard in Baltimore.

Early death

Appraisal. Poe's work owes much to the concern of romanticism with the occult and satanic. It owes much also to his own feverish dreams, to which he applied a rare faculty of shaping plausible fabrics out of impalpable materials. With an air of objectivity and spontaneity, his productions are closely dependent on his own powers of imagination and an elaborate technique. His keen and sound judgment as appraiser of contemporary literature, his idealism and musical gift as a poet, his dramatic art as a storyteller, considerably appreciated in his lifetime, secured him a prominent place among universally known men of letters.

The outstanding fact in Poe's character is a strange duality. The wide divergence of contemporary judgments on the man seems almost to point to the coexistence of two persons in him. With those he loved he was gentle and devoted. Others, who were the butt of his sharp criticism, found him irritable and self-centred and went so far as to accuse him of lack of principle. Was it, it has been asked, a double of the man rising from harrowing nightmares or from the haggard inner vision of dark crimes or from appalling graveyard fantasies that loomed in Poe's unstable being?

Much of Poe's best work is concerned with terror and sadness, but in ordinary circumstances the poet was a pleasant companion. He talked brilliantly, chiefly of literature, and read his own poetry and that of others in a voice of surpassing beauty. He admired Shakespeare and Alexander Pope. He had a sense of humour, apologizing to a visitor for not keeping a pet raven. If the mind of Poe is considered, the duality is still more striking. On one side, he was an idealist and a visionary. His yearning for the ideal was both of the heart and of the imagination. His sensitiveness to the beauty and sweetness of women inspired his most touching lyrics ("To Helen," "Annabel Lee," "Eulalie," "To One in Paradise") and the full-toned prose hymns to beauty and love in "Ligeia" and "Eleonora." In "Israfel" his imagination carried him away from the material world into a dreamland. This Pythian mood was especially characteristic of the later years of his life.

More generally, in such verses as "The Valley of Unrest," "Lenore," "The Raven," "For Annie," and "Ulalume" and in his prose tales his familiar mode of evasion from the universe of common experience was through eerie thoughts, impulses, or fears. From these materials he drew the startling effects of his tales of death ("The Fall of the House of Usher," "The Masque of the Red Death," "The Facts in the Case of M. Valdemar," "The Premature Burial," "The Oval Portrait," "Shadow"), his tales of wickedness and crime ("Berenice," "The Black Cat," "William Wilson," "Imp of the Perverse," "The Cask of Amontillado," "The Tell-Tale Heart"), his tales of survival after dissolution ("Ligeia," "Morella," "Metzengerstein"), and his tales of fatality ("The Assignation," "The Man of the Crowd"). Even when he does not hurl his characters into the clutch of mysterious forces or onto the untrodden paths of the beyond, he uses the anguish of imminent death as the means of causing the nerves to quiver ("The Pit and the Pendulum"), and his grotesque invention deals with corpses and decay in an uncanny play with the aftermath of death.

Prose tales

On the other side, Poe is conspicuous for a close observation of minute details, as in the long narratives and in many of the descriptions that introduce the tales or constitute their settings. Closely connected with this is his power of ratiocination. He prided himself on his logic and carefully handled this real accomplishment so as to impress the public with his possessing still more of it than he had; hence the would-be feats of thought reading, problem un-

ravelling, and cryptography that he attributed to his Legrand and Dupin. This suggested to him the analytical tales, which created the detective story, and his science fiction tales.

Duality in his art

The same duality is evinced in his art. He was capable of writing angelic or weird poetry, with a supreme sense of rhythm and word appeal, or prose of sumptuous beauty and suggestiveness, with the apparent abandon of compelling inspiration; yet he would write down a problem of morbid psychology or the outlines of an unrelenting plot in a hard and dry style. In Poe's masterpieces the double contents of his temper, of his mind, and of his art are fused into a oneness of tone, structure, and movement, the more effective, perhaps, as it is compounded of various elements.

As a critic, Poe laid great stress upon correctness of language, metre, and structure. He formulated rules for the short story, in which he sought for the ancient unities: *i.e.*, the short story should relate a complete action and take place within one day in one place. To these unities he added that of mood or effect. He was not extreme in these views, however. He praised longer works and sometimes thought allegories and morals admirable if not crudely presented. Poe admired originality, often in work very different from his own, and was sometimes an unexpectedly generous critic of decidedly minor writers.

Poe's genius was early recognized abroad. No one did more to persuade the world and, in the long run, the United States, of Poe's greatness than the French poets Charles Baudelaire and Stéphane Mallarmé. Indeed his role in French literature was that of a poetic master-model and guide to criticism. French Symbolism relied on his "Philosophy of Composition," borrowed from his imagery, and used his examples to generate the modern theory of "pure poetry."

MAJOR WORKS

POEMS: *Tamerlane and Other Poems* (1827); *Al Aaraaf, Tamerlane, and Minor Poems* (1829); *Poems by Edgar Allan Poe* (1831); *The Raven and Other Poems* (1845); "Ulalume" (1847); "Annabel Lee" (1849); "The Bells" (1849).

SHORT STORIES: *Tales of the Grotesque and Arabesque* (1840); *The Prose Romances of Edgar A. Poe* (1843), includes "The Murders in the Rue Morgue" and "The Man That Was Used Up"; *Tales* (1845), includes "The Fall of the House of Usher," "The Mystery of Marie Roget," "The Gold Bug," "The Black Cat," and "The Purloined Letter"; "The Masque of the Red Death" (1842); "The Pit and the Pendulum" (1843); "The Tell-Tale Heart" (1843); "The Premature Burial" (1844); "The Cask of Amontillado" (1846).

BIBLIOGRAPHY

General works: ARTHUR H. QUINN, *Edgar Allan Poe: A Critical Biography* (1941, reprinted 1969), long the standard work; N. BRYLLION FAGIN, *The Histrionic Mr. Poe* (1949), emphasizes the theatrical quality of his life; PHILIP LINDSAY, *The Haunted Man: A Portrait of Edgar Allan Poe* (1953); GEOFFREY RANS, *Edgar Allan Poe* (1965), introductory; EDWARD WAGENKNECHT, *Edgar Allan Poe: The Man Behind the Legend* (1963), his character and personality; FRANCES WINWAR, *The Haunted Palace: A Life of Edgar Allan Poe* (1959); C. ALPHONSO SMITH, *Edgar Allan Poe: How to Know Him* (1921), introductory; JAMES A. HARRISON, *Life of Edgar Allan Poe* (1903, reprinted 1970); VINCENT BURANELLI, *Edgar Allan Poe* (1961), primarily a critical introduction; UNA CONSTANCE POPE-HENNESSY, *Edgar Allan Poe, 1809–1849: A Critical Biography* (1934, reprinted 1969); WILLIAM ROBERT BITTNER, *Poe: A Biography* (1962), includes a summary of the basic points of scholarly controversy over his life; HERVEY ALLEN, *Israfel: The Life and Times of Edgar Allan Poe*, 2 vol. (1926), popular, partly fictionalized; ROGER ASSELINEAU, *Edgar Allan Poe* (1970), a brief introduction; HALDEEN BRADDY, *Glorious Incense: The Fulfillment of Edgar Allan Poe*, 2nd ed. (1968), an overview of his achievement and influence.

Critical analyses: E.H. DAVIDSON, *Poe: A Critical Study* (1957); PERRY MILLER, *The Raven and the Whale* (1956); LOUIS BROUSSARD, *The Measure of Poe* (1969), emphasis on his work as allegory and symbolism.

Critical essays: ERIC W. CARLSON (ed.), *The Recognition of Edgar Allan Poe: Selected Criticism Since 1829* (1966); WILLIAM L. HOWARTH (ed.), *Twentieth Century Interpretations of Poe's Tales* (1971); ROBERT CHARLES REGAN (ed.), *Poe: A Collection of Critical Essays* (1967); FLOYD STOVALL, *Edgar Poe the Poet: Essays New and Old on the Man and His Work* (1969); KILLIS CAMPBELL, *The Mind of Poe, and Other Studies* (1933, reprinted 1962), each deals with some matter that has been in dispute among students of Poe.

Scholarly aids: CHARLES F. HEARTMAN and JAMES R. CANNY (comps.), *A Bibliography of First Printings of the Writings of Edgar Allan Poe . . .*, rev. ed. (1943); JOHN ALBERT ROBBINS, *The Merrill Checklist of Edgar Allan Poe* (1969); BURTON POLLIN, *Dictionary of Names and Titles in Poe's Collected Works* (1968); J.L. DAMERON, *Edgar Allan Poe: A Checklist of Criticism, 1942–1960* (1966); JOHN W. ROBERTSON, *Bibliography of the Writings of Edgar Allan Poe*, 2 vol. (1934, reprinted 1969); BRADFORD BOOTH and CLAUDE E. JONES, *A Concordance of the Poetical Works of Edgar Allan Poe* (1941); R.L. GALE, *Plots and Characters in the Fiction and Poetry of Edgar Allan Poe* (1970).

Specialized topics: E.W. PARKS, *Edgar Allan Poe As Literary Critic* (1964); CHARLES BAUDELAIRE, *Baudelaire on Poe: Critical Papers*, trans. by LOIS and FRANCIS E. HYSLOP (1952), includes a list of Poe's works translated by Baudelaire; JOHN WALSH, *Poe the Detective: The Curious Circumstances Behind "The Mystery of Marie Roget"* (1967), reinvestigation of a crime used in a Poe story; AGNES M. BONDURANT, *Poe's Richmond* (1942), on the city in which he spent his youth; DAVID K. JACKSON, *Poe and the Southern Literary Messenger* (1934, reprinted 1970), the periodical he edited; HERVEY ALLEN and T.O. MABBOTT, *Poe's Brother* (1926); PALMER COBB, *The Influence of E.T.A. Hoffmann on the Tales of Edgar Allan Poe* (1908); SIDNEY P. MOSS, *Poe's Literary Battles: The Critic in the Context of His Literary Milieu* (1963); CELESTIN P. CAMBIAIRE, *The Influence of Edgar Allan Poe in France* (1927, reprinted 1971); SAMUEL MOSKOWITZ (ed.), *The Man Who Called Himself Poe* (1969), an anthology of fiction and poetry in which Poe appears as an integral character; MICHAEL ALLEN, *Poe and the British Magazine Tradition* (1969); PATRICK FRANCIS QUINN, *The French Face of Edgar Poe* (1957); JUNICHI NAKAMURA, *Edgar Allan Poe's Relations with New England* (1957).

Psychoanalytic: JOSEPH WOOD KRUTCH, *Edgar Allan Poe: A Study in Genius* (1926); JOHN W. ROBERTSON, *Edgar Allan Poe: A Psychopathic Study* (1922); DAVID M. REIN, *Edgar Allan Poe: The Inner Pattern* (1960); ARTHUR LERNER, *Psychoanalytically Oriented Criticism of Three American Poets: Poe, Whitman, and Aiken* (1970); MARIE BONAPARTE, *Edgar Poe, étude psychoanalytique*, 2 vol. (1933; Eng. trans., *The Life and Works of Edgar Allan Poe*, with an introduction by SIGMUND FREUD, 1949), often farfetched.

(C.Ce./T.O.M./J.Ba.)

Poetry

Poetry is a vast subject, as old as history and older, present wherever religion is present, possibly—under some definitions—the primal and primary form of languages themselves. The present article means only to describe in as general a way as possible certain properties of poetry and of poetic thought regarded as in some sense independent modes of the mind. Naturally, not every tradition nor every local or individual variation can be—or need be—included, but the article illustrates by examples of poetry ranging between nursery rhyme and epic. (For particular information about various types of poetry, see PROSODY and FOLK LITERATURE and articles on poetic genres, such as BALLAD and EPIC; for specific national poetic forms, see AFRICAN PEOPLES, ARTS OF; AMERICAN INDIAN PEOPLES, ARTS OF; CENTRAL ASIAN PEOPLES, ARTS OF; ISLAMIC PEOPLES, ARTS OF; JEWISH PEOPLES, ARTS OF; LITERATURE, EAST ASIAN; LITERATURE, WESTERN; OCEANIC PEOPLES, ARTS OF; SOUTH ASIAN PEOPLES, ARTS OF; and SOUTHEAST ASIAN PEOPLES, ARTS OF.) This article will consider the difficulty or impossibility of defining poetry; man's nevertheless familiar acquaintance with it; the differences between poetry and prose; the idea of form in poetry; poetry as a mode of thought; and what little may be said in prose of the spirit of poetry.

Attempts to define poetry. Poetry is the other way of using language. Perhaps in some hypothetical beginning of things it was the only way of using language or simply was language *tout court*, prose being the derivative and younger rival. Both poetry and language are fashionably thought to have belonged to ritual in early agricultural societies; and poetry in particular, it has been claimed, arose at first in the form of magical spells recited to ensure a good harvest. Whatever the truth of this hy-

pothesis, it blurs a useful distinction: by the time there begins to be a separate class of objects called poems, recognizable as such, these objects are no longer much regarded for their possible yam-growing properties, and such magic as they may be thought capable of has retired to do its business upon the human spirit and not directly upon the natural world outside.

Formally, poetry is recognizable by its greater dependence on at least one more parameter, the *line*, than appears in prose composition. This changes its appearance on the page; and it seems clear that people take their cue from this changed appearance, reading poetry aloud in a very different voice from their habitual voice, possibly because, as Ben Jonson said, poetry "speaketh somewhat above a mortal mouth." If, as a test of this description, people are shown poems printed as prose, it most often turns out that they will read the result as prose simply because it looks that way; which is to say that they are no longer guided in their reading by the balance and shift of the line in relation to the breath as well as the syntax.

That is a minimal definition but perhaps not altogether uninformative. It may be all that ought to be attempted in the way of a definition: Poetry is the way it is because it looks that way, and it looks that way because it sounds that way and vice versa.

Poetry and prose. People's reason for wanting a definition is to take care of the borderline case, and this is what a definition, as if by definition, will not do. That is, if a man asks for a definition of poetry, it will most certainly not be the case that he has never seen one of the objects called poems that are said to embody poetry; on the contrary, he is already tolerably certain what poetry in the main is, and his reason for wanting a definition is either that his certainty has been challenged by someone else or that he wants to take care of a possible or seeming exception to it: hence the perennial squabble about distinguishing poetry from prose, which is rather like distinguishing rain from snow—everyone is reasonably capable of doing so, and yet there are some weathers that are either-neither.

Distinguishing prose from poetry

Sensible things have been said on the question. The poet T.S. Eliot suggested that part of the difficulty lies in the fact that there is the technical term "verse" to go with the term "poetry," while there is no equivalent technical term to distinguish the mechanical part of prose and make the relation symmetrical. The French poet Paul Valéry said that prose was walking, poetry dancing. Indeed, the original two terms, *prosus* and *versus*, meant, respectively, "going straight forth" and "returning"; and that distinction does point up the tendency of poetry to incremental repetition, variation, and the treatment of many matters and different themes in a single recurrent form such as couplet or stanza.

Robert Frost said shrewdly that poetry was what got left behind in translation, which suggests a criterion of almost scientific refinement: when in doubt, translate; whatever comes through is prose, the remainder is poetry. And yet to even so acute a definition the obvious exception is a startling and a formidable one: some of the greatest poetry in the world is in the Authorized Version of the Bible, which is not only a translation but also, as to its appearance in print, identifiable neither with verse nor with prose in English but rather with a cadence owing something to both.

There may be a better way of putting the question by the simple test alluded to above. When people are presented with a series of passages drawn indifferently from poems and stories but all printed as prose, they will show a dominant inclination to identify everything they possibly can as prose. This will be true, surprisingly enough, even if the poem rhymes and will often be true even if the poem in its original typographical arrangement would have been familiar to them. The reason seems to be absurdly plain: the reader recognizes poems by their appearance on the page, and he responds to the convention whereby he recognizes them by reading them aloud in a quite different tone of voice from that which he applies to prose (which, indeed, he scarcely reads aloud at all). It should be added that he makes this distinction also without reading aloud; even in silence he confers upon a piece of poetry an attention that differs from what he gives to prose in two ways especially: in tone and in pace.

In place of further worrying over definitions, it may be both a relief and an illumination to exhibit certain plain and mighty differences between prose and poetry by a comparison. In the following passages a prose writer and a poet are talking about the same subject, growing older.

> Between the ages of 30 and 90, the weight of our muscles falls by 30 percent and the power we can exert likewise . . . . The number of nerve fibres in a nerve trunk falls by a quarter. The weight of our brains falls from an average of 3.03 lb. to 2.27 lb. as cells die and are not replaced. . . . (Gordon Rattray Taylor, *The Biological Time Bomb*, 1968.)

> Let me disclose the gifts reserved for age
> To set a crown upon your lifetime's effort.
> First, the cold friction of expiring sense
> Without enchantment, offering no promise
> But bitter tastelessness of shadow fruit
> As body and soul begin to fall asunder.
> Second, the conscious impotence of rage
> At human folly, and the laceration
> Of laughter at what ceases to amuse.
> And last, the rending pain of re-enactment
> Of all that you have done, and been. . . .
> (T.S. Eliot, *Four Quartets.*)

Before objecting that a simple comparison cannot possibly cover all the possible ranges of poetry and prose compared, the reader should consider for a moment what differences are exhibited. The passages are oddly parallel, hence comparable, even in a formal sense; for both consist of the several items of a catalog under the general title of growing old. The significant differences are of tone, pace, and object of attention. If the prose passage interests itself in the neutral, material, measurable properties of the process, while the poetry interests itself in what the process will signify to someone going through it, that is not accidental but of the essence; if one reads the prose passage with an interest in being informed, noting the parallel constructions without being affected by them either in tone or in pace, while reading the poetry with a sense of considerable gravity and solemnity, that too is of the essence. One might say as tersely as possible that the difference between prose and poetry is most strikingly shown in the two uses of the verb "to fall":

Differences of tone, pace, and object of attention

> The number of nerve fibres in a nerve trunk falls by a quarter
> As body and soul begin to fall asunder

It should be specified here that the important differences exhibited by the comparison belong to the present age. In each period, speaking for poetry in English at any rate, the dividing line will be seen to come at a different place. In Elizabethan times the diction of prose was much closer to that of poetry than it later became, and in the 18th century authors saw nothing strange about writing in couplets about subjects that later would automatically and compulsorily belong to prose—for example, horticulture, botany, even dentistry. This brief article is not the place for entering into a discussion of so rich a chapter in the history of ideas; but it should be remarked that the changes involved in the relation between poetry and prose are powerfully influenced by the immense growth of science, commerce, and number in man's ways of describing, even of viewing, the world.

Returning to the comparison, it is observable that though the diction of the poem is well within what could be commanded by a moderately well-educated speaker, it is at the same time well outside the range of terms in fact employed by such a speaker in his daily occasions; it is a diction very conscious, as it were, of its power of choosing terms with an effect of peculiar precision and of combining the terms into phrases with the same effect of peculiar precision and also of combining sounds with the same effect of peculiar precision. Doubtless the precision of the prose passage is greater in the more obvious property of dealing in the measurable; but the poet attempts a precision with respect to what is not in the same sense measurable nor even in the same sense accessible to observation; the distinction is perhaps just that made by the French scientist and philosopher Blaise Pascal in discrim-

Precision of poetic diction

inating the spirits of geometry and finesse; and if one speaks of "effects of precision" rather than of precision itself, that serves to distinguish one's sense that the art work is always somewhat removed from what people are pleased to call the real world, operating instead, in Immanuel Kant's shrewd formula, by exhibiting "purposefulness without purpose." To much the same point is what Samuel Taylor Coleridge remembers having learned from his schoolmaster:

> I learnt from him, that Poetry, even that of the loftiest and, seemingly, that of the wildest odes, had a logic of its own, as severe as that of science; and more difficult, because more subtle, more complex, and dependent on more, and more fugitive causes. In the truly great poets, he would say, there is a reason assignable, not only for every word, but for the position of every word. (*Biographia Literaria*, ch. I.)

Perhaps this is a somewhat exaggerated, as it is almost always an unprovable, claim, illustrating also a propensity for competing with the prestige of science on something like its own terms—but the last remark in particular illuminates the same author's terser formulation: "prose = words in the best order, poetry = the best words in the best order." This attempt at definition, impeccable because uninformative, was derived from Jonathan Swift, who had said, also impeccably and uninformatively, that style in writing was "the best words in the best order." Which may be much to the same effect as Louis Armstrong's saying, on being asked to define jazz, "Baby, if you got to ask the question, you're never going to know the answer." Or the painter Marcel Duchamp's elegant remark on what psychologists call "the problem of perception": "If no solution, then maybe no problem?" This species of gnomic, riddling remark may be determinate for the artistic attitude toward definition of every sort; and its skepticism is not confined to definitions of poetry but extends to definitions of anything whatever, directing one not to dictionaries but to experience and, above all, to use: "Anyone with a watch can tell you what time it is," said Valéry, "but who can tell you what is time?"

Experiencing poetry

Happily, if poetry is almost impossible to define, it is extremely easy to recognize in experience; even untutored children are rarely in doubt about it when it appears:

> Little Jack Jingle,
> He used to live single,
> But when he got tired of this kind of life,
> He left off being single, and liv'd with his wife.

It might be objected that this little verse is not of sufficient import and weight to serve as an exemplar for poetry. It ought to be remembered, though, that it has given people pleasure so that they continued to say it until and after it was written down, nearly two centuries ago. The verse has survived, and its survival has something to do with pleasure, with delight; and while it still lives, how many more imposing works of language—epic poems, books of science, philosophy, theology—have gone down, deservedly or not, into dust and silence. It has, obviously, a form, an arrangement of sounds in relation to thoughts that somehow makes its agreeable nonsense closed, complete, and decisive. But this somewhat muddled matter of form deserves a heading and an instance all to itself.

Form in poetry. People nowadays who speak of form in poetry almost always mean such externals as regular measure and rhyme, and most often they mean to get rid of these in favour of the freedom they suppose must follow upon the absence of form in this limited sense. But in fact a poem having only one form would be of doubtful interest even if it could exist. In this connection, the poet J.V. Cunningham speaks of "a convergence of forms, and forms of disparate orders," adding: "It is the coincidence of forms that locks in the poem." For a poem is composed of internal and intellectual forms as well as forms externally imposed and pre-existing any particular instance, and these may be sufficient without regular measure and rhyme; if the intellectual forms are absent, as in greeting-card verse and advertising jingles, no amount of thumping and banging will supply the want.

Form, in effect, is like the doughnut that may be made either by putting some nothing into a circle of something or by closing some something around some nothing; it is either the outside of an inside, as when people speak of "good form" or "bourgeois formalism," or the inside of an outside, as in the scholastic saying that "the soul is the form of the body." Taking this principle, together with what Cunningham says of the matter, one may now look at a very short and very powerful poem with a view to distinguishing the forms, or schemes, of which it is made. It was written by Rudyard Kipling—a great poet at present somewhat sunken in reputation, probably on account of misinterpretations having to do more with his imputed politics than with his poetry—and its subject, one of a series of epitaphs for the dead of World War I, is a soldier shot by his comrades for cowardice in battle.

> I could not look on Death, which being known,
> Men led me to him, blindfold and alone.

The aim of the following observations and reflections is to distinguish as clearly as possible—distinguish without dividing—the feelings evoked by the subject, so grim, horrifying, tending to helpless sorrow and despair, from the feelings, which might better be thought of as meanings, evoked by careful contemplation of the poem in its manifold and somewhat subtle ways of handling the subject, leading the reader on to a view of the strange delight intrinsic to art, whose mirroring and shielding power allows him to contemplate the world's horrible realities without being turned to stone.

Kinds of poetic form

There is, first, the obvious external form of a rhymed, closed couplet in iambic pentameter (that is, five poetic "feet," each consisting of one stressed and one unstressed syllable, per line). There is, second, the obvious external form of a single sentence balanced in four grammatical units with and in counterpoint with the metrical form. There is, third, the conventional form belonging to the epitaph and reflecting back to antiquity; it is terse enough to be cut in stone and tight-lipped also, perhaps for other reasons, such as the speaker's shame. There is, fourth, the fictional form belonging to the epitaph, according to which the dead man is supposed to be saying the words himself. There is, fifth, especially poignant in this instance, the real form behind or within the fictional one, for the reader is aware that in reality it is not the dead man speaking, nor are his feelings the only ones the reader is receiving, but that the comrades who were forced to execute him may themselves have made up these two lines with their incalculably complex and exquisite balance of scorn, awe, guilt, and consideration even to tenderness for the dead soldier. There is, sixth, the metaphorical form, with its many resonances ranging from the tragic through the pathetic to irony and apology: dying in battle is spoken of in language relating it to a social occasion in drawing room or court; the coward's fear is implicitly represented as merely the timorousness and embarassment one might feel about being introduced to a somewhat superior and majestic person, so that the soldiers responsible for killing him are seen as sympathetically helping him through a difficult moment in the realm of manners. In addition, there is, seventh, a linguistic or syntactical form, with at least a couple of tricks to it: the second clause, with its reminiscence of Latin construction, participates in the meaning by conferring a Roman stoicism and archaic gravity on the saying; remembering that the soldiers in the poem had been British schoolboys not long before, the reader might hear the remote resonance of a whole lost world built upon Greek and Roman models; and the last epithets, "blindfold and alone," while in the literal acceptation they clearly refer to the coward, show a distinct tendency to waver over and apply mysteriously to Death as well, sitting there waiting "blindfold and alone." One might add another form, the eighth, composed of the balance of sounds, from the obvious likeness in the rhyme down to subtleties and refinements beneath the ability of coarse analysis to discriminate. And even there one would not be quite at an end; an overall principle remains, the compression of what might have been epic or five-act tragedy into two lines, or the poet's precise election of a single instant to carry what the novelist, if he did his business properly, would have been hundreds of pages arriving at.

The coincidence of poetic forms

It is not at all to be inferred that the poet composed his poem in the manner of the above laborious analysis of its strands; the whole insistence, rather, is that he did not catalog eight or 10 forms and assemble them into a poem; more likely it "just came to him." But the example may serve to indicate how many modes of the mind go together in this articulation of an implied drama and the tension among many possible sentiments that might arise in response to it.

In this way, by the coincidence of forms that locks in the poem, one may see how to answer a question that often arises about poems: though their thoughts are commonplace, they themselves mysteriously are not. One may answer on the basis of the example and the inferences produced from it that a poem is not so much a thought as it is a mind: talk with it, and it will talk back, telling you many things that you might have thought for yourself but somehow didn't until it brought them together. Doubtless a poem is a much simplified model for the mind. But it might still be one of the best man has available. On this great theme, however, it will be best to proceed not by definition but by parable and interpretation.

Poetry as a mode of thought: the Protean encounter. In the fourth book of the *Odyssey* Homer tells the following strange tale. After the war at Troy, Menelaus wanted very much to get home but was held up in Egypt for want of a wind because, as he later told Telemachus, he had not sacrificed enough to the gods. "Ever jealous the Gods are," he said, "that we men mind their dues." But because the gods work both ways, it was on the advice of a goddess, Eidothea, that Menelaus went to consult Proteus, the old one of the sea, as one might consult a travel agency.

Proteus was not at all easy to consult. He was herding seals, and the seals stank even through the ambrosia Eidothea had thoughtfully provided. And when Menelaus crept up close, disguised as a seal, and grabbed him, Proteus turned into a lion, a dragon, a leopard, a boar, a film of water, and a high-branched tree. But Menelaus managed to hang on through all this until at last Proteus gave up and was himself again; whereupon Menelaus asked him the one great question: How do I get home? And Proteus told him: You had better go back to Egypt and sacrifice to the gods some more.

This story may be taken as a parable about poetry. A man has an urgent question about his way in the world. He already knows the answer, but it fails to satisfy him. So at great inconvenience, hardship, and even peril, he consults a powerful and refractory spirit who tries to evade his question by turning into anything in the world. Then, when the spirit sees he cannot get free of the man, and only then, he answers the man's question, not simply with a commonplace but with the same commonplace the man had been dissatisfied with before. Satisfied or not, however, the man now obeys the advice given him.

A foolish story? All the same, it is to be observed that Menelaus did get home. And it was a heroic thing to have hung onto Proteus through those terrifying changes and compelled him to be himself and answer up. Nor does it matter in the least to the story that Menelaus personally may have been a disagreeable old fool as well as a cuckold.

A poet also has one great and simple question, simple though it may take many forms indeed. Geoffrey Chaucer put it as well as anyone could, and in three lines at that:

> What is this world? what asketh men to have?
> Now with his love, now in his colde grave,
> Allone, with-outen any companye.
> ("The Knights Tale," 1919–1921)

And a poet gets the simple answer he might expect, the one the world grudgingly gives to anyone who asks such a question: The world is this way, not that way, and you ask for more than you will be given, which the poet, being scarcely more fool than his fellowmen, knew already. But on the path from question to answer, hanging onto the slippery disguiser and shape-shifter Proteus, he will see many marvels; he will follow the metamorphoses of things in the metamorphoses of their phrases, and he will be so elated and ecstatic in this realm of wonders that the voice in which he speaks these things, down even to the stupid, obvious, and commonplace answer, will be to his hearers a solace and a happiness in the midst of sorrows:

> When I do count the clock that tells the time,
> And see the brave day sunk in hideous night;
> When I behold the violet past prime,
> And sable curls, all silver'd o'er with white;
> When lofty trees I see barren of leaves,
> Which erst from heat did canopy the herd,
> And summer's green all girded up in sheaves,
> Borne on the bier with white and bristly beard,
> Then of thy beauty do I question make,
> That thou among the wastes of time must go,
> Since sweets and beauties must themselves forsake
> And die as fast as they see others grow;
> And nothing 'gainst Time's scythe can make defence
> Save breed, to brave him when he takes thee hence.
> (Shakespeare, Sonnet 12.)

The poetic question and answer

Like Menelaus, the poet asks a simple question, to which, moreover, he already knows the unsatisfying answer. Question and answer, one might say, have to be present, although of themselves they seem to do nothing much; but they assert the limits of a journey to be taken. They are the necessary but not sufficient conditions of what really seems to matter here, the Protean encounter itself, the grasping and hanging on to the powerful and refractory spirit in its slippery transformations of a single force flowing through clock, day, violet, graying hair, trees dropping their leaves, the harvest in which, by a peculiarly ceremonial transmutation, the grain man lives by is seen without contradiction as the corpse he comes to. As for the answer to the question, it is not surprising nor meant to be surprising; it is only just.

On this point—that the answer comes as no surprise—poets show an agreement that quite transcends the differences of periods and schools. Alexander Pope's formula, "What oft was thought, but ne'er so well expresst," sometimes considered as the epitome of a shallow and parochial decorum, is not in essence other than this of John Keats:

> I think Poetry should surprise by a fine excess, and not by Singularity—it should strike the Reader as a wording of his own highest thoughts, and appear almost a Remembrance. (Letter to John Taylor, 1818.)

In the present century, Robert Frost is strikingly in agreement:

> A word about recognition: In literature it is our business to give people the thing that will make them say, "Oh yes I know what you mean." It is never to tell them something they dont know, but something they know and hadnt thought of saying. It must be something they recognize. (Letter to John Bartlett, in *Modern Poetics*, ed. James Scully, 1965.)

And the poet and critic John Crowe Ransom gives the thought a cryptically and characteristically elegant variation: "Poetry is the kind of knowledge by which we must know that we have arranged that we shall not know otherwise." Perhaps this point about recognition might be carried further, to the extreme at which it would be seen to pose the problem of how poetry, which at its highest has always carried, at least implicitly, a kind of Platonism and claimed to give, if not knowledge itself, what was more important, a "form" to knowledge, can survive the triumph of scientific materialism and a positivism minded to skepticism about everything in the world except its own self (where it turns credulous, extremely). The poets' adjustment, over two or three centuries, to a Newtonian cosmos, Kantian criticism, the spectral universe portrayed by physics has conspicuously not been a happy one and has led alternately or simultaneously to the extremes of rejection of reason and speaking in tongues on the one hand and the hysterical claim that poetry will save the world on the other. But of this let the Protean parable speak as it will.

There is another part to the story of Menelaus and Proteus, for Menelaus asked another question: What happened to my friends who were with me at Troy? Proteus replies, "Son of Atreus, why enquire too closely of me on

this? To know or learn what I know about it is not your need: I warn you that when you hear all the truth your tears will not be far behind. . . ." But he tells him all the same: "Of those others many went under; many came through. . . ." And Menelaus does indeed respond with tears of despair, until Proteus advises him to stop crying and get started on the journey home. So it sometimes happens in poetry, too: the sorrowful contemplation of what is, consoles, in the end, and heals, but only after the contemplative process has been gone through and articulated in the detail of its change:

> When to the sessions of sweet silent thought
> I summon up remembrance of things past,
> I sigh the lack of many a thing I sought,
> And with old woes new wail my dear time's waste;
> Then can I drown an eye, unused to flow,
> For precious friends hid in death's dateless night,
> And weep afresh love's long since cancell'd woe,
> And moan the expense of many a vanish'd sight.
> Then can I grieve at grievances foregone,
> And heavily from woe to woe tell o'er
> The sad account of fore-bemoaned moan,
> Which I new pay as if not paid before.
> But if the while I think on thee, dear friend,
> All losses are restor'd and sorrows end.
>
> (Shakespeare, Sonnet 30)

This poem, acknowledged to be a masterpiece by so many generations of readers, may stand as an epitome and emblem for the art altogether, about which it raises a question that must be put, although it cannot be satisfactorily and unequivocally answered: the question of whether poetry is a sacrament or a confidence game or both or neither. To reply firmly that poetry is not religion and must not promise what religion does is to preserve a useful distinction; nevertheless, the religions of the world, if they have nothing else in common, seem to be based on collections of sacred poems. Nor, at the other extreme, can any guarantee that poetry is not a confidence game be found in the often-heard appeal to the poet's "sincerity." One will never know whether Shakespeare wept all over the page while writing the 30th sonnet, though one inclines to doubt it, nor would it be to his credit if he did, nor to the reader's that he should know it or care to know it.

The nature of poetry

For one thing, the sonnet is obviously artful—that is, full of artifice—and even the artifice degenerates here and there into being artsy. "Then can I drown an eye, unused to flow." Surely that is poesy itself, at or near its worst, where the literal and the conventional, whatever their relations may have been for Shakespeare and the first reader of these sugar'd sonnets among his friends, now live very uncomfortably together (Ben Jonson's "Drink to me only with thine eyes" is a like example of this bathetic crossing of levels), though perhaps it has merely become unattractive as a result of changing fashions in diction.

Moreover, while the whole poem is uniquely Shakespearean, the bits and pieces are many of them common property of the age, what one writer called "joint stock company poetry." And the tricks are terribly visible, too; art is not being used to conceal art in such goings-on as "grieve at grievances" and "fore-bemoaned moan." "He who thus grieves will excite no sympathy," as Samuel Johnson sternly wrote of John Milton's style in the elegy "Lycidas," "he who thus praises will confer no honour."

Nor is that the worst of it. This man who so powerfully works on the reader's sympathies by lamenting what is past contrives to do so by thinking obsessively about litigation and, of all things, money; his hand is ever at his wallet, bidding adieu. He cannot merely "think" sweet silent thoughts about the past; no, he has to turn them into a court in "session," whereto he "summons" the probably culprit "remembrance"; when he "grieves," it is at a "grievance"—in the hands of the law again; finally, as with the sinners in Dante's *Divine Comedy*, his avarice and prodigality occupy two halves of the one circle: he bemoans his expenses while paying double the asking price.

And still, for all that, the poem remains beautiful; it continues to move both the young who come to it still innocent of their dear time's waste and the old who have sorrows to match its sorrows. As between confidence game and sacrament there may be no need to decide, as well as no possibility of deciding: elements of play and artifice, elements of true feeling, elements of convention both in the writing and in one's response to it, all combine to veil the answer. But the poem remains.

Effect of the poem as a whole

If it could be plainly demonstrated by the partisans either of unaided reason or revealed religion that poetry was metaphorical, mythological, and a delusion, while science, say, or religion or politics were real and true, then one might throw poetry away and live honestly though poorly on what was left. But, for better or worse, that is not the condition of man's life in the world; and perhaps men care for poetry so much—if they care at all—because, at last, it is the only one of man's many mythologies to be aware, and to make him aware, that it, and the others, are indeed mythological. The literary critic I.A. Richards, in a deep and searching consideration of this matter, concludes: "It is the privilege of poetry to preserve us from mistaking our notions either for things or for ourselves. Poetry is the completest mode of utterance."

The last thing Proteus says to Menelaus is strange indeed:

> You are not to die in Argos of the fair horse-pastures, not there to encounter death: rather will the Deathless Ones carry you to the Elysian plain, the place beyond the world. . . . There you will have Helen to yourself and will be deemed of the household of Zeus.

So the greatest of our poets have said, or not so much said, perhaps, as indicated by their fables, though nowadays people mostly sing a different tune. To be as the gods, to be rejoined with the beloved, the world forgotten. . . . Sacrament or con game? Homer, of course, is only telling an old story and promises mankind nothing; that is left to the priests to do; and in that respect poetry, as one critic puts it, must always be "a ship that is wrecked on entering the harbor." And yet the greatest poetry sings always, at the end, of transcendence; while seeing clearly and saying plainly the wickedness and terror and beauty of the world, it is at the same time humming to itself, so that one overhears rather than hears: All will be well.

BIBLIOGRAPHY. The most convenient way to get to know poetry is to read poetry. It would be invidious for the writer of a general article on the subject to prejudice the reader by making a selection of poems or poets; in experience, anyhow, one's acquaintance with poetry comes about chiefly by love and accident, supported, when not undermined, by schools, colleges, and libraries. Beyond that, the bibliographical temptation is to put before the reader numerous learned works that are not poetry but about poetry; whatever their usefulness at various stages of study, and it may be great, they must not substitute for poetry itself. Therefore no such list is attempted.

The beginning reader, however, may well be able to use some help in interpreting, such as a critical or explanatory anthology. CLEANTH BROOKS and ROBERT PENN WARREN, *Understanding Poetry*, 3rd ed. (1960), is still probably the best of its kind, as numerous imitations amply attest. Useful instruction in method may be found in I.A. RICHARDS, *Practical Criticism: A Study of Literary Judgment* (1929, reprinted 1964).

(H.Ne.)

Pogonophora

Members of the phylum Pogonophora are invertebrate animals that live a sedentary life in long protective tubes on sea floors throughout the world. Pogonophora, which have neither mouth, anus, nor gut, bear one to many pinnate (featherlike) tentacles at their anterior end; hence their common name beardworms. An intestine, which forms in embryos, disappears as development progresses. Males of the phylum are generally similar in appearance to females.

Pogonophora were first classified as a distinct phylum in the middle of the 20th century. The first species, *Siboglinum weberi*, described in 1914, came from the seas of the Malayan Archipelago; the second species, *Lamellisabella zachsi*, which came from the Okhotsk Sea, was described

in 1933. In 1937 a new class called Pogonophora was established for *Lamellisabella*. In 1955 a close affinity between *Siboglinum* and *Lamellisabella* was proved, and the members were placed in the newly established phylum Pogonophora.

Pogonophora usually inhabit marine waters to depths that exceed 1,000 metres; some species have been found at depths ranging from 7,000 to 10,000 metres. A native population of Pogonophora inhabits the trenches of the Pacific and the Indian oceans. Many genera have a discontinuous distribution.

Feeding

Natural history. The tentacles, probably used during feeding, vary in number from 1 to 2,000; the number is correlated with body size. The primitive Pogonophora have a few tentacles. Among members of the families Lamellisabellidae, Spirobrachiidae, and Lamellibrachiidae, the neighbouring tentacles are fused through most of their length. The tentacles, long processes containing blood vessels, are continuous with the body cavity, or coelom. Rows of very thin single-celled units called pinnules are found on the tentacles. The pinnules, which extend into the intertentacular cavity formed by the free or fused tentacles, intermesh to form a filter. Beside each pinnule base is a ciliary tract. In each intertentacular region, ciliary tracts produce a current of water that carries in micro-organisms and other nutrients; these are filtered through the pinnule filter. Digestive enzymes probably are secreted by gland cells located at the pinnule bases. Water leaves by an opening at the base of certain tentacles. The digested food may be absorbed by the pinnules. The pinching off of small food particles in the outer layers of the pinnules during a process known as pinocytosis also has been observed. Direct penetration of amino acids from sea water into the tentacles also has been reported to take place.

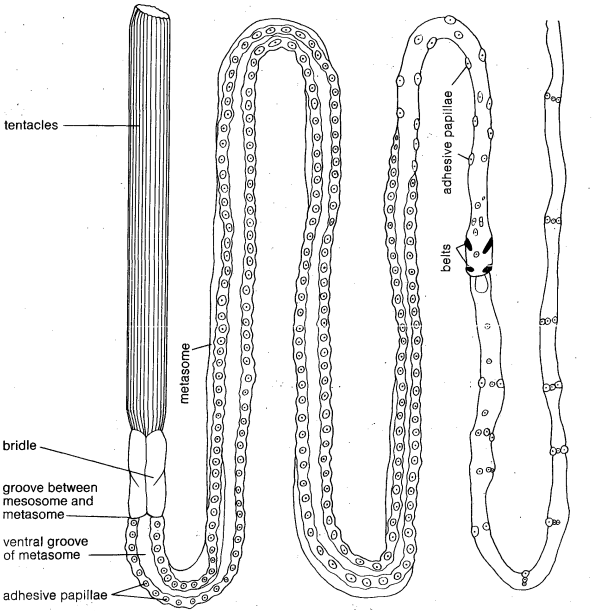
Reproduction in Pogonophora is sexual. Eggs, which have a high yolk content, are laid by the female in the anterior part of the tube, in which fertilization and early development to formation of a swimming larva also occur. Cleavage leads to the formation of a bilaterally symmetrical embryo. A cell layer called the endoderm first is represented by a few yolk-rich cells, which form a primary intestine; the intestine disappears as the yolk is utilized. Both mouth and anus fail to develop. A secondary body cavity is formed by separation of coelomic pockets from the primary intestine. The embryo eventually develops into a larva having two ciliary belts and three body segments.

Body plan

Form and function. The wormlike body varies in length from several centimetres to 0.5 metre, the body diameter, from 0.06 millimetre to 4 millimetres. *Lamellibrachia barhami* is one of the largest species. The body consists of three segments; two small anterior regions are called protosome and mesosome; the long trunk section is called the metasome. Each segment has its own coelom. The small protosome bears tentacles. The mesosome contains a structure known as a bridle, also called a frenulum, a pair of oblique cuticular ridges that extend backward to meet in the midventral line (see drawing). The bridle supports the protruding worm on the edge of its tube. The metasome is divided into two sections by a pair of parallel ridges called belts; these have rows of small platelets containing minute teeth. One part of the metasome, in front of the belts, has a groove bordered by low folds, each bearing a row of papillae. In *Siboglinum* the papillae contain glands or in other pogonophores may be topped with chitin-like platelets; in this case, papillae probably help the animal to cling to the inner surface of the tube. Behind the groove, the papillae decrease in number and are arranged irregularly.

The protective tube is submerged in mud at the lower end; the very thin tubes of *Sclerolinum*, however, pierce pieces of sunken wood. The tube, secreted by special glands, contains a hard substance called chitin. Although the animals never leave their tubes, they are able to move inside them.

The body wall consists of an outer layer of epidermal cells covered by a cuticle. Beneath the epidermal cells is a layer of longitudinal muscle cells. The nervous system, found within the epidermis, forms an almost continuous network. There are no sense organs. The circulatory system consists of two long vessels; in one, blood flows forward; in the other, backward. At the base of the tentacles, the vessel through which blood flows forward is differentiated into a muscular heart, which serves to push blood into the tentacle vessels. The blood, red in colour, contains the pigment hemoglobin. The excretory system consists of a pair of nephridia, or coelomoducts, which connect the body cavity of the protosome with the exterior. The reproductive system consists of a pair of sexual glands that lie in the trunk. In females, ovaries are found in the anterior part of the metasome, while the sexual openings, or oviducts, are found in the middle of the body. In males, the testes are located in the posterior part of the body, and the sexual openings, called sperm ducts, in the anterior part. Numerous capsules containing sperm (spermatophores) form in the sperm ducts of the males.



Female *Spirobrachia beklemischevi*, ventral view with posterior one-sixth of trunk not shown.
From A.V. Ivanov, *Zoologische Jahrbucher Abt. 1 Systematik*, 85 (1957); Gustav Fischer Verlag

Evolution and classification. The Pogonophora may be separated into two classes, Afrenulata and Frenulata. The Afrenulata contains one species—*Lamellibrachia barhami*, which has been found in the Pacific Ocean near the coast of California. The class Frenulata contains 16 genera in six families. A total of 110 species of Pogonophora thus far have been identified; it is probable that many more species remain to be identified.

According to general body plan and embryological development, Pogonophora should be included with the large group of animals that have a secondary body cavity; this group also includes Chaetognatha, Tentaculata, Hemichordata, and Echinodermata. A rich fossil Pogonophora fauna has been found in Cambrian (about 500,000,000 to 570,000,000 years ago) and Ordovician (about 430,000,000 to 500,000,000 years ago) deposits near Leningrad, in Poland, and in Siberia.

BIBLIOGRAPHY. L.H. HYMAN, "The Enterocoelous Coelomates—Phylum Pogonophora," in *The Invertebrates*, vol. 5, pp. 208-227 (1959); A.V. IVANOV, "Embranchement des Pogonophores," in P.P. GRASSE (ed.), *Traité de zoologie*, vol. 5, pp. 1521-1622 (1960), in French, and *Pogonophora* (Eng. trans. 1963); K.E. JOHANSSON, "Pogonophora," in J.G. HELMCKE *et al.* (eds.), *Handbuch der Zoologie*, 3, 2, 18, pp. 1-50 (1968), in German.

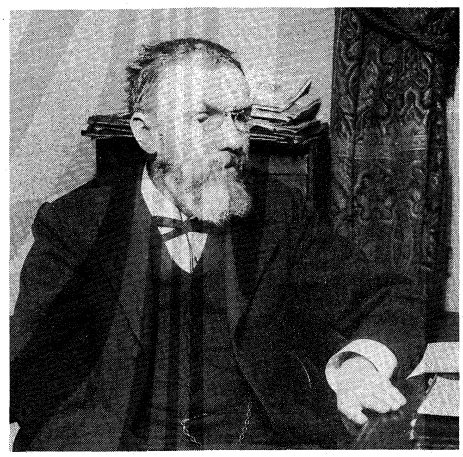
(A.V.I.)

Poincaré, Henri

The French mathematician, theoretical astronomer, and philosopher of science Jules-Henri Poincaré was generally acknowledged as the greatest living mathemati-

cian at the turn of the 20th century. He substantially influenced modern ideas concerning cosmogony, relativity, and topology. Read by thousands in all walks of life, he was a master of French prose in his writings on philosophical topics.

H. Roger-Viollet



Poincaré, 1909.

Education

Poincaré was born at Nancy on April 29, 1854, into a family distinguished by its contributions to government and administration. His first cousin was Raymond Poincaré, president of the French Republic during World War I. Poincaré was ambidextrous and was nearsighted; during his childhood he had poor muscular coordination and was seriously ill for a time with diphtheria. He received special instruction from his gifted mother, learned German during the Franco-Prussian War—at which time he also witnessed the suffering of the soldiers and populace with his father, who was a physician—and excelled in written composition in elementary school. Becoming deeply interested in mathematics during adolescence, he attended in 1872–75 the École Polytéchnique in Paris, where he easily won top honours in mathematics—but was undistinguished in physical exercise and in art. He had an unusually retentive memory for everything he read; moreover, he could visualize what he heard, a useful faculty because he could not clearly see at a distance the mathematical symbols that were on the blackboard. Throughout his life he was able to perform complex mathematical calculations in his head and could quickly write a paper without extensive revisions. He received, in 1879, a doctorate from the École Nationale Supérieure des Mines with a thesis on differential equations.

Following a brief appointment in mathematical analysis at the Université de Caen, in 1881 Poincaré joined the Université de Paris, where, during the rest of his life, he lectured and wrote prolifically—almost 500 papers—on mechanics and experimental physics, in all branches of pure and applied mathematics, and in theoretical astronomy. Changing his lectures every year, he would review optics, electricity, the equilibrium of fluid masses, the mathematics of electricity, astronomy, thermodynamics, light, and probability. Many of these lectures appeared in print shortly after they were delivered at the university. In his extensive writings on probability, Poincaré anticipated the concept of ergodicity that is basic to statistical mechanics.

Applying his mastery of analysis to the question of the solvability of algebraic equations, Poincaré developed before age 30 the idea of the automorphic function—one that is invariant under a group of transformations that are characterized algebraically by ratios of linear terms. He showed how these functions can be used to integrate linear differential equations with rational algebraic coefficients and how to express the coordinates of any point on an algebraic curve as uniform functions of a single algebraic variable (or parameter). Some of these automorphic functions he called Fuchsian, after the German mathematician Immanuel Lazarus Fuchs, who was one of the founders of the theory of differential equations; he found that they were associated with transformations arising in non-Euclidean geometry. In recognition of his fundamental contributions to mathematics, he was elected, in 1887, to membership in the Académie des Sciences in Paris.

Work in celestial mechanics

In celestial mechanics, Poincaré made substantial contributions to the theory of orbits, particularly the classical three-body problem (for example, the system involving the Sun, Moon, and Earth). This was part of the "problem of n bodies" (planets, stars, etc.), set for a prize by King Oscar II of Sweden: given the present masses, velocities, motions, and mutual distances of "n bodies," how long will they remain stable in their present spatial relationships, or will their orbits change at some future date? In his solution, Poincaré developed powerful new mathematical techniques, including the theories of asymptotic expansions and integral invariants, and made fundamental discoveries on the behaviour of the integral curves of differential equations near singularities. He was awarded the prize in 1889, even though his solution to the problem was only partially correct; in the same year he was also made a knight of the French Legion of Honour.

Poincaré summarized his new mathematical methods in astronomy in *Les Méthodes nouvelles de la mécanique céleste* (3 vol., 1892, 1893, 1899). Another result of this work, his *Analysis situs* in 1895, was an early systematic treatment of topology, which deals with properties of a system that endure when metric distortion occurs—that is, topology deals with the qualitative characteristics of spatial configurations that do not vary during cumulative transformations. He also contributed to the theory of numbers by demonstrating how the conception of binary quadratic forms, which was developed by the German mathematician Carl Friedrich Gauss, could be cast in geometric form. In 1904 he lectured at the St. Louis Exposition.

In mathematical analysis, Poincaré made important contributions to the theory of equilibrium of rotating fluid masses. In particular, he described the conditions of stability of the pear-shaped figures that played so prominent a part in the researches of later cosmogony, with reference to the evolution of celestial bodies. He attempted an application of these ideas to the stability of Saturn's rings and to the origin of binary stars. In 1906, in a paper on the dynamics of the electron, he obtained, independently of Einstein, many of the results of the special theory of relativity. The principal difference was that Einstein developed the theory from elementary considerations concerning light signalling, whereas Poincaré's treatment was based on the full theory of electromagnetism and was restricted to phenomena associated with the concept of a universal ether that functioned as the means of transmitting light, a concept that Einstein shared.

Books on the philosophy of science

After Poincaré achieved prominence as a mathematician, he turned his superb literary gifts to the challenge of describing for the general public the meaning and importance of science and mathematics. Always deeply interested in the philosophy of science, he wrote *La Science et l'hypothèse* (1903; *Science and Hypothesis*, 1905), *La Valeur de la science* (1905; *The Value of Science*, 1907), and *Science et méthode* (1908; *Science and Method*, 1914), all of which reached a wide public of nonprofessionals. His works were translated into English, German, Hungarian, Japanese, Spanish, and Swedish. He emphasized the subconscious, while probing the psychology of mathematical discovery and invention. He was a forerunner of the modern intuitionist school in that he believed that some mathematical induction is a priori and independent of logic. In his view, sudden illumination, following long subconscious work, is a prélude to mathematical creation. But his greatest contribution to philosophy was to emphasize the role played in scientific method by convention—*i.e.*, by the arbitrary choice of concepts.

Poincaré's prestige and influence increased in French science, and in 1906 he was elected president of the

Académie des Sciences and in 1908 to membership in the Académie Française, the highest honour accorded a French writer.

He died in Paris on July 17, 1912, from an embolism following prostate surgery.

BIBLIOGRAPHY. There is no definitive biography of Henri Poincaré, but ERIC T. BELL, *Men of Mathematics*, ch. 28 (1937, reprinted 1961), contains a biographical study of his work. A useful obituary notice is in the *Proc. R. Soc.*, series A, 91:5–16 (1915). The bibliography of ERNST LEBON, *Henri Poincaré: biographie, bibliographie analytique des écrits* (1909), was supplemented by R.C. ARCHIBALD, "Jules Henri Poincaré," *Bull. Am. Math. Soc.*, 22:125–136 (1915). A biographical article by VITO VOLTERRA, "Henri Poincaré," appeared as a Rice Institute pamphlet, vol. 1, no. 2, pp. 113–162 (1915). The collected works, *Oeuvres de Henri Poincaré*, were published by the *Académie des Sciences de Paris*, 10 vol. (1916–54). His writings on the philosophy of science and of mathematics can be found in *The Foundations of Science*, containing *Science and Hypothesis, The Value of Science*, and *Science and Method*, trans. by GEORGE BRUCE HALSTED (1913); and in *Dernières pensées* (1913; Eng. trans., *Mathematics and Science: Last Essays*, 1963).

(G.J.W.)

Poisonous Animals and Plants

The study of poisonous animals and plants, in the broadest sense, deals with the poisons produced by all living organisms. The science concerned with poisons produced by biological organisms is known as biotoxicology. The term poison refers to any substance that, when ingested, injected, absorbed, or applied to the body of an animal in relatively small quantities, may cause, by its chemical action, damage to a structure, a disturbance of function, or, if the quantity is sufficient, death.

GENERAL FEATURES

Properties and occurrence of biotoxins. The poisons produced by biological organisms are referred to as biotoxins, or biological poisons. Numerous micro-organisms, plants, and animals produce a vast array of toxic chemical substances, most of which have not yet been adequately characterized as to their chemical nature or biological effects. The classical definition of the term toxin referred only to poisons of microbial origin with high molecular weights (based on the weight of a hydrogen atom as 1) and antigenic properties (*i.e.*, substances to which the body reacts by producing antibodies, proteins that attack and neutralize the antigens). In recent years, however, the definition has been expanded to include poisons derived from higher plants and animals.

Definition of toxic substances

It is virtually impossible to assign a precise meaning to the term toxic. Obviously, almost any substance may be toxic if used in sufficient quantity. One property of a toxin that permits it to be so categorized, however, is that it is harmful to sensitive hosts when administered in relatively small amounts. The quantitative relationship between the amount of toxin administered and its effect on a host, or experimental subject, can be depicted by a graph called the dose–response curve. A potent toxin produces a steep slope in the linear portion of the curve (*i.e.*, a small amount of toxin has a great effect), whereas a less toxic substance results in a shallow-sloped linear portion of the curve.

Results of the various methods used to measure the dose response of a given host to a toxin depend upon a statistical treatment of one type or another. In most instances the effect measured is the death of the host. The response is greatest in the region of 50 percent mortality; the 50 percent lethal dose (LD_{50}), or the dosage that will result in the death of 50 percent of the tested animals, thus has been established as the most useful way to estimate toxicity. Unfortunately, animal hosts do not act as completely homogeneous populations, and considerable variation between groups usually occurs. Even when homogeneous groups of animals are used, individual variation in response to toxins is encountered from such factors as the health of the individual host.

Biotoxins can be conveniently grouped into three major categories: (1) microbial toxins, poisons produced by bacteria, blue-green algae, dinoflagellates, golden-brown algae, etc.; (2) phytotoxins, poisons produced by higher plants; and (3) zootoxins, poisons produced by animals. The term phytotoxin is also commonly used by plant pathologists to refer to substances that are toxic to plants. The term as used in this article, however, is restricted to poisons derived from plants that are toxic to animals and particularly to man. The geographical distribution of poisonous organisms varies greatly; poison-producing micro-organisms tend to be ubiquitous in their distribution. Poisonous plants and animals are found in greatest abundance and varieties in warm-temperate and tropical regions. Relatively few toxic organisms of any kind are found in polar latitudes.

Evolutionary aspects of biotoxins

Knowledge of the evolutionary significance and development of most biotoxins is largely speculative and poorly understood. In some instances they may have developed during the evolution of certain animal species as part of the food procurement mechanism (*e.g.*, in snakes; cnidarians, jellyfishes, and their relatives; mollusks, octopuses, and others; and spiders). Biotoxins may also function as defensive mechanisms, as in some snakes, fishes, arthropods (*e.g.*, insects, millipedes), and others. The defense may be quite complex—as in the protection of territorial rights for reproductive purposes—and inhibitory or antibiotic substances may be produced that result in the exclusion of competitive animal or plant species. Certain marine organisms and terrestrial plants may release into the water, air, or soil inhibitory substances that discourage the growth of other organisms; well-known examples include the production of antibiotic substances by micro-organisms. Similar chemical-warfare mechanisms are used in battles for territorial rights among the inhabitants of a coral reef, a field, or a forest. Thus biotoxins play important roles in the regulation of natural populations. Of increasing interest has been the discovery that certain substances, which may be toxic to one group of organisms, may serve a vital function in the life processes of the source organism.

The complex ecological relationships between some of the more toxic marine invertebrates and fishes are as yet poorly understood. Interactions between species may, in the presence of certain chemical substances, contribute directly to the production of highly toxic biochemical products, an interaction that is sometimes referred to as ecological biogenesis.

Importance of biotoxins to man. Toxic plants and animals are of importance to man in a variety of ways. Many biotoxins are potent agents capable of adversely affecting man directly by causing disease, pain, or death or indirectly by damaging his plants or animals. Toxic organisms thus constitute a potential public-health problem with economic and sociological implications.

Venom-producing animals and stinging and dermatogenic (*i.e.*, skin-poisoning) plants capable of inflicting pain and sometimes death by means of parenteral contact (*i.e.*, by bringing poisons into the body other than through the digestive tract) constitute environmental hazards. Biotoxic agents may produce their injurious effects by becoming involved in man's food supply; ingestion of a poisonous microbial organism, plant, or marine animal or one of their toxic by-products may cause intoxication. An example is that of the shore fishes of many tropical islands; otherwise valuable food fishes are frequently contaminated by a poison called ciguatoxin. The poison, a potent neurotoxin (nerve poison), is accidentally ingested by the fishes in their food; such fish can no longer be used for either human or animal consumption.

Entry methods for biological poisons

Some of the effects produced by biotoxins on man are of an acute nature, and the injuries they cause are readily discernible. The effects of some of the mycotoxins (poisons produced by fungi) and poisons produced by higher plants, however, are long-term and chronic; they result in the development of cancerous growths and other chronic degenerative changes that are sometimes difficult to detect.

The scientific importance of biotoxicology is frequently misunderstood. One popular attitude is that, because poisons are potentially lethal substances capable of causing

intoxication and death, they are therefore to be avoided and that research is justified to determine their pharmacological and chemical properties with the hope of ultimately developing useful antitoxins. Another aspect equally deserving of consideration, however, is the exploration of biologically active substances and chemical structures in order to determine their physical nature and biological properties. Many of the biotoxins studied even in a cursory manner have been found to have diverse biological properties, some of which have application to medicine: fungicidal, growth inhibitory, antiviral, antibiotic, antitumour, hemolytic (red-blood-cell destroying), anti-inflammatory, analgesic (pain controlling), cardioinhibitory (heart arresting), psychopharmacological (affecting mental processes), and others. Toxicity thus provides important information regarding valuable pharmacological properties of biotoxins.

In the area of natural-products research, of which biotoxicology is an important segment, ethnic data provides insight into biological activity. Ethnobotanical and ethnozoological studies reveal the way in which plants and animals have been utilized by primitive peoples; crude data of this type serve as useful indicators of the general pharmacological properties of biotoxins. Biotoxins also offer a spectrum of new or little-known chemical substances, which the chemist may ultimately be able to synthesize and manipulate into useful products. Highly active biotoxins may eventually be useful in neurophysiological research, especially in elucidating the nature of both nervous activity and muscular function.

Medical uses of biotoxins

The study of biotoxicology ultimately will yield a vast amount of valuable data in such diverse areas as nutrition, medicine, chemistry, pharmacology, biology, ecology, military science, and drug development.

HISTORICAL BACKGROUND

The story of poisonous plants and animals and their relationship to mankind dates to the dawn of man's written language; poisons and poisonous organisms played a dominant role in the affairs of early man, especially in those of ancient magician-physicians and priests. Not only was it of value for them to know the means of treating a poisonous sting but it was equally useful that they be familiar with the biological poisons capable of causing death of an enemy. Because of the profound and little-understood physiological effects resulting from contact with plant and animal poisons, ancient man often surrounded biotoxins with an aura of mystery, superstition, and religion. The pre-Grecian era of Mesopotamian medicine was also dominated by magicians and priests. Little is known about their knowledge of specific poisonous animals, but they were keen observers of natural history and were undoubtedly aware of many noxious pests. Medical records of ancient Egypt include a practicing physician's formulary containing about 260 prescriptions, some of which are for various kinds of animal stings, including the poisonous puffer fish. The Bible and Talmudic records document the toxic effects of dinoflagellates (a group of algae). Other records list some 760 medicinal plants and mention venomous animals and antidotes for bites and stings.

Out of ancient Chinese science, which was based on plant and animal lore, demoniac beliefs, and faith healing, arose a legendary emperor revered as god of medicine and father of agriculture and herbal therapy; he is credited with the authorship of the predecessor of Chinese pharmacopeias. Ancient Chinese phytotoxicology was a comparatively well-developed science, and poisonous plants and their antidotes were described in elaborate detail.

In the Western world the ancient Greeks were probably the first to dissociate medicine from magic and religion. Aristotle was familiar with the venom of jellyfishes and scorpion fishes. Mithradates VI, king of Pontus in the 2nd century BC, experimented with the giving and taking of poisons and is said to have been one of the first to study intensively the art of poisoning and the preparation of antidotes. He aspired to develop a universal antidote, which became known as mithridates and theriacs. In the 1st century AD, Pedanius Dioscorides, a Greek army surgeon, wrote on many terrestrial plants used as drugs and also recorded the toxic nature of sea hares, polychaete worms, stingrays, and sea vipers. Pliny the Elder, the Latin naturalist, dealt with poisonous plants and animals in his *Natural History* (*Historia naturalis*) and wrote about the dangerous wounds inflicted by the venomous spines of the weever fishes and stingrays and about the toxicity of the sea hare. He also recommended that a powder of the sting of the stringray be used as a potion to treat toothache.

From the 5th to the 8th century the centre of medical culture was Constantinople; Paulus of Aegina wrote an *Epitome* on medical subjects comprised of seven books, one of which was concerned with venomous stings, bites, and toxicology. During the early years of the 7th century, the Arabs wrote extensively about poisons, the effect of venoms, and antidotal remedies.

From the 9th through the 15th century, when there was neither induction nor experiment, biotoxicology languished along with most other sciences. One of the few important works was that of the Persian physician Avicenna. His *Qānūn fī aṭ-ṭibb* (*Canon of Medicine*) was a compilation of his experiences in the practice of medicine and a compendium of all that was known about medicine at the time. The *Canon* is concerned with, among other things, the composition and preparation of drugs. He recognized both oral and parenteral poisons, discussed venomous stings and their treatment, and dealt with animal, plant, and mineral poisons. For more than 500 years his work was considered the outstanding authority on poisons and their antidotes.

Avicenna's *Canon of Medicine*

Many printed herbals, which began to appear around 1470, including discussions about poisonous plants. The great work in biotoxicology of the Renaissance was that of the father of modern biotoxicology, the French physician Jacques Grevin, who published his book *Deux livres des venins* in 1568. Another important work that appeared in the latter half of the 17th century dealt with various types of poisonous plants.

In 1756 the British physician Richard Mead described snake poisoning in his *Mechanical Account of Poisons* in several essays and how he had swallowed the venom, thereby proving that parenteral administration was in this case essential to produce intoxication. Other important works of this period were influential because they paid particular attention to poisonings resulting from the injudicious use of medicinal plants and to some of the effects of poisons on animals.

The important toxicological contributions of the French physician M.J.B. Orfila, who has been designated the father of experimental toxicology, marked the opening of the 19th century. His *Traité de toxicologie*, first published in 1814, went through many editions and translations. Most of his investigations were concerned with inorganic agents and plant poisons and, to a lesser extent, with the venoms of snakes, spiders, bees, wasps, and poisonous fishes. In 1888 occurred the first demonstration that the disease-producing mechanism of a micro-organism in man could be explained in terms of a toxic substance released by bacteria. Tetanus toxin was discovered in 1890.

Experimental research based on sophisticated analytical procedures accelerated in the early part of the 20th century. This period also saw the discovery of hypersensitivity reactions and a better understanding of anaphylaxis and allergies. With the advent of World War II, environmental biological poisons became medically important in many military operations. A rapidly expanding world population in the postwar years further pointed up the need for the utilization of food resources in heavily populated tropical regions in which oral biotoxins are most likely to be encountered. There have also been increased demands for new therapeutic agents, and biotoxins have served as a rich and varied resource.

MAJOR CATEGORIES OF BIOTOXINS

Microbial toxins. Microbial poisons are produced by the so-called Monera (bacteria and blue-green algae) and

Protista (fungi, dinoflagellates, golden-brown algae, and others). Various classifications have been proposed for the microbial poisons, but none is entirely satisfactory. The problems encountered when dealing with these organisms result from a lack of precise knowledge concerning their biological nature and their phylogenetic relationships; in addition, their poisons show great diversity and chemical complexity. The following outline, however, is useful in dealing with this subject.

Bacterial toxins. The prefixes "exo-" and "endo-" are retained in classifying the bacterial toxins mainly for historical reasons rather than because they are found either outside or inside the bacterial cell. The main differences in these toxins lie in their chemical structure. A list of representative species of toxic bacteria and their toxins is given in Table 1.

Table 1: Representative Toxic Bacteria and Their Toxins

	toxin	chemical nature	comments
Gram-positive bacteria			
Corynebacterium diphtheriae	diphtheria toxin	protein	involved in diphtheria in man
Staphylococcus aureus	α-toxin, β-hemolysin, enterotoxins, leukocidin	protein	causes pyogenic (pus-forming) infections in animals and in man
Clostridium botulinum	botulinus toxin	protein	involved in botulism in man
Clostridium perfringens	α-toxin, θ-toxin	protein	causes gas gangrene in man
Clostridium tetani	tetanospasmin	protein	causes tetanus in man
Streptococcus pyogenes	streptolysin O, erythrogenic toxin	protein	causes scarlet fever and other streptococcal infections in animals and man
Gram-negative bacteria			
Shigella dysenteriae	neurotoxin	protein	causes bacillary dysentery in man
Vibrio cholerae	cholera toxin	protein	causes cholera in man
Escherichia coli, *Salmonella* species, *Shigella* species, *Brucella* species, and others	"conventional endotoxins"	complex of protein, lipid, and polysaccharide	most gram-negative bacteria produce toxic complex chemical mixtures that are categorically referred to as "endotoxins"

Bacterial exotoxins

Poisonous proteins from bacteria are sometimes referred to as bacterial exotoxins. The exotoxins are generally produced by gram-positive organisms (*i.e.*, bacteria that react in certain ways to the staining procedure known as gram staining); at least two bacteria, *Shigella dysenteriae* and *Vibrio cholerae*, that produce exotoxins are gram-negative, however. The exotoxins usually do not contain any nonprotein substances, and most are antigenic; *i.e.*, they stimulate the formation of antibodies. The exotoxins may appear in the culture medium in which the bacteria are growing during the declining phases of growth; in some cases they are released at the time of normal destruction of the cells after death (autolysis). The exotoxins, less stable to heat than the endotoxins, may be detoxified by agents that do not affect endotoxins. They are more toxic than endotoxins, and each exotoxin exerts specific effects, collectively known as pharmacological properties. Exotoxins are neutralized by homologous antibodies—*i.e.*, the active agents in blood serum produced by a process involving the bacteria against which the serum is to be used.

Endotoxins are antigens composed of complexes of proteins, polysaccharides (large molecules built up of numerous sugars), and lipids (fats). The protein part determines the antigenicity, or quality of being reacted against as a foreign substance in a living organism. The polysaccharide part determines the immunological specificity, or limitations on the types of antibodies that can react with the endotoxin molecule and neutralize it (the immunological reaction). Some of the lipids possibly determine the toxicity. Endotoxins are derived from the bacterial cell wall and, when cells are grown in culture, are released only on autolysis. Endotoxins are not neutralized by homologous antibodies and are relatively stable to heat; all of them have the same pharmacological properties.

Mycotoxins. Fungi are plantlike members of the phylum Mycota that do not contain any chlorophyll. It has been estimated that there are about 85,000 species of fungi, about 55,000 of which are microscopically small in size. A significant number are known to produce poisons of various types. Toxic fungi can be roughly divided into two main categories on the basis of their size: the smaller microfungi and the larger mushrooms. The toxic microfungi are members of one of two classes: Ascomycetes, or the sac fungi, and the Deuteromycetes, or the imperfect fungi (*i.e.*, fungi in which no sexual reproductive stages are known). The large toxic mushrooms, or toadstools, are mostly members of the class Basidiomycetes, although some Ascomycetes, such as the poisonous false morel (*Gyromitra esculenta*), may attain a size as large as some of the mushrooms.

The ability of fungi, such as ergot (*Claviceps purpurea*) and some mushrooms, to produce intoxication has long been known. During the 19th century it was recognized that molds are responsible for such diseases as yellow-rice toxicoses in Japan and alimentary toxic aleukia in Russia. The eruption of so-called turkey X disease in England in 1960 and the resulting discovery of the substance known as aflatoxin (see Table 2) stimulated study of the subject of mycotoxicology. Despite the vast amount of research on mycotoxins, much remains to be done. Mycotoxins are continuously being discovered, and their chemical and biological properties, which vary greatly from one poison to the next, are being studied with great interest. Because mycotoxins have now been recognized as potential cancer-producing agents (carcinogens) that become in-

Table 2: Representative Toxic Microfungi

	toxin	comments
Claviceps purpurea (ergot)	"ergotoxine," a complex of toxic alkaloids, ergocryptine, ergocornine, ergocristine, and others	causes ergot poisoning in animals and humans; produces vomiting, abdominal pain, numbness, nervous disorders, convulsions, gangrene, and abortion
Stachybotrys alternans	stachybotryotoxin	causes a toxicosis in animals and humans; produces stomatitis (inflammatory diseases of the mouth), rhinitis (inflammation of the mucous membranes of the nose), conjunctivitis (inflammation of the inner surface of the eyelid); failure of blood to clot, blood abnormalities, neurological disturbances, and death
Aspergillus flavus and other species *Penicillium* species	aflatoxin complex (16 or more known toxins)	causes toxicosis in animals and possibly man; toxins damage liver and kidneys; aflatoxin is one of the most potent liver-cancer producing agents known
Fusarium sporotrichioides and other *Fusarium* species	fusariogenin, epicladosporic acid, fagicladosporic acid	causes alimentary toxic aleukia in animals and man; produces burning sensation of the mouth, tongue, throat, and stomach, nausea, vomiting, headache, cold extremities, hemorrhagic spots, convulsions, anemia, gangrene, death
Cladosporium epiphyllum and other species of *Cladosporium*	same as *Fusarium* species	
Pithomyces chartarum (*Sporidesmium bakeri*)	sporidesmin	causes facial eczema (an eruptive severe rash) in cattle; produces sensitization of the skin to sunlight resulting in scab formation and sores; there may also be severe liver damage
Fusarium species, *Rhizopus* species, *Aspergillus* species, *Penicillium islandicum*, and others	luteoskyrin, "islanditoxin," citrinin, citreoviridin, and others; a large complex of poisons involved	causes moldy or yellowed rice, which is toxic to animals and man; the effects in man have not been well defined; may cause nausea, vomiting, diarrhea, prostration, liver damage, and death; the effects vary greatly because of the various poisons involved

Table 3: Representative Poisonous Mushrooms

	toxin	type of poisoning
Lorchel, or false morel (*Helvella esculenta*)	gyromitrin	toxicity to people is variable; causes severe liver damage accompanied by nausea, vomiting, abdominal pain, jaundice, enlarged and tender liver, coma, convulsions; fatality rate about 15 percent
Fly mushroom, or fly agaric (*Amanita muscaria*)	muscarine	symptoms develop rapidly, are severe, consisting of severe gastrointestinal disturbances, delirium, hallucinations, convulsions; rarely causes death
Panther mushroom (*Amanita pantherina*)	muscarine	symptoms are similar to *A. muscaria* poisoning
Destroying angel, or death cup (*Amanita phalloides*)	amanitine, phalloidine	symptoms develop slowly, about 6–15 hours after eating: extreme abdominal pain, nausea, vomiting, excessive thirst, anuria (absence or defective excretion of urine), prostration, weakness, jaundice, cyanosis, convulsion, death; fatality rate up to 90 percent; no known antidote
Boletus miniato-olivaceus	muscarine (?)	causes gastrointestinal and visual disturbances
Jack-o'-Lantern fungus (*Clitocybe illudens*)	muscarine	causes gastrointestinal upset, not fatal
Inky cap (*Coprinus atramentarius*)	unknown	some people experience a peculiar type of intoxication after eating this mushroom then drinking an alcoholic beverage: giddiness, gastrointestinal upset, prostration, and tachycardia (rapid heart action); the alcohol is believed to increase the solubility and absorption of the poison
Entoloma lividum	unknown	causes gastrointestinal upset, usually not fatal
Inocybe patouillardi	muscarine	symptoms are similar to *A. muscaria* poisoning
Lepiota morgani	unknown	causes gastrointestinal upset; fatalities have been reported
Mexican hallucinogenic mushroom (*Psilocybe mexicana*)	psilocybin, psilocin	causes euphoria, loss of sense of distance and size, and hallucinations

volved in man's food supply, they are important in the study of environmental carcinogenesis.

Poisonous mushrooms

Poisonous mushrooms, or toadstools as they are commonly called, are the widely distributed members of the class Basidiomycetes. Although there are more than 30,000 species of large mushrooms, only a few are known to be poisonous when eaten (see Table 3); some of the poisons, however, are deadly. About 90 percent of the deaths attributed to mushroom poisoning result from eating members of the genus *Amanita*. When eating wild mushrooms, they should be accurately identified by an experienced person; the safest procedure is to eat only cultivated species. The problem of toxicity in mushrooms is complex; no single rule or test method exists by which the toxicity of a mushroom can be determined. The most poisonous species closely resemble some of the most prized edible species; in addition, toxicity within a given wild species may vary from one set of ecological conditions or from one geographical locality to the next. Moreover, although some mushrooms that are poisonous when fresh are edible when cooked, dried, salted, or preserved in some other way, others remain poisonous in spite of all preparation procedures. It has also been observed that some people may become poisoned by eating mushrooms that apparently do not affect others. As with microfungi, the mushroom poisons vary in their chemical and biological properties from species to species.

Algal poisons. The study of poisonous algae is known as phycotoxicology. Two major groups of algae produce toxic substances: the Cyanophyta, or blue-green algae, and the Pyrrophyta, or dinoflagellates.

Environmental extremes for blue-green algae

Algae comprise some of the smallest forms of chlorophyll-bearing organisms. The approximately 1,500 species of blue-green algae are among the most primitive organisms; they resemble the bacteria in their mode of reproduction and in certain physiological functions. The blue-green algae, among the most widely distributed of all organisms, live in all parts of the aquatic environment that receive light, including the exceedingly saline waters of the Dead Sea. They have extreme temperature tolerances and inhabit waters whose temperatures vary from near the freezing point to those of thermal springs as hot as 82° C (180° F). Blue-green algae have a remarkable ability to survive drying for years under desert conditions. About twice as many species are found inhabiting marine waters as freshwaters. Some strains of a species are toxic; other strains of the same species are not. Water blooms of blue-green algae have been responsible for the death of fishes, waterfowl, cattle, horses, swine, and other animals. Blue-green algae have also been implicated as causes of human intoxications. A list of representative poisonous blue-green algae and their poisons appears in Table 4.

The approximately 1,100 species of dinoflagellates, important producers of the primary food supply of the sea, are microscopic one-celled organisms that are dependent upon various inorganic nutrients in the water and upon radiant energy for photosynthesis, the process by which they produce their own food supplies. Although dinoflagellates inhabit both marine waters and freshwaters, most species are marine. Dinoflagellates are most often found in cool or temperate waters. During periods of planktonic blooms (times of high concentrations of microscopic organisms in the water) dinoflagellates multiply in large numbers. These planktonic blooms, sometimes referred to as red tide because they discolour the water, are often associated with weather disturbances that may bring about changes in water masses or upwellings. During periods of bloom large numbers of toxic dinoflagellates may be ingested by shellfish; the poisons accumulate in their digestive glands. Man and animals may in turn be poisoned by eating poisoned shellfish. Certain species of dinoflagellates are capable of producing some of the most toxic substances known. The two species of dinoflagellates most commonly involved in human intoxications have been *Gonyaulax catenella* along the Pacific coast of North America and *G. tamarensis* along the eastern coast of North America. Intoxications from these organisms are known as paralytic shellfish poisoning. The symptoms, which begin with a tingling or burning sensation, then numbness of the lips, gums, tongue, and face, gradually spread. Gastrointestinal upset may be present. Other symptoms include weakness, joint aches, and muscular paralysis; death may result. There is no specific treatment nor antidote. The poison, variously called paralytic shellfish poison, mussel poison, and saxitoxin, is a complex nonprotein nitrogen-containing compound. Paralytic shellfish poisoning is best avoided by following local public-health quarantine regulations. Despite the numerous folk tales as to methods for differentiating poisonous from edible mussels, there is no reliable substitute for laboratory toxicity tests.

Table 4: Representative Blue-green Algae and Their Poisons

	toxin	toxic effects
Anacystis cyanea (also called *Microcystis aeruginosa*), a freshwater alga	neurotoxin (a polypeptide)	ingestion causes mass mortality of fish; strychnine-like convulsions in cattle and death; respiratory distress, convulsions and death in laboratory animals; gastroenteritis (inflammation of intestinal tissues) in humans; poison is rapid acting
Anabaena flos-aquae, a freshwater alga	neurotoxin (a polypeptide)	toxic effects are similar to those above; poison is rapid acting
Aphanizomenon flos-aquae, a freshwater alga	neurotoxin (a polypeptide ?)	toxic effects are similar to above, but the poison is usually slower acting
Lyngbya majuscula, a marine and freshwater blue-green alga	dermatogenic toxin, chemistry unknown	causes an acute inflammatory reaction when brought in contact with human skin; toxin has antiviral, antibacterial, and antifungal activity

Symptoms of paralytic shellfish poisoning

Table 5: Representative Poisonous Plants

name and distribution	toxic principle	toxic effects and comments
Plants poisonous to eat		
Rosary pea, or jequirity bean (*Abrus precatorius*); tropical regions	abrin (*N*-methyltryptophan) and abric acid	onset of symptoms may be delayed several hours to two days, vomiting, diarrhea, acute gastroenteritis, chills, convulsions, death from heart failure; one seed chewed may be fatal to a child
Aconite, or monkshood (*Aconitum napellus*); North America, Europe	aconite and a complex of other alkaloids	tingling, burning sensation in tongue, throat, skin, restlessness, respiratory distress, muscular uncoordination, vomiting, diarrhea, convulsions, possible death; an extremely poisonous plant
Corn cockle (*Agrostemma githago*); North America, Europe	githagin, agrostemmic acid (saponins)	dizziness, diarrhea, respiratory distress, vomiting, headache, sharp pains in spine, coma, death; frequent ingestion of small amounts results in chronic githagism (a disease similar to lathyrism, that results in pain, burning and prickling sensations in lower extremities, and increasing paralysis); milled seeds may be found in wheat flour
Purple loco weed (*Astragulus mollissimus*); Northern Hemisphere	locoine	dullness, weakness, irregular behaviour, impaired vision, edema of eyelids, loss of muscular control, loss of appetite, emaciation, starvation, death in sheep, horses, and cattle
Belladonna (*Atropa belladonna*); United States, Europe, Asia	hyoscyamine, atropine, hyoscine, and a complex of other alkaloids	dryness of the skin, mouth, throat, difficulty in swallowing, flushing of the face, cyanosis (a bluish discoloration of skin due to insufficient oxygen), nausea, vomiting, slurred speech, coma, death; children and animals frequently poisoned by eating fruit
Akee (*Blighia sapida*)	hypoglycin A, B	sudden vomiting, drowsiness, muscular and nervous exhaustion, prostration, coma, death
Rape (*Brassica napus*)	glycosides (isothiocyanates)	pulmonary emphysema, respiratory distress, anemia, constipation, irritability, blindness in cattle
Marihuana (*Cannabis sativa*); United States, Mexico, tropical America	cannabinol, canabidiol, and related compounds	exaltation, inebriety, confusion, followed by central nervous system depression; prolonged addiction may produce dullness or mania; excessive quantities may cause death as a result of cardiac depression
Water hemlock (*Cicuta maculata*); northern temperate regions	cicutoxin	abdominal pain, nausea, vomiting, diarrhea, respiratory distress, hypersalivation, convulsions, death; among the most poisonous plants
Poison hemlock (*Conium maculatum*); temperate United States, South America, northern Africa, Asia	coniine, conhydrine, *N*-methyleoniine, coniceine, and other alkaloids	muscular weakness, paralysis of extremities, blindness, respiratory paralysis, death; responsible for many human fatalities; leaves most toxic when plant is flowering
Purging croton (*Croton tiglium*); Asia, Pacific Islands, Africa	croton, croton resin, ricinine	vomiting, violent purging, collapse, death; croton oil is also a skin irritant, causing reddening, swelling, and pustules
Daphne (*Daphne mezereum*); temperate regions	glycoside involving aglycone dihydroxycoumarin	vomiting, burning sensation of the mouth, ulceration of the oral mucosa, diarrhea, stupor, weakness, convulsions, and death
Jimson weed or thornapple (*Datura stramonium*); temperate and tropical regions	hyoscine, hyoscyamine, atropine	headache, nausea, vomiting, dizziness, thirst, dry and burning sensation in skin, mental confusion, mania, loss of memory, convulsions, death; children are often poisoned by eating seeds or sucking flowers
Larkspur (*Delphinium* species); northern temperate regions	delphinine, delphinoidine, delphisine, and other alkaloids	burning and inflammation of mouth, nausea, vomiting, respiratory distress, itching, cyanosis; one of the greatest causes of death in grazing livestock
Dumbcane (*Dieffenbachia seguine*); widely cultivated in temperate regions, tropical regions	protoanemonine, calcium oxalate	irritation and burning of the mouth, tongue, and lips, hypersalivation, swelling of the tongue, difficulty in swallowing and breathing
Foxglove (*Digitalis purpurea*); Europe, North America	glycosides, digitoxigenin, and others	loss of appetite, nausea, vomiting, slow pulse and irregular heartbeat, diarrhea, abdominal pain, headache, fatigue, drowsiness, convulsions, death
Wild yam (*Dioscorea hispida*); southern Asia, Pacific Islands	dioscorine	discomfort, then burning of the throat, giddiness, vomiting of blood, respiratory distress, drowsiness, exhaustion, paralysis of the nervous system, death; raw tubers are a frequent cause of death in the Philippines
Huanuco cocaine (*Erythroxylon coca*); tropics of both hemispheres	cocaine and other alkaloids	central nervous system stimulation followed by depression, numbness of tongue, paralysis of respiratory centres, cyanosis, respiratory distress, death; leaves are commonly chewed by Indians of Peru and Bolivia as a stimulant
Manchineel (*Hippomane mancinella*); Florida, Central America, South America, West Indies	physostigmine or a similar alkaloid plus a sapogenin	fruit causes gastroenteritis, which may be fatal, and causes ulceration of intestinal tract; sap causes burning of skin, swelling and hemorrhage of the eyes; sap is used as an arrow poison
Black henbane (*Hyoscyamus niger*); North America, Europe, Asia, Oceania	hyoscyamine, hyoscine, atropine, and other alkaloids	similar to belladonna poisoning caused by *Atropa belladonna;* children are poisoned by eating seeds and pods
Barbados nut (*Jatropha curcas*); tropics	curcin	burning of the throat, bloating, dizziness, vomiting, diarrhea, drowsiness, dysuria, leg cramps, violent purgative action; may be fatal to children
Mountain laurel (*Kalmia latifolia*); North America	andromedotoxin	hypersalivation, tears, impaired vision, tingling of skin, dizziness, vomiting, muscular paralysis, convulsions, coma, death; children are poisoned by eating leaves
Grass peavine (*Lathyrus sativus*); North America, Europe, northern Africa, Asia	β-aminopropionitrile	back pain, weakness in legs, paralysis; has caused death in children

Respiratory irritation may result from the inhalation of toxic products in the windblown spray from red-tide areas containing the toxic dinoflagellate *Gymnodinium brevis*, which is found in the Gulf of Mexico and Florida; the nature of the poison is unknown. Deaths of large numbers of brackish-water pond fishes because of *Prymnesium parvum* have been reported in Israel; the poison is known as prymnesin.

Plant poisons (phytotoxins). The study of plant poisons is known as phytotoxicology. Most of the poisonous higher plants are angiosperms, or flowering plants. Of the 300 families of angiosperms representing about 200,000 species, only a small percentage are recognized as poisonous. Several systems have been devised for the classification of poisonous plants, none of which is completely satisfactory. Poisonous plants may be classified according to the chemical nature of their toxic constituents, their phylogenetic relationship, or their botanical characteristics. The following classification, which is based on their toxic effects, has been found to be useful: (1) plants that are poisonous to eat; (2) plants that are poisonous upon contact; (3) plants that produce photosensitization; and (4) plants that produce airborne allergies (see Table 5).

Plant poisons, or phytotoxins, comprise a vast range of biologically active chemical substances, such as alkaloids, polypeptides, amines, glycosides, oxalates, resins, toxal-

Table 5: Representative Poisonous Plants (continued)

name and distribution	toxic principle	toxic effects and comments
Cassava (*Manihot esculenta*); tropics	cyanophoric glycosides	nausea, respiratory distress, twitching, staggering, convulsions, coma, death
Chinaberry (*Melia azedarach*); North America, southern Africa, Asia	azadarin	stomatitis with violent and bloody vomiting, paralysis
Opium poppy (*Papaver somniferum*); Europe, Asia, tropics	morphine, codeine, thebaine, papavarine, narcotine	central nervous system depression, pinpoint pupils, depressed respiration, cyanosis, coma, death
Pokeberry (*Phytolacca americana*); North America, Europe, southern Africa	phytolaccine	burning, bitterness in mouth, vomiting, purging, spasms, convulsions, death
Castor bean (*Ricinus communis*); United States, tropics	ricin, a toxalbumin	burning of mouth, throat, and stomach, vomiting, diarrhea, abdominal cramps, dulled vision, convulsions, respiratory distress, paralysis, death; one to three seeds may be fatal in children
Deadly nightshade (*Solanum nigrum*); North America, Europe	solanine, a glycoalkaloid	nausea, vomiting, abdominal pain, diarrhea, trembling, paralysis, coma, death
Nuxvomica tree (*Strychnos nux-vomica*); India, Hawaii	strychnine, brucine, and other alkaloids	extreme irritability, tonic convulsions, exhaustion, sweating, muscular rigidity, respiratory distress, death
Thevetia (*Thevetia peruviana*); tropics	thevetin, thevetoxin, neriifolin	vomiting, diarrhea, high blood pressure, erratic heartbeat, respiratory distress, heart paralysis, death
False hellebore (*Veratrum viride*)	protoveritrine, germerine, jervine, and other alkaloids	vomiting, abdominal pain, muscular weakness, convulsions, respiratory distress, coma, death
Death camas (*Zigadenus* species)	zygadenine	vomiting, hypersalivation, staggering, paralysis, convulsions, coma, death
Plants poisonous by contact		
Euphorbia, spurge (*Euphorbia* species); worldwide	a complex of substances including alkaloids, glycosides, and others	eye irritation, blindness, blistering of the skin, swelling around the mouth, burning of the mouth, unconsciousness, death; milky sap is used as an arrow poison
Spurge nettle (*Jatropha stimulosus*); North America, Europe, Asia	toxin unknown	contact produces instant, intense stinging and itching due to an irritating substance injected into the skin by the stinging hairs; results in a skin eruption of minute red papular (small conical elevations of the skin) rash, which wears off in about 30 minutes; a dull purplish discoloration of the skin may remain for several weeks
Shiney-leaf stinging tree, tree nettle (*Laportea photiniphylla*); Australia	5-hydroxytryptamine (and other toxic substances?)	contact with the stinging hairs of this plant produces intense rapidly spreading pain, reddened rash, and later a severe skin eruption; severe stings may result in the victim dashing about, rolling on the ground, and crying out with pain; fatalities have been reported; dried leaves may cause intense sneezing
Poisonwood (*Metopium toxiferum*); West Indies, Florida	similar to poison ivy	contact with any part of the tree, especially sap, turns the skin black, causes a rash, blisters, etc; smoke from a burning tree is very irritating causing illness and temporary blindness
Strophanthus (*Strophanthus* species); Florida, tropical America, Africa	an alkaloid, trigonelline, and a large number of cardiac glycosides and aglycones	vomiting, slow and irregular pulse, blurred vision, delirium, circulatory failure, death; used as an arrow poison
Curare (*Strychnos toxifera*); Central America and northern South America	toxiferines, caracurines, and other alkaloids	haziness of vision, relaxation of facial muscles, inability to raise head, loss of muscle control of arms, legs, and respiratory muscles, death; used as a poison for arrows and for blow-gun darts
Poison ivy (*Toxicodendron radicans*, also called *Rhus toxicodendron*); North America	urushiol	skin irritation, swelling, blistering, itching; may be fatal to young children; smoke from burning plant is toxic
Plants that produce photosensitization		
Buckwheat (*Fagopyrum sagittatum*); North America, Europe	fagopyrin, a naphthodianthrone derivative	ingestion of the leaves by animals causes liver dysfunction, thereby resulting in deposition of a photosensitizing pigment in the skin; sunlight then causes redness of the skin, nervousness, swelling of the eyelids, convulsions, and prostration in farm animals
St. Johnswort (*Hypericum perforatum*); North America, Europe	hypericin, a naphthodianthrone derivative	similar to buckwheat
Plants that produce airborne allergies		
Box elder (*Acer negundo*); Northern Hemisphere	oleoresin and a water soluble antigen	hay fever (respiratory allergy) may also cause an eczematous dermatitis of the exposed parts of the body
Poplar, cottonwood (*Populus* species); almost worldwide	same as box elder	
Ash (*Fraxinus* species); temperate regions	same as box elder	
Elm (*Ulmus* species); Northern Hemisphere, tropics	same as box elder	

Chemical variability of plant poisons

bumins, and a large group of miscellaneous compounds whose chemical structure has not yet been determined. Alkaloids, most of which are found in plants, are characterized by the presence of nitrogen and their ability to combine with acids to form salts. They are usually bitter in taste. It has been estimated that about 10 percent of the plant species contain some type of alkaloid. Only a few of the 5,000 alkaloids characterized thus far do not produce any biological activity; most cause a strong physiological reaction when administered to an animal. Amines are organic compounds containing nitrogen. A polypeptide is a string of three or more amino acids. A few polypeptides and amines are toxic to animals. Some glycosides, which are compounds that yield one or more sugars and one or more other compounds—aglycones (nonsugars)—when hydrolyzed (chemically degraded by the introduction of water molecules between adjacent subunits), are extremely toxic to animals. Toxicity resides in the aglycone component or a part of it. Oxalates are salts of oxalic acid, which under natural conditions is not toxic but becomes so because of the oxalate ion. Resins, a heterogeneous assemblage of complex compounds, differ widely in chemical properties but have certain similar physical properties. Some resins are physiologically very active, causing irritation to nervous and muscle tissue. Toxalbumins are highly toxic protein molecules that are

Table 6: Representative Animals Poisonous When Eaten

name and distribution	toxic principle	toxic effects and comments
Protozoans—one-celled animals		
Dinoflagellate (*Gymnodinium brevis*); Gulf of Mexico, Florida	unknown	irritation of mucous membranes of nose and throat; causes sneezing, coughing, respiratory distress due to inhalation of windblown spray from red tide areas
Dinoflagellate (*Gonyaulax catenella*); Pacific coast of North America	paralytic shellfish poison, saxitoxin ($C_{10}H_{15}N_7O_8 \cdot 2\ HCl$)	tingling, burning sensation and numbness of lips, tongue, face, spreading elsewhere to the body; weakness, dizziness, joint aches, hypersalivation, intense thirst, difficulty in swallowing, muscular paralysis, and death; extremely toxic; usually involved with the eating of shellfish
Dinoflagellate (*Gonyaulax tamarensis*); east coast of North America	same as *Gonyaulax catenella*	that have been feeding on toxic dinoflagellates
Cnidarians—jellyfish, sea anemones, and relatives		
Matamalu samasama, a sea anemone (*Radianthus paumotensis*); southern Polynesia to Micronesia	poison unknown	nausea, vomiting, abdominal pain, cyanosis, prostration, coma, death; this species of sea anemone eaten raw or cooked is poisonous
Echinoderms—starfishes, sea cucumbers, and relatives		
Sea cucumber (*Holothuria* species); tropical Pacific	holothurin	little is known concerning the symptoms produced by poisonous sea cucumbers; ingestion of toxic sea cucumbers may be fatal; sea cucumbers under most circumstances are safe to eat, but some species are toxic; species are difficult to identify
Mollusks—octopus, squid, shellfish, and others		
California mussel (*Mytilus californianus*); Pacific coast of North America	paralytic shellfish poison, saxitoxin	these mollusks become poisonous to eat because of feeding on toxic dinoflagellates; symptoms same as for dinoflagellates (*Gonyaulax* species)
Butter clam (*Saxidomus giganteus*); Alaska to California	same as California mussel	
Surf clam (*Spisula solidissima*); Labrador to North Carolina	same as California mussel	
Whelk (*Neptunea* species); Europe, Pacific region	tetramine	nausea, vomiting, diarrhea, weakness, fatigue, dizziness, photophobia (intolerance to light), impaired vision, and dryness of the mouth; poison is believed to be restricted to the salivary glands of the whelk
Turban shell (*Turbo argyrostoma*); tropical Pacific Ocean	poison believed to be related to ciguatoxin	diarrhea, weakness of the legs, fatigue, cold water produces a painful stinging sensation, itching; the illness closely resembles ciguatera fish poisoning
Callista shellfish (*Callista brevisiphonata*); Japan	a histamine-like substance, choline	flushing of the face, itching, urticaria (stinging sensation of skin), sensation of constriction of the chest, abdominal pain, nausea, respiratory distress, asthmatic attacks, paralysis, hypersalivation, numbness of the tongue, throat; recovery usually within 10 days
Japanese dosinea (*Dosinia japonica*); Japan	venerupin	nausea, vomiting, nervousness, bleeding from the mucous membranes of the nose, mouth, and gums, halitosis, hemorrhages into the skin, enlarged liver, delirium, coma, death
Japanese littleneck cockle (*Tapes semidecussata*); Japan	same as Japanese dosinea	
Arthropods—joint-legged animals: crabs, spiders, and others		
Shore crab (*Demania toxica*); Philippines	unknown	nausea, vomiting, diarrhea, muscular weakness, respiratory distress, difficulty in speaking, hypersalivation, muscular paralysis, convulsions, death
Crab (*Zozymus aeneus*); Indo-Pacific	similar to tetrodotoxin; toxicity of these crabs variable	tingling about the mouth, nausea, vomiting, muscular paralysis, coma, convulsions, death
Crab (*Eriphia sebana*); tropical Pacific Ocean	same as *Z. aeneus*	
Asiatic horseshoe crab (*Tachypleus tridentatus*); Southeast Asia	unknown	dizziness, headache, nausea, vomiting, abdominal pain, cardiac palpitation, numbness of the lips, weakness, muscular paralysis, hypersalivation, loss of consciousness, death
Sharks, eels, and other fish		
Greenland shark (*Somniosus microcephalus*); Arctic	unknown; flesh toxic; liver of tropical sharks very toxic and may also cause death	nausea, vomiting, diarrhea, abdominal pain, tingling and burning sensation of the tongue, throat, and esophagus, muscular cramps, respiratory distress, coma, death
Moray eel (*Gymnothorax javanicus*); Indo-Pacific†	ciguatoxin*	symptoms may develop rapidly or slowly; tingling about the lips, tongue, and throat, followed by numbness, nausea, vomiting, abdominal cramps, muscular weakness, paralysis, convulsions, teeth feel loose, visual impairment, skin rash, hot objects feel cold and vice versa ("Dry Ice" or "electric shock" sensation); loss of muscular coordination, coma, death in about 12 percent of the cases; known as ciguatera fish poisoning, this is one of the most common forms of fish poisoning
Blue crevally (*Caranx melampygus*); Indo-Pacific†	same as Moray eel	
Red snapper (*Lutjanus bohar*); Indo-Pacific†	same as Moray eel	
Spotted grouper (*Cephalopholis argus*); Indo-Pacific†	same as Moray eel	

*Fish poisoning is categorically referred to as "ichthyosarcotoxism," but there are several different forms of fish poisoning such as ciguatera fish poisoning, clupeotoxism, scombrotoxism, and others. †More than 400 species of tropical reef fishes have been involved in ciguatera fish poisoning. These fish are normally edible, but under certain conditions may become toxic.

produced by only a small number of plants. Ricin, a toxalbumin from the castor bean (*Ricinus communis*), is one of the most toxic substances known.

Under certain ecological conditions plants may become poisonous as a result of the accumulation of toxic inorganic minerals such as copper, lead, cadmium, fluorine, manganese, nitrates, or selenium. Photosensitization, an unusual toxic reaction resulting from the ingestion of certain plants, may be of two types. The toxic substance may be obtained directly from the plant, which thereupon acts on the skin (primary photosensitivity), or the toxicity may result from liver damage caused by the metabolism of a toxic plant and failure of the breakdown products to be eliminated by the liver (hepatic photosensitivity). In either case the animal reacts by becoming restless; in addition, the skin reddens, and a severe sloughing of the skin develops. Death seldom occurs.

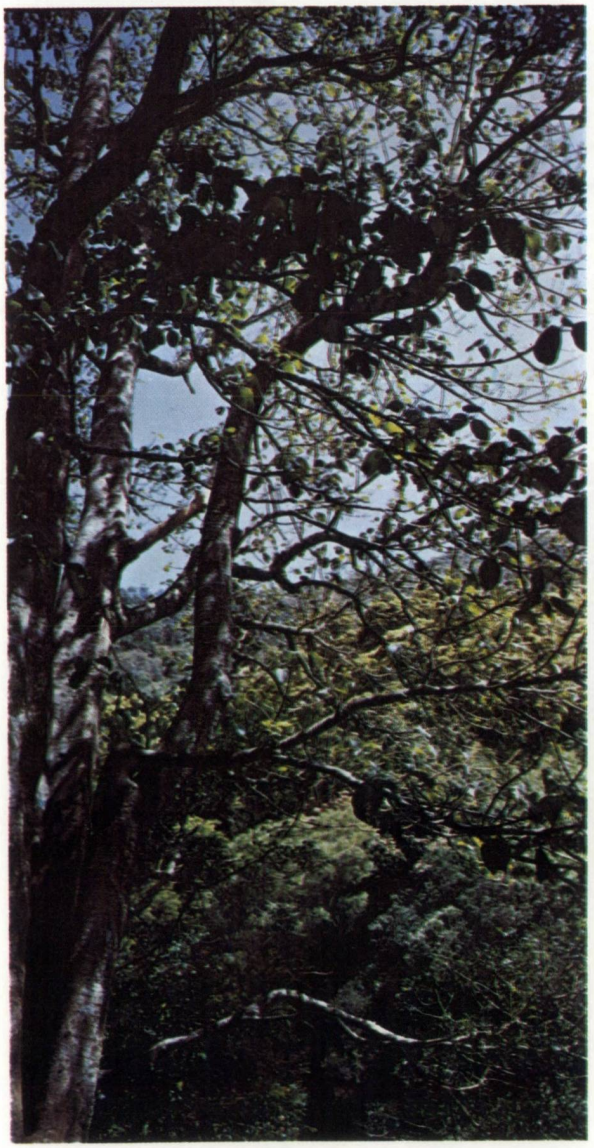
A large number of poisonous plants occur throughout the world; a few representative species and their poisons are listed in Table 5.



Queen hornet *(Vespa crabro)*.



Black widow *(Latrodectus mactans)*.





Poison hemlock (*Conium maculatum*).

Amanita pantherina.



Laportea gigas, (left) entire tree, (below) underside of a leaf showing stinging hairs.



Plate 1: (Top left) S.C. Bisserot—Bruce Coleman Inc., (top right) J.A.L. Cooke, (bottom left, bottom right) K. Gillett, World Life Research Institute, (centre left) Kitty Kahout from Root Resources—EB Inc., (centre right) L. Hugh Newman from The Natural History Photographic Agency—EB Inc.



Castor bean *(Ricinus communis)* with details of (left) seeds and (right) fruit.



Rosary pea *(Abrus precatorius)* with enlarged view of the poisonous seeds.



Larkspur *(Delphinium anthiscifolium)* with details of flower and fruit.

Monkshood *(Aconitum japonicum)* with details of tuberous root and flower.

Deadly nightshade *(Solanum nigrum)* with enlarged views of flower and fruit.



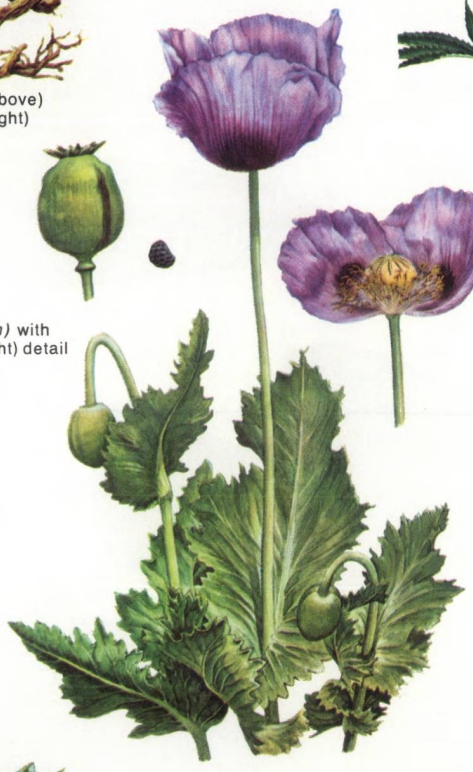




Belladonna *(Atropa belladonna)* showing (above) details of the flower, (below) the fruit, and (right) the root.



Marihuana *(Cannabis sativa)* with (left) male plant, (right) female plant



Opium poppy *(Papaver somniferum)* with (left) mature fruit and seed and (right) detail of flower.

Jimson weed *(Datura stramonium)* with enlarged view of seeds.



Foxglove *(Digitalis purpurea)*.



Plate 2: B.W. Halstead, World Life Research Institute; artwork, (top right) Luon Kwan-Ming Li, (others) J. Fujishima

Plate 3: B.W. Halstead, World Life Research Institute; artwork, (top right) H. Matsuura, (others) J. Fujishima



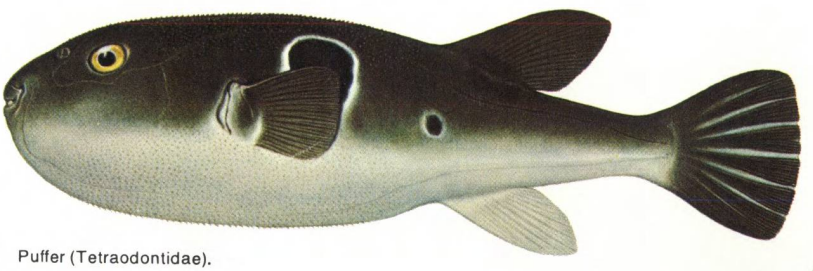
Crab *(Zozymus aeneus).*



Lionfish *(Pterois volitans).*



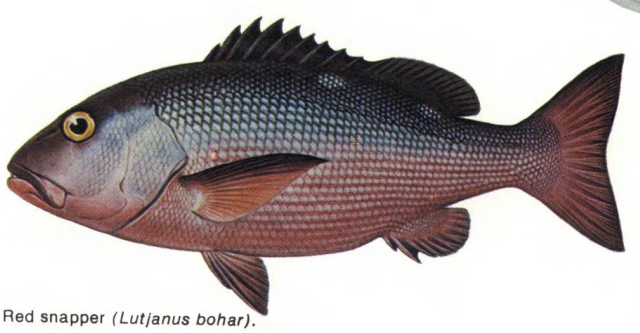
Moray eel *(Gymnothorax javanicus).*



Puffer (Tetraodontidae).

Blue crevally *(Caranx melampygus).*





Red snapper *(Lutjanus bohar).*

Stonefish *(Synanceja horrida).*

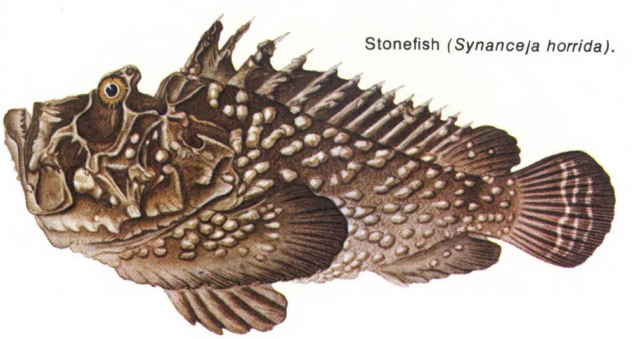
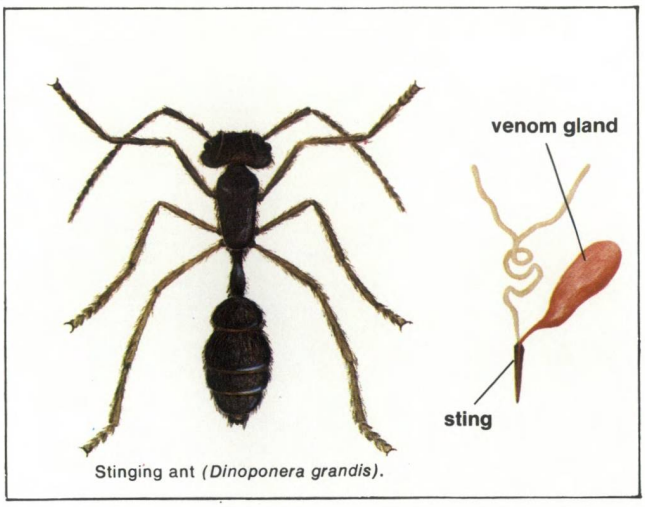
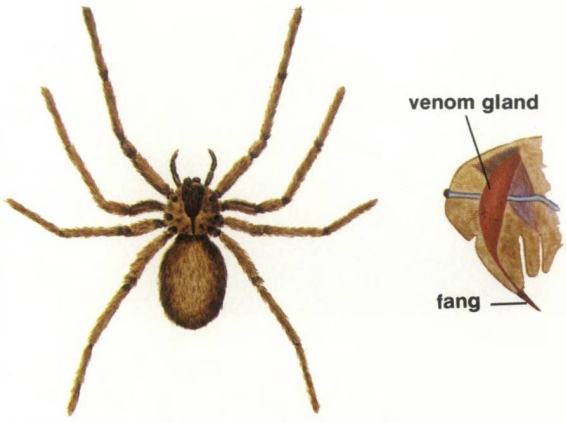


Plate 4: B.W. Halstead—World Life Research Institute; artwork, (centre right above) T. Kumada, (centre right below) K. Tomita, (bottom right) S. Arita, (others) M. Shirao

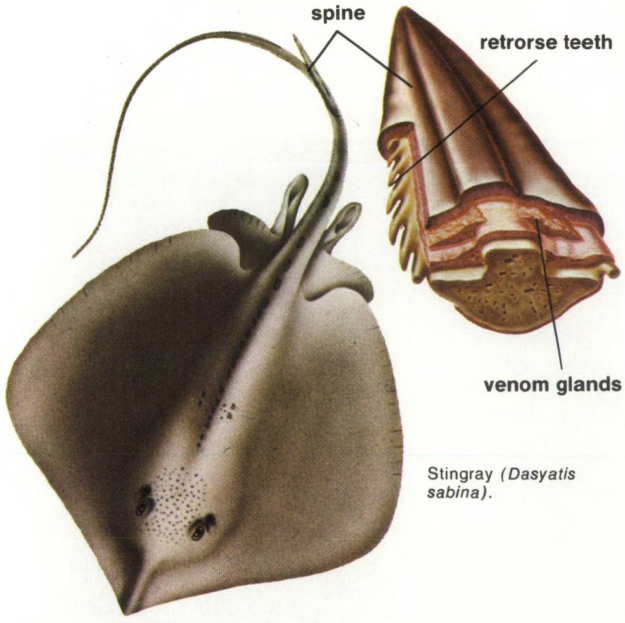
Plate 5: B.W. Halstead—World Life Research Institute; artwork, (centre left) L. Barlow, (bottom right) L. Barlow and M. Shirao, (others) K. Fogassy



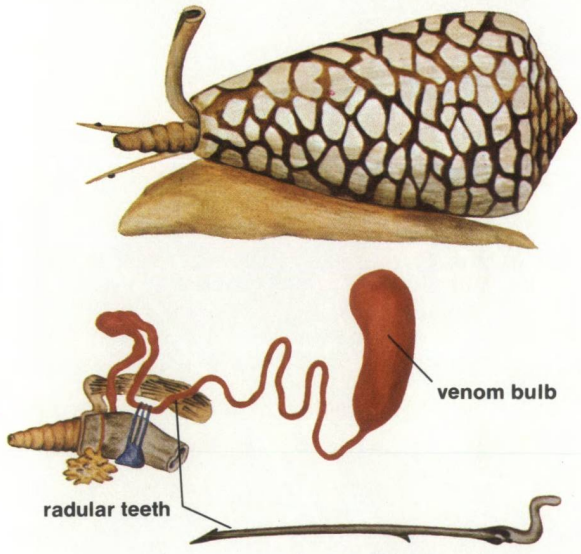
Stinging ant *(Dinoponera grandis)*.



Brown recluse spider *(Loxosceles reclusa)*.



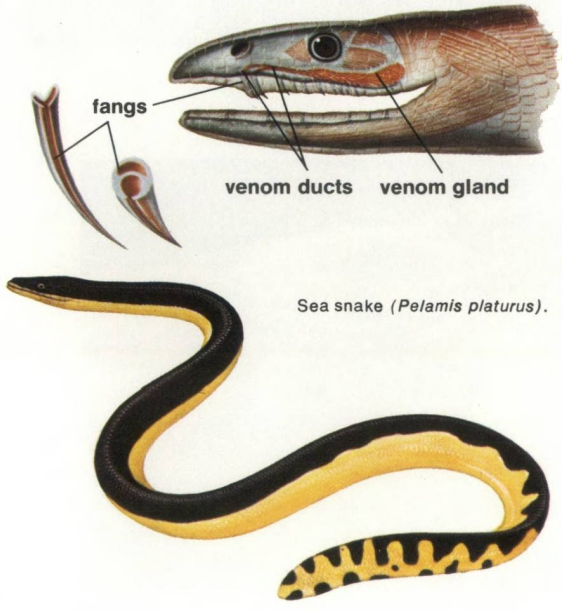
Stingray *(Dasyatis sabina)*.



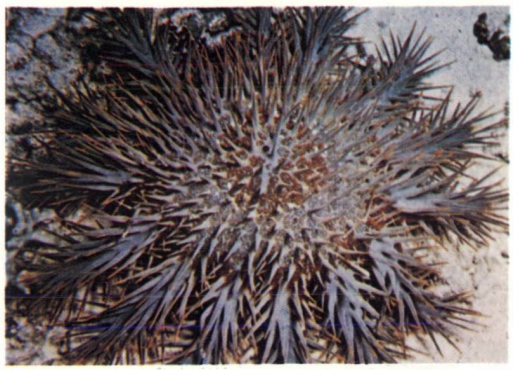
Marbled cone snail *(Conus marmoreus)*.



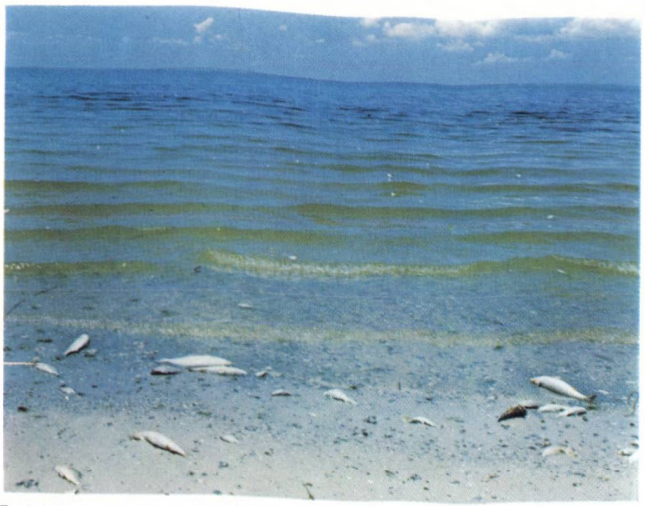
Blue-spotted octopus *(Octopus maculosus)*.



Sea snake *(Pelamis platurus)*.

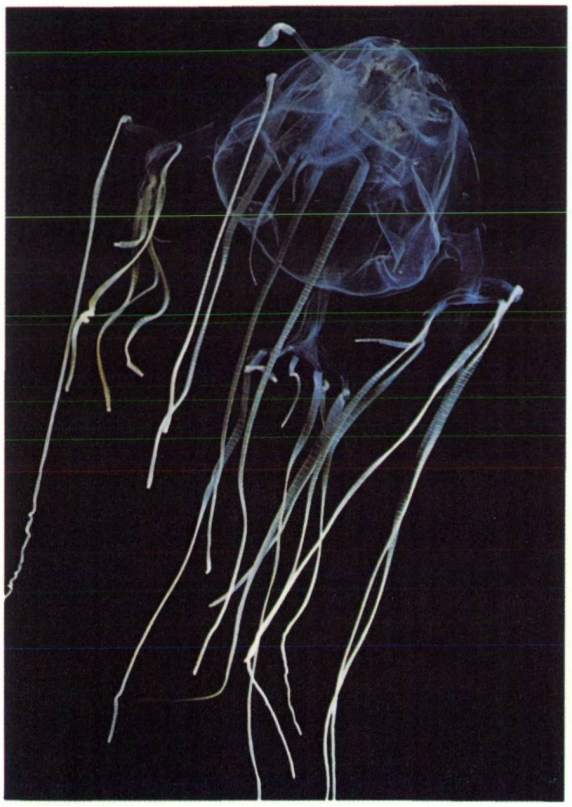


Crown of thorns starfish *(Acanthaster planci)*.

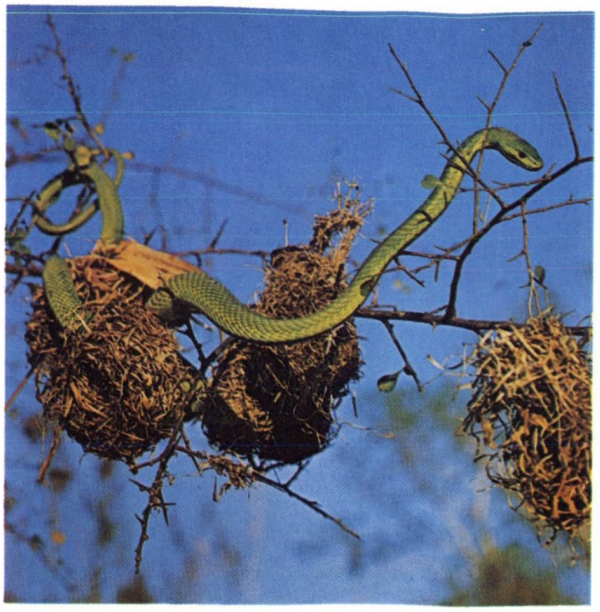


Red tide, Tampa Bay, Florida, showing fish kill and red coloration caused by dinoflagellates.

Sea wasp *(Chironex fleckeri)*.



Green mamba *(Dendroaspis viridis)*.



Kokoa frog or South American arrow poison frog (*Dendrobates* species).



Gila monster *(Heloderma suspectum)*.

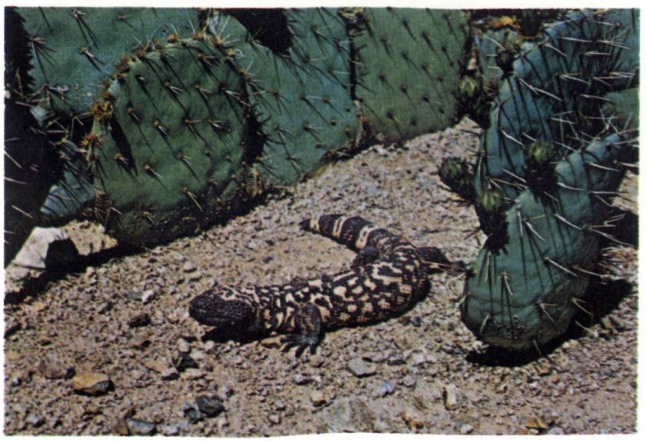


Plate 6: (Top left, centre left) World Life Research Institute, (top left) K. Gillett, (top right, bottom left) National Audubon Society, (top right) R.F. Head, (bottom left) George Porter, (center right) N. Myers—Bruce Coleman Inc., (bottom right) Richard Weymouth Brooks—Photo Researchers

Table 6: Representative Animals Poisonous When Eaten (continued)

name and distribution	toxic principle	toxic effects and comments
Yellowfin grouper (*Mycteroperca venenosa*); west tropical Atlantic†	same as Moray eel*	
Emperor sea bass (*Variola louti*); Indo-Pacific†	same as Moray eel	
Great barracuda (*Sphyraena barracuda*); tropical seas†	same as Moray eel	
Triggerfish (*Balistes vetula*); tropical Atlantic†	same as Moray eel	
Trunkfish (*Lactoria cornutus*); Indo-Pacific†	same as Moray eel	
Thread herring (*Clupanodon thrissa*); Indo-Pacific	clupeotoxin, chemical nature unknown	metallic taste, nausea, vomiting, abdominal pain, vascular collapse, hypersalivation, numbness, muscular paralysis, convulsions, coma, death; this form of poisoning develops rapidly and violently; mortality rate is about 50 percent and death may come within a few minutes; this form of fish poisoning is known as clupeotoxism; no known antidote
Atlantic thread herring, or yellow-billed sprat (*Opisthonema oglinum*); West Indies	same as thread herring	
Red-ear sardine, also called yellow-billed sprat (*Harengula humeralis*); West Indies	same as thread herring	
Castor-oil fish (*Ruvettus pretiosus*); tropical Atlantic, Indo-Pacific	oleic acid	produces a painless diarrhea; poisoning known as gempylotoxism; no treatment needed
Skipjack tuna (*Euthynnus pelamis*); tropical seas Bluefin tuna (*Thunnus thynnus*); subtropical and temperate seas	saurine, a histamine-like substance; scombroid fishes (mackerals, tunas, swordfishes, and allies), contain a chemical constituent in their flesh called histidine; when histidine is acted upon, it forms a histamine-like substance called saurine; this occurs when the fishes are permitted to stand at room temperature for several hours; scombroid fishes are more susceptible to the development of saurine poison than most other kinds of fishes	the symptoms of acute scombroid poisoning resemble those of a severe allergy: sharp, peppery taste, headache, throbbing of the large blood vessels of the neck, nausea, vomiting, massive red welts, and intense itching; recovery after 8–12 hours; this is probably the most common and cosmopolitan form of fish poisoning; antihistamines are used for treatment
Deadly death puffer (*Arothron hispidus*); tropical Pacific, Indian Ocean, Red Sea	tetrodotoxin ($C_{11}H_{17}O_8N_3 \pm y\ H_2O$)	tingling of lips and tongue, loss of motor coordination, floating sensation, hypersalivation, numbness of the entire body, muscular paralysis, difficulty in swallowing, weakness, nausea, vomiting, convulsions, about 60 percent fatality in humans; no known antidote
White-spotted puffer (*Arothron meleagris*); tropical Pacific	same as deadly death puffer	
Amphibians		
California newt (*Taricha torosa*); California	tarichatoxin; poison said to be identical to tetrodotoxin; the eggs of this newt are extremely toxic	effects in man are unknown; no known antidote
Reptiles		
Hawksbill turtle (*Eretmochelys imbricata*); tropical seas Leatherback turtle (*Dermochelys coriacea*); temperate and tropical seas	chelonitoxin, chemistry unknown; the flesh of some species of marine turtles is extremely poisonous	nausea, vomiting, diarrhea, burning sensation of lips, tongue, mouth, tightness of the chest, difficulty in swallowing, hypersalivation, foul breath, skin rash, sloughing of the skin, enlargement of the liver, coma, death; fatality rate is high; no known antidote
Mammals		
Sei whale (*Balaenoptera borealis*); North Pacific and North Atlantic oceans	unknown; livers of many marine mammals are toxic	intense headache, neck pain, photophobia, desquamation (peeling in scales) around the mouth and face, flushing of the face; antihistamines are used in treatment
White whale (*Delphinapterus*); Arctic seas	unknown	flesh is poisonous and has caused fatalities in humans; no known antidote
Polar bear (*Thalarctos maritimus*); Arctic	vitamin A, and possibly other toxic substances	intense throbbing headaches, nausea, vomiting, diarrhea, abdominal pain, dizziness, drowsiness, irritability, muscle cramps, visual disturbances, collapse, coma, rarely death

*See footnote for *ciquatoxin*. †See footnote for *Moray eel*.

Animal poisons (zootoxins). Poisonous animals are widely distributed throughout the approximately 1,000,000 species of animals, ranging from protozoans to polar bears; the only major group that seems to be exempt is the birds. The percentage of organisms that produces toxic substances is not known, but a significant number are poisonous. There are probably many more poisonous animals inhabiting the marine environment than are found on land.

Zootoxins can be divided into several categories: (1) oral poisons—those that are poisonous when eaten; (2) parenteral poisons, or venoms—those that are produced by a specialized poison gland and administered by means of a venom apparatus; and (3) crinotoxins—those that are produced by a specialized poison gland but are merely released into the environment, usually by means of a pore. Categories of zootoxins

Oral zootoxins (see Table 6) are generally thought to be small molecules; most venoms (Table 7) are believed to be large molecules, usually a protein or a substance in

Table 7: Representative Venomous Animals That Inflict a Sting

name and distribution	toxic principle	toxic effects and comments
Cnidarians*		
Fire coral (*Millepora* species); tropical seas	unknown	skin irritation, mild prickly or stinging sensation similar to a nettle (plant) sting
Portuguese man-of-war (*Physalia* species); tropical seas	tetramine, 5-hydroxytryptamine	immediate, intense stinging, throbbing, or burning sensation, shooting sensation, inflammatory rash, blistering of the skin, shock, collapse, in very rare cases death
Sea wasp (*Chironex fleckeri*); northern and northeast Australia	cardiotoxin	immediate extreme painful stinging sensation, seared reddened lines wherever the tentacles touch the skin, large indurated wheallike lesions, prostration, dizziness, circulatory failure, respiratory distress, rapid death in a high percentage of cases; no specific antidote available
Sea anemone (*Actinia equina*); Mediterranean, Black Sea, etc.	nature of venom unknown	burning, stinging sensation, itching, swelling, redness, ulceration, nausea, vomiting, prostration; no specific antidote available
Mollusks		
Cone shell (*Conus* species); tropical Indo-Pacific region	quarternary ammonium compounds and others	blanching at the site of injection, cyanosis of the surrounding area, numbness, stinging or burning sensation, blurring of vision, loss of speech, difficulty in swallowing, nausea, extreme weakness, coma, and death in some cases; no specific antidote
Spotted octopus (*Octopus maculosus*); Indo-Pacific, Indian Ocean	cephalotoxin, a neuromuscular poison	sharp burning pain, similar to a bee sting, numbness of the mouth and tongue, blurring of vision, loss of tactile sensation, difficulty in speech and swallowing, paralysis of legs, nausea, prostration, coma, and death in a high percentage of cases
Insects		
Kissing bug (*Triatoma* species); Latin America, United States	unknown	bite usually painless; later itching, edema about the bite, nausea, palpitation, redness; the bite is of relatively minor importance but spreads Chagas disease caused by a trypanosome (protozoan)
Hawkmoth (*Acanthosphinx guessfelotii*)	unknown	tibial spurs are capable of inflicting a painful scratch; a poison may be present
Moth (*Anaphe* species); Europe	unknown	female moth possesses hollow barbed setae or hairs in the anal tuft which are released at time of egg laying; during the breeding season they become so numerous they are released in the atmosphere and produce a severe dermatitis; the hairs contain a poison
Puss caterpillar (*Megalopyge* species); United States, Latin America	unknown	stinging hairs of the caterpillar associated with poison-secreting glands; contact with the hairs produces an intense burning pain, itching, pustules, redness, nausea, fever, numbness, swelling, and paralysis; recovery usually within about six days
Honeybee (*Apis* species); worldwide	neurotoxin, hemolytic, melittin, hyaluronidase, phospholipase A, histamine, and others	sting produces acute local pain or burning sensation, blanching at the site of the sting surrounded by zone of redness, and itching; local symptoms usually disappear after 24 hours; severe cases may develop massive swelling, redness, shock, prostration, vomiting, rapid heart beat, respiratory distress, trembling, coma, and death; it is estimated that 500 stings delivered in a short period of time can provide a lethal dose to a human; bees kill more people in the United States than do venomous reptiles
Bumblebee (*Bombus* species); temperate regions	similar to honeybee (*Apis*) venom	stings are similar to honeybee (*Apis*) stings; bumblebees are not as vicious as honeybees
Carpenter bee (*Xylocopa* species); temperate regions	similar to honeybee venom	similar to honeybee stings
Wasp (*Seliphron* species); temperate regions	similar to honeybee venom	similar to honeybee stings; stings may be fatal
Yellow jacket, hornet (*Vespula* species); temperate regions	similar to bee venom; also acetylcholine	they can both bite and sting; the sting is similar to that of the honeybee but more painful; yellow jackets are quite aggressive; stings may be fatal
Wasp (*Polistes* species and *Vespa* species); worldwide	similar to bee venom; also acetylcholine	wasps are less aggressive than hornets and their sting is similar to the honeybee but generally less painful than hornets; stings may be fatal
Velvet ant (*Dasymutilla* species); United States	similar to wasp venom	velvet ants are actually female wingless wasps; stings are similar to other wasp stings
Harvester ant (*Pogonomyrmex* species); United States	bradykinin, formic acid, hyaluronidase, hemolytic, phospholipase A, and others.	ant stings cause immediate intense burning, pain, blanched area at site of sting surrounded by redness, ulceration, fever, blistering, itching, hemorrhaging into the skin, eczematoid dermatitis, pustules, respiratory distress, prostration, coma, and death in some instances
Fire ant (*Solenopsis* species); United States, Latin America	similar to harvester-ant venom	similar to above; stings are very painful, burning sensation, etc.
Ponerine ant (*Paraponera clavata*); Latin America	same as fire ant	
Spiders		
Brown spider (*Loxosceles* species); United States, South America, Europe, Asia	cytotoxic, hyaluronidase, hemolytic, and others	bite causes stinging sensation or burning pain, blanching at site of bite surrounded by redness, blistering, hemorrhages into the skin and internal organs, ulceration, vomiting, fever, cardiovascular collapse, convulsions, sometimes death
Black widow (*Latrodectus* species); tropical and temperate regions	neurotoxic	bite may or may not be painful, two tiny red dots at site of bite, localized swelling after a few minutes; intense cramping pain of abdomen, legs, chest, back; pain may become agonizing causing patient to scream; rigidity of muscles, lasting 12–48 hours, nausea, sweating, respiratory distress, priapism (abnormal, painful erection of the penis) in males, chills, skin rash, restlessness, fever, numbness, tingling, about 4 percent are fatal; antiserum is available
Tarantula (*Dugesiella* and *Lycosa* species); temperate and tropical regions	venom varies, usually mild	most of the large tarantulas found in the United States, Mexico, and Central America are harmless to man; their bites are similar to a pin prick and relatively mild; their danger to man has been grossly exaggerated; some of the large tropical species may be more poisonous, but their effects are largely localized
Scorpions†		
Scorpion (species of *Centruroides*, *Tityus*, and *Leiurus*); warm temperate and tropical regions	neurotoxin, cardiotoxin, hemolytic, lecithinase, hyaluronidase, and others	symptoms vary depending upon the species of scorpion; sting from the tail stinger causes a sharp burning sensation, swelling, sweating, restlessness, salivation, confusion, vomiting, abdominal pain, chest pain, numbness, muscular twitching, respiratory distress, convulsions, death; the mortality rate from stings from certain species of scorpions is very high; antiserum is available

*About 9,000 species of jellyfish, sea anemones, etc.; stings are produced by means of nematocysts, specialized stinging cells. †About 650 species.

Table 7: Representative Venomous Animals That Inflict a Sting (continued)

name and distribution	toxic principle	toxic effects and comments
Echinoderms‡		
Crown-of-thorns starfish (*Acanthaster planci*); Indo-Pacific	nature of poison unknown	penetration of the spines produces a painful wound, redness, swelling, vomiting, numbness, and paralysis
Long-spined sea urchin (*Diadema setosum*); Indo-Pacific	nature of poison unknown	penetration of the spines produces an immediate and intense burning sensation, redness, swelling, numbness, muscular paralysis
Sea urchin (*Toxopneustes pileolus*); Indo-Pacific	nature of poison unknown	bites from the stinging jaws or pedicellariae (small pincer-like organs) produce an immediate intense radiating pain, faintness, numbness, muscular paralysis, respiratory distress, and occasionally death
Sharks and rays		
Stingray (*Dasyatis* species); warm temperate and tropical seas	stingray venom, cardiotoxin, chemistry unknown	penetration of the tail spines inflicts jagged painful wounds that produce sharp, shooting, throbbing pain, fall in blood pressure, nausea, vomiting, cardiac failure, muscular paralysis, rarely death; no known antidote; stingrays are among the most common causes of envenomations in the marine environment
Bony fishes		
Weever fish (*Trachinus draco*); Mediterranean Sea	weeverfish venom, chemistry unknown	opercular and dorsal fin spines can produce instant pain, burning, stabbing or crushing sensation; pain spreads and becomes progressively more intense causing the victim to scream with anguish, suffer loss of consciousness, numbness about the wound, swelling, redness, nausea, delirium, difficulty in breathing, convulsions, and death; no known antidote
Scorpion fish (*Scorpaena* species); temperate and tropical seas	scorpionfish venom, chemistry unknown	fin spines can inflict painful stings and intense immediate pain that may cause victim to scream followed by redness, swelling, loss of consciousness, ulceration of the wound, paralysis, cardiac failure, delirium, convulsions, nausea, prostration, and respiratory distress, but rarely death; no known antidote
Stonefish (*Synanceja* species); Indo-Pacific region	stonefish venom, chemistry unknown	produces an extremely painful sting by means of the dorsal fin spines; symptoms similar to other scorpionfish stings but more serious
Reptiles		
Gila monster (*Heloderma suspectum*); southwestern United States	heloderma venom, primarily a neurotoxin	all of the teeth are venomous; bite causes local pain, swelling, weakness, ringing of the ears, nausea, respiratory distress, cardiac failure, may cause death; no antiserum available
Rear-fanged snakes		
Boomslang (*Dispholidus typus*); Africa	coagulates the blood; proteolytic and other enzymes are present	local pain, swelling, hemorrhages, bleeding from nose, mouth, and skin, headache, vomiting, collapse, death; antiserum is available
Front-fixed-fanged snakes		
Death adder (*Acanthophis antarcticus*); Australia, New Guinea	coagulates blood; proteolytic and other enzymes present	effects of envenomation develop slowly (15–60 minutes); nausea, vomiting, faintness, drowsiness, staggering, slurred speech, respiratory distress, hemorrhages, death; antiserum available
Indian krait (*Bungarus candidus caeruleus*); India, Burma, Malaya, Indonesia	cholinesterase, protease, and other enzymes present	little pain or local reaction, latent period up to 12 hours; abdominal pain, staggering gait, difficulty in swallowing, stiffness of jaws, coma, respiratory distress, cardiac failure, death; antiserum available
Brown snake (*Demansia textilis*); Australia	cholinesterase, ophio-adenosinetriphosphatase, phospholipase, and other enzymes	latent period up to 12 hours; abdominal pain, vomiting, headache, dizziness, blood in the urine, weakness, respiratory and circulatory collapse, death; antiserum available; this snake is responsible for most of the deaths due to snake bite in Australia; very dangerous
Black mamba (*Dendroaspis polylepis*); Ethiopia, Somalia, Natal, southwest Africa	cholinesterase, L-amino-acid oxidase, ophio-adenosine, triphosphatase, and other substances	local pain, swelling, salivation, paralysis of vocal cords, sweating, vomiting, restlessness, drowsiness, collapse, coma, respiratory distress, death; one of the most dangerous of all living snakes; antiserum available; a large black mamba secretes enough venom to kill ten men
Ringhals (*Hemachatus hemachatus*); South Africa	coagulates blood; cholinesterase, L-amino-acid oxidase, 5-nucleotidase, lecithinase, protease, and other enzymes	pain, weakness, respiratory distress, may cause blindness, cyanosis, collapse, death; venom sprayed at eyes; this is one of the spitting cobras; antiserum available
Coral snake (*Micrurus corallinus*); subtropical South America	cholinesterase, hyaluronidase, L-amino-acid oxidase, and other enzymes	numbness without pain at bite, headache, swelling of face and lips, hyperesthesia (unusual sensitivity of skin), sore throat, drooping of eyelids, photophobia, vomiting, rapid heart rate, backache, irritability, salivation, death; there are about 40 species in the genus *Micrurus;* all are dangerous; antiserum available for some species but not others
Indian cobra (*Naja naja*); southern Asia, Indonesia, Taiwan, Philippines	carboxypolypeptidase, cholinesterase, diastase, dipeptidase, polypeptidase, protease, and other enzymes; one of the most complex snake venoms	pain radiating from site of bite, edema, numbness, drooping of the eyelids, head, salivation, difficulty in speech, muscular incoordination, weakness, respiratory distress, blindness, incontinence, convulsions, death; antiserum available
Black-necked cobra (*Naja nigricollis*); Africa	cholinesterase, L-amino-acid oxidase, phospholipase, protease, and other enzymes	similar to poisonings by the Indian cobra; venom sometimes sprayed at eyes; produces an intense irritation but permanent damage to eyes is rare; antiserum available
Tiger snake (*Notechis scutatus*); Australia	coagulates blood; cholinesterase, hyaluronidase, ophio-adenosine-triphosphatase, phosphatase, and other enzymes	latent period of 15–60 minutes; nausea, vomiting, faintness, sweating, drowsiness, staggering, slurred speech, difficulty in swallowing, drooping of eyelids, respiratory distress, death; antiserum available
King cobra (*Ophiophagus hannah*); India, Burma, Philippines, Thailand	cholinesterase, L-amino-acid oxidase, ophio-adenosinetriphosphotase	symptoms similar to those produced by other cobras; symptoms develop rapidly, death often occurs in 30–60 minutes; one of the largest of venomous snakes—obtains a length of 16 feet; antiserum available
Taipan (*Oxyuranus scutellatus*); Australia, New Guinea	lecithinase, phospholipase	symptoms similar to those produced by the tiger snake; its long fangs and potent venom make this one of the world's most dangerous snakes; produces flaccid paralysis, respiratory paralysis, often rapidly fatal; antiserum available
Australian black snake (*Pseudechis porphyriacus*); Australia	L-amino-acid oxidase, lecithinase, phospholipase	local pain and swelling, vomiting, hemorrhages from nose and mouth, prostration, hematuria (presence of blood in the urine), and sometimes death; usually the bites are nonfatal; antiserum available
Sea snakes		
Beaked sea snake (*Enhydrina schistosa*); Indo-Pacific; all sea snakes are venomous, but only a few species have caused death in man	lecithinase	no local reaction; latent period minutes to hours; giddiness, muscle ache, muscular weakness, drooping of the eyelids, lockjaw, respiratory failure, acute renal (kidney) failure, death

‡Sea urchins, starfish, etc., about 4,000 species.

Table 7: Representative Venomous Animals That Inflict a Sting (continued)

name and distribution	toxic principle	toxic effects and comments
Erectile-fanged snakes		
Copperhead (*Ancistrodon contortrix*); eastern and southern United States	cholinesterase, L-amino-acid oxidase, phosphodiesterase, and other enzymes	local pain, swelling, necrosis, nausea, vomiting, shock, petechiae (minute bleeding spots), bloody stools, death; antiserum available
Eastern cottonmouth (*Ancistrodon piscivorus*); southeastern United States to central Texas	cephalinase, cholinesterase, hyaluronidase, lacithinase, and other enzymes	similar to poisoning by the copperhead but more severe; local necrosis more marked; death; antiserum available
Fer-de-lance (*Bothrops atrox*); central Mexico, south into South America	cholinesterase, deoxyribonuclease, L-amino-acid oxidase, protease, and other enzymes	local pain, bleeding from bite, gums, nose, mouth, rectum, blood fails to clot, shock, respiratory distress, hemorrhages into muscles and nervous system, death; antiserum available
Eastern diamondhead rattlesnake (*Crotalus adamanteus*); southeastern United States	cephalinase, cholinesterase, hyaluronidase, and other enzymes	local pain, edema, hemorrhaging into the tissues, dryness of the mouth, vomiting, shock, anemia, tingling sensation, blood in feces, difficulty in speech, yellow vision, loss of consciousness, death; antiserum available
Western diamondback rattlesnake (*Crotalus atrox*); southwestern United States	bradykininogen, cholinesterase, hyaluroindase, L-amino-acid oxidase, and other substances	symptoms similar to poisoning by *C. adamanteus*, but neurological symptoms less marked; antiserum available
Cascabel (*Crotalus durissus terrificus*); southern Mexico south to Argentina	cholinesterase, deoxyribonuclease, hyaluronidase, lecithinase, and other enzymes	symptoms similar to cobra bites, except that no edema occurs; antiserum available
Bushmaster (*Lachesis muta*); Costa Rica to northern South America	protease	very few bites recorded and little is known concerning the effects of its venom; severe prostration, rapid death; antiserum available
Taiwan habu (*Trimeresurus mucrosquamatus*); southeastern China, Taiwan	neurotoxin, amino-acid oxidase, hemorrhagin, and other substances	local pain, blistering, bleeding into the tissues, shock, death; antiserum available
True vipers		
Puff adder (*Bitis arietans*); Africa, Saudia Arabia	fibrinolytic, bradykininogen, cholinesterase, L-amino-acid oxidase, protease, and other substances	severe local edema, necrosis and sloughing of the tissues, restlessness, respiratory distress, gastrointestinal hemorrhages, death; antiserum available
Saw-scaled viper (*Echis carinatus*); India, Iraq, Saudia Arabia, Africa	carboxypolypeptidase, cholinesterase, dipeptidase, endopeptidase, and other enzymes	local pain and edema, hemorrhages into the skin, mucous membranes, and gastrointestinal tract, liver damage, shock, death; antiserum available
European viper (*Vipera berus*); Europe, Asia, Japan	bradykininogen, L-amino-acid oxidase	local pain and edema, hemorrhages along lymphatics, very little general reaction, vomiting, abdominal pain, shock, death; antiserum available
Russell's viper (*Vipera russelli*); Asia, Indonesia	carboxypolypeptidase, cholinesterase, dipeptidase, lipase, lecithinase, protease, and other enzymes	rapidly spreading edema with hemorrhages into the tissues, abdominal pain, vomiting, blood does not clot, loss of consciousness, circulatory failure, shock, death; antiserum available
Mammals		
Duck-billed platypus (*Ornithorhynchus anatinus*); Australia	protease and other enzymes	may inflict envenomation by means of the spurs on the hindlegs of the male platypus; immediate intense pain, swelling, redness in the area of the sting, faintness, numbness about the wound, restlessness, prostration; no death recorded; recovery is uneventful after a few days
Short-tailed shrew (*Blarina brevicauda*); North America	neurotoxic, proteolytic	the lower incisor teeth are associated with submaxillary poison glands; bites produce localized pain, redness in the area of the bite, shooting pains, and general discomfort of the affected part; gradual recovery after several days; no fatalities recorded
Solenodon (*Solenodon paradoxus*); Haiti	same as short-tailed shrew	
European water shrew (*Neomys fodiens*); Europe, Asia	same as short-tailed shrew	

close association with one. Venoms, which are produced by specialized poison glands, are injected by means of a mechanical device that is able to penetrate the flesh of the victim. Little is known about the biological or chemical properties of most crinotoxins (Table 8). The term poisonous may be used in the generic sense to refer to all three categories of zootoxins.

It is difficult to generalize when dealing with poisonous animals because they involve such a complex assemblage of unrelated organisms. Although there are poisonous animals in polar regions, by far the greatest number and variety inhabit warm temperate or tropical areas. One of the most dangerous of all toxic organisms probably is the sea wasp (*Chironex fleckeri*), which inhabits the northeast coastal waters of Australia. In less than one minute this jellyfish is able to inflict death in an adult human by means of its nematocyst, or stinging apparatus.

Some of the most complex relationships in biotoxicology are found in the marine environment. Certain marine biotoxins, such as ciguatera fish poison, apparently originate in marine plants, are ingested by herbivores and then passed on to carnivores and eventually to man. The extremely complex mechanism by which this is accomplished has eluded scientists for many years. With the build-up of toxic industrial chemical pollutants in the marine environment, the problems of toxicity in marine organisms are becoming increasingly more serious. There is evidence that under certain conditions chemical pollutants may trigger biotoxicity cycles in marine organisms. The outbreaks in Japan of Minamata disease were the result of such a cycle: micro-organisms, algae, shellfishes, and fishes ingested or absorbed industrial wastes with highly toxic organic compounds containing mercury and were in turn consumed by humans, causing a number of deaths among the population.

The relationships of representative poisonous animals and their position in the total framework of the animal kingdom can best be appreciated by categorizing them according to the group in which they belong. They are further grouped as to whether they are poisonous to eat, venomous, or crinotoxic in Tables 6, 7, and 8.

BIBLIOGRAPHY. S.J. AJL *et al.*, *Microbial Toxins*, 8 vol. (1970–72), the most comprehensive series of review articles ever assembled on the subject of bacterial protein toxins, bacterial endotoxins, and algal and fungal toxins; G. BRESADOLA, *Funghi mangerecci e velenosi*, 2 vol. (1932–33), one of the great classical works on poisonous mushrooms—although primarily European, most of the species discussed have a somewhat cosmopolitan distribution (magnificently illustrated); W. BUCHERL, E. BUCKLEY, and V. DEULOFEU, *Venomous Animals and Their Venoms*, vol. 1 (1968), and W. BUCHERL and E. BUCKLEY, vol. 2–3 (1971), a comprehensive series of review articles on venomous animals, their biology, effects of the venom, and treatment; UNITED STATES NAVY, BUREAU OF MEDICINE AND SURGERY, *Poisonous Snakes of the World: A Manual for Use by U.S. Amphibious Forces* (1968), a valuable, well-illustrated compendium on the poisonous snakes of the world, including their identification,

Table 8: Representative Crinotoxic Animals*

name and distribution	toxic principle	toxic effects and comments
Sponges		
Red moss (*Microciona prolifera*); eastern United States coastal waters	unknown	contact with the sponge produces a chemical irritation of the skin, redness, stiffness of the finger joints, swelling, blisters, and pustules
Flatworms		
Flatworm (*Leptoplana tremellaris*); European coastal waters	unknown	poison is produced by epidermal skin glands; no human intoxications recorded; extracts from the skin of these worms injected into laboratory animals produces cardiac arrest
Arthropods—joint-legged animals		
Blister beetles (*Cantharis vesicatorea*); United States	cantharidin	toxic substance does not seem to be produced by special glands but is found throughout the body of the beetle; no discomfort at time of initial contact; after about 8–10 hours large blisters develop on the skin accompanied by slight burning or tingling sensation; swallowing of the beetles may cause kidney damage; cantharidin is used as an aphrodisiac known as Spanish fly—a very dangerous substance to use; ingestion can cause severe gastroenteritis, kidney damage, blood in the urine, priapism, profound collapse, and death
Millipedes (species of *Orthoporus*, *Rhinocrichus*, *Julus*, and *Spirobolus*); temperate and tropical regions	unknown	repugnatorial (distasteful to enemies) fluid may be exuded or forcefully squirted from body pores a distance up to 30 inches or more; contact with the skin causes mild to moderately intense burning pain, redness, and pigmentation of the skin; toxic fluid squirted in the eyes may cause temporary blindness, an inflammatory reaction, and pain
Venomous ticks (species of *Ixodes* and *Ornithodoros*); temperate and tropical regions	unknown	tick bites result in swelling, redness, intense pain, headache, muscle cramps, loss of memory, etc.
Fishes		
Sea lamprey (*Petromyzon marinus*); Atlantic Ocean	unknown	slime of the lamprey is toxic, ingestion may cause diarrhea
Pacific sea bass (*Grammistes sexlineatus*); Indo-Pacific	antibiotic, neurotoxic	slime of fish is toxic; effects on man not known
Soapfish (*Rypticus saponaceus*); tropical and subtropical Atlantic	neurotoxic	slime of fish is toxic; produces an irritation of the mucous membrane
Amphibians†		
European earth salamander (*Salamandra maculosa*); Europe	skin glands of the salamander are poisonous; contains the alkaloids samandarine, samandenone, samandine, samanine, samandarone, samandaridine, and others	effects on humans not known; affects the heart and nervous system; causes in animals convulsions, cardiac irregularity, paralysis, and death
Toads (*Bufo* species); temperate and tropical regions	bufotoxin, bufogenins, and 5-hydroxytryplanime; poison includes a complex of many substances	produces a poisonous secretion in the parotid glands and skin; handling of some toads may cause a skin irritation; ingestion causes nausea, vomiting, numbness of the mouth and tongue, and tightness of the chest; the poison has a digitalis-like action
Frogs (some species of *Dendrobates*, *Physalaemus*, and *Rana*); northern South America and Central America	skin secretions are poisonous; histamine, bufotenine, physalaemin, serotonin, and other substances; composition varies with the species	skin secretions produce a burning sensation when handled; used by Indians as an arrow poison
Tree frogs (some species of *Hyla* and *Phyllobates*); northern South America, Central America	skin secretions are poisonous; batrachotoxin, steroidal alkaloids, serotonin, histamine, and other substances; bufotenine varies with the species	some frog species produce a burning sensation and a skin rash when handled; skin secretions in the eye may produce a severe inflammatory reaction; if ingested poison causes vomiting and abdominal pain; batrachotoxin is extremely toxic if injected; used by Indians as an arrow poison

*Animals in which poison glands are present and poison is released into environment through a pore. †Poisonous amphibians are sometimes referred to as "venomous," but they do not possess a true venom apparatus. They possess only poison glands.

distribution, habits, and first aid measures and treatment for their bites; C.A. FRAZIER, *Insect Allergy* (1969), one of the most complete works available concerning allergies and the poisonous effects of terrestrial arthropods; B.W. HALSTEAD, *Poisonous and Venomous Marine Animals of the World*, 3 vol. (1965–70), an exhaustive and profusely illustrated compendium on the toxic marine animals of the world from protozoans to polar bears, including their historical background, names, geographical distribution, biology, mechanism of intoxication, and medical, toxicological, pharmacological, and chemical aspects; R. HEIM, *Les Champignons toxiques et hallucinogènes* (1963), a valuable handbook on poisonous and hallucinogenic mushrooms, their biology, and poisonous effects; H.L. KEEGAN and W.V. MACFARLANE, *Venomous and Poisonous Animals and Noxious Plants of the Pacific Region* (1963), a useful compendium of technical data on poisonous organisms of the Pacific area, including their biology, poisons, and toxic effects; J.M. KINGSBURY, *Poisonous Plants of the United States and Canada* (1964), a valuable review of the literature on poisonous plants and the toxicology of plant poisoning in man and animals; L.M. KLAUBER, *Rattlesnakes: Their Habits, Life Histories, and Influence on Mankind*, 2 vol. (1956), the most comprehensive work ever written on North American rattlesnakes, containing valuable information on their identification, venom apparatus, snake bites, and treatment; M. PHISALIX-PICOT, *Animaux venimeux et venins*, 2 vol. (1922), one of the great classical works on poisonous animals and their toxins.

(B.W.H.)

Poisons and Poisoning

Poisons are substances that in small amounts are capable of producing serious injury or death. In truth, however, the poison is in the dose. It is a commonplace that drugs of unchallenged benefit in appropriate doses may become seriously poisonous when such doses are exceeded. It must not be forgotten, also, that variation in response of different individuals may be so great that an average dose can induce serious adverse reactions in the susceptible. Toxins are poisons of natural origin, whether bacterial, plant, or animal, and comprise some of the most lethal materials known; *e.g.*, aconite, botulinus toxin, and fish poisons.

The sciences of pharmacology and toxicology, the latter of which has poisons and poisoning in its province, share common principles, common procedures, and a common vocabulary; they have so much in common that university departments of pharmacology and toxicology grant advanced degrees in either or both disciplines, and pharmacology–toxicology research centres have been created under government sponsorship. The differences, practical and important, are based on points of view. Toxicology may be defined as the science of the injurious effects of chemicals on living systems or simply as the science of poisons; pharmacology, as the science of the useful, advantageous effects of drugs and chemicals on living sys-

tems, especially the effects related to disease, whether preventive, curative, or diagnostic.

Scope of the article. This review of poisons and poisoning will (1) define the field of toxicology and the mission of the toxicologist; (2) describe some of the problems attendant on the recognition, treatment, and prevention of poisonings; and (3) consider in general terms the fate of poisons introduced into the body and the sites and mechanisms of toxic actions. Material relevant to this subject may also be found in two related articles, POISONOUS ANIMALS AND PLANTS; DRUG AND DRUG ACTION.

Background. In the adaptive struggle for survival, members of every kingdom—unicelluiar forms, fungi, plants, and animals alike—evolved chemical agents to discourage competition, to defend themselves, and to secure prey. Primitive chemical-warfare agents are well known to man; for example, the antibiotics produced by bacteria and fungi, the growth inhibitors that suppress competitors of higher plants, and the venoms of bacteria, parasites, insects, and predatory animals. Some of these chemicals are synthesized by the organism—*e.g.*, the nicotine formed in the tobacco root—and some are obtained from the food—*e.g.*, the toxic cardiac glycosides (carbohydrates that affect the heart) in the monarch butterfly that come from the milkweed diet of its caterpillar; these are apparently the reason that birds shun the monarch. Presumably, curiosity led prehistoric man to touch, to taste, and to rub on his skin everything that came to hand and to discover the irritating, emetic, and purging properties of certain earths, molds, plants, and animal tissues. Witch doctors learned to practice these effects on their fellows. Through hallucinations, they established rituals and ordeals; through illness or death, they punished enemies or tribal wrongdoers; they tipped arrows with malignant extracts. A survival of this practice is our word "toxic," from the Greek words *toxon* ("bow") and *toxikon* (*pharmikon*), "arrow (poison)." Poisons, like infectious diseases, may inflict their injuries subtly. It is easy to imagine that, long ago, witch doctors also employed toxins against disease, an image repeated by Shakespeare, "In poison there is physic." Into medieval times, medical remedies included many of these physiologically potent substances, and some remain today as irreplaceable. The extent of the fear and hatred of poisons can be judged from the pronouncement in 1450 by James II forbidding all persons to bring home poisons "Under pain of treason." The persistence into modern times of the king's taster attests to the grim success of poison as a political weapon.

Natural poisons

Organisms also developed defenses against the poisons of everyday life. Painful experience (man and poison ivy) teaches individuals, and instinct (birds and monarch butterflies, horses and rattlesnakes) guides species to avoid exposures. Rejection mechanisms (regurgitation, purgation) expel noxious matter. Inside the cells, gene-controlled, special metabolic processes detoxify life-endangering molecules.

Modern man has borrowed practices and, in some instances, chemical substances from other species. In addition, man has developed a broad array of specialized chemicals to protect himself, to enhance his health, and to reduce competition for food or for living space. The remedies made use of by today's physician include many substances with astonishing healing powers in therapeutic doses but also capable of devastating injury. Highly-irritant substances disperse mobs. Agriculture uses millions of pounds annually of compounds selectively toxic for pests of many sorts, from bacteria, molds, and weeds to insects and coyotes.

THE FIELD OF TOXICOLOGY

The field of toxicology comprises many areas of service and research; selected examples are listed in the accompanying chart. Even casual inspection of the list invites the conclusion that toxicology is destined to play an increasingly prominent role in man's future development. The toxicological activities under "medical," "environmental," and "industrial" in Table 1 lie mainly in the field of preventive medicine, in which the mission of the toxicologist is to prevent injury and to insure safety. The sum total of toxicological problems is staggering. The advent in the past score of years of waves of new, frequently highly specific, and highly potent drugs has produced a flood of physician-caused illness. From other sources, the number of poisonings annually, including ingestions however slight, is estimated at 1,600,000 in the United States; of these perhaps one-fifth to one-half are associated with obvious acute illnesses; about 3,000 are fatal. The death toll from alcohol-associated motor accidents adds 25,000 or more deaths annually. Threats to the quality of the environment have led to serious exploration of the enormous field of ecologic toxicology. Among industrial employees, certain groups, such as farm workers, thousands of whom are repeatedly poisoned, are not yet covered by the widely enforced and increasingly effective occupational health measures. During the past 75 years, for example, illnesses among those handling toxic metals in Great Britain have decreased from over 1,000 annually, with nearly 40 deaths, to less than 60 illnesses, with no deaths. The toxicological activities listed under "pesticidal" and "military," in contrast, seek more effective poisons. Ideally, pesticides selectively control pests without injury to man or to other nontarget organisms, and military agents incapacitate without causing death or lasting injury. In either field, adequate defenses must protect the users.

The importance of toxicology

Table 1: Service Areas of the Toxicologist

medical	environmental	industrial	pesticidal	military
Toxicologic evaluation	air and water pollution	occupational health	insecticides	agents
Adverse drug reactions	radiation toxicology	industrial hygiene	rodenticides	lacrimators
Treatment of acute poisonings	space toxicology	metals	herbicides	nerve gases
Forensic toxicology	ecologic toxicology	solvents	fungicides	vesicants
Public health toxicology	food additives	gases	defoliants	incapacitating agents
Chemotherapy	household toxicology	dusts	fumigants	defense

To many people, the coroner's chemist of 30 years ago, the forensic toxicologist of today, represents the field of toxicology. These specialists, armed with the penetrating methods of modern analytical chemistry as applied to body tissues and fluids, with a knowledge of poisons, their doses and effects, and trained in the rules of evidence, provide expert testimony when a death, injury, or disease is due to poisoning. Although, formerly, toxicology was concerned largely with the recognition and treatment of poisonings, today, this important work consumes a small fraction of the professional talent and money expended. The emphasis now is safety, the prevention of injury by such means as: (1) providing working conditions such that earning a livelihood does not cost the worker his health or his life, whether the work is in the factory, the field, or the house; (2) providing foods that nourish and medicines that heal without causing illnesses that destroy; and (3) helping to provide food, clothing, comfort, and recreation for a burgeoning population while maintaining an environment of optimal quality for the human race. The knowledge of toxic effects furnishes the basis for this work.

CLASSIFICATIONS OF TOXICITY

Several levels of effect may be defined: (1) doses that produce no detectable effects—*i.e.*, cause no symptoms; (2) doses that produce useful or therapeutic effects; (3) doses that produce toxic or injurious effects; and (4) lethal doses. As a rule of thumb, toxicologists assume that doses increase about tenfold between each of the four levels. Such a prediction holds fairly well for morphine and for salicylates, for example, but is by no means universally applicable. Many substances are less lethal than predicted—*e.g.*, the sulfonamide drugs, atropine, and especially penicillin—and others are more lethal; *e.g.*, digi-

Table 2: Selected Household Products and Drugs of Significant Toxicity

	supertoxic (class 6)	extremely toxic (class 5)	very toxic (class 4)	moderately toxic (class 3)	slightly toxic (class 2)	practically nontoxic (class 1)
Probable lethal dose (human)						
Amount per kg	less than 5 mg	5–50 mg	50–500 mg	0.5–5 g	5–15 g	more than 15 g
For a 70 kg (154 lb) man	a taste (less than 7 drops)	7 drops–1 tsp	1 tsp–1 oz	1 oz–1 pt	1 pt–1 qt	more than 1 qt
Household products and farm chemicals	a few insecticides, fungicides, rodenticides	drain and sewer cleaners (caustics), fireplace flame colours, some pesticides	disinfectants, degreasers, rust removers, some depilatories, fire extinguishing liquids, many agricultural chemicals	polishes, hair dyes, permanent waves, motor fuels, antifreeze, bleaches, turpentine	most cosmetics, lubricating oils, soap products, waxes, incense, some insecticides	foods, candles, mucilage, pure soap, lead pencils, modelling clay
Drugs	digitoxin atropine cyanide heroin nicotine strychnine	amphetamines antihistamines morphine mercury bichloride ephedrine amobarbital (Amytal) ergot	aspirin phenobarbital boric acid fluoride quinidine chlorpromazine	ether kerosene isopropyl alcohol meprobamate methanol Enovid Dilatin	ethyl alcohol milk of magnesia saccharin	kaolin liquid petrolatum talc aluminum silicate

talis. Quantitative toxicity must be measured for each substance, however; toxicity is unpredictable even for substances with nearly identical chemical structures.

Degrees of toxicity

To permit rapid estimates of the chances for recovery of poisoned patients when the ingested material has been identified, systems of toxicity ratings have been proposed. One such system defines toxicity classes based on six commonly-used measures, such as a taste, a mouthful, and an ounce. Substances so poisonous that a taste (less than seven drops) probably would kill a man are rated 6, "supertoxic." At the other end of the scale are found substances in class 1, "practically nontoxic," with probably lethal doses of a quart or more. Examples of selected household products and of drugs are offered in Table 2. A few household products warrant ratings of 5 to 6; *e.g.*, concentrated caustic cleaners, some insecticides, and other pesticides. About 70 percent of the consumer products considered in one survey were rated 3 or lower. Most cosmetics and toiletries are found in class 2 (slightly toxic); others—*e.g.*, those with high concentrations of alcohol—were given toxicity ratings of 3. Among drugs, many well-known natural products are rated 5 and 6, including digitoxin, nicotine, morphine, and strychnine. The toxicity of aspirin (rating, 4) and its ready availability in many homes confers on aspirin a special hazard for children. The distinction between toxicity and hazard should be noted. Toxicity is the ability of a substance to injure; hazard is the probability that the circumstances of exposure are such as to produce injury.

The search for increasingly effective pesticides, antipersonnel agents, and drugs that has included efforts to isolate and identify naturally occurring toxins, has also revealed substances (Table 3) more toxic than any cited in Table 2. These toxicities can be contrasted with those of the insecticide parathion, the oral lethal dose of which is three or more milligrams per kilogram in rats, and the fungicide, methyl mercury dicyandiamide, the intraperitoneal lethal dose of which is 0.4 milligram per kilogram in mice.

Table 3: Some Extremely Toxic Substances

	estimated average lethal dose (mg/kg)	species	route
Botulinus toxin	0.00015	man	oral
Palytoxin*	0.00015	mouse	intravenous
Tetrodotoxin	0.01	mouse	intravenous
Aflatoxin	0.03–3	duckling	oral
Sarin	0.05	mouse	intravenous

*From a plant growing on coral.

Five types of classification

Poisons are sometimes classified on the basis of origins, of form, of chemical nature, of activity, and of use: (1) With respect to origins: (*a*) toxins of plant and animal origin are of special interest to veterinary toxicologists, medicinal herbalists, chemists, metabolic biochemists, and nutritionists; (*b*) industrial chemicals and minerals are of concern to occupational physicians and industrial hygienists; (*c*) drugs are the focus of pharmacologists and clinical toxicologists. (2) With respect to form: (*a*) radiation: radioactive nuclides (*e.g.*, radioactive iodine); (*b*) gases: carbon monoxide, hydrogen sulfide, war gases; (*c*) liquids: solvents; (*d*) solids: asbestos, metals. (3) With respect to chemical nature: (*a*) acids: oxalic; (*b*) alkalies: lye, ammonia; (*c*) alcohols: ethyl, methyl; (*d*) amines: aniline; (*e*) heavy metals: mercury, lead, cadmium; (*f*) hydrocarbons: kerosene, benzene; (*g*) chlorinated hydrocarbons: dichlorodiphenyltrichloroethane (DDT), dieldrin, carbon tetrachloride, polychlorinated biphenyls; (*h*) organic phosphates: parathion, malathion, and others. (4) With respect to activity: (*a*) corrosives: acids, alkalis, phenols; (*b*) irritants: (this class formerly comprised toxic elements but is now obsolete); (*c*) systemic poisons: convulsants, central-nervous-system stimulants and depressants, peripheral-nerve poisons, muscle poisons, metabolic poisons, blood poisons. (5) With respect to use: (*a*) pesticides or economic poisons: herbicides, rodenticides, insecticides, acaricides, fungicides, fumigants, germicides; (*b*) chemotherapeutic agents: antibiotics, anthelminthics, and parasiticides in general; (*c*) homicidal agents: chemicals used in executions (cyanide), in homicide (arsenic, strychnine), in suicide (barbiturates, etc.), in war gases (organic phosphates, mustards); (*d*) riot-control agents: chlorobenzylidine malononitrile, chloracetophenone—"tear gas."

ACUTE AND CHRONIC POISONING

Acute poisoning is characterized by a sudden onset and a rapid rise to intense effect; the symptoms run a short course that does not necessarily terminate in death. Acute poisonings frequently follow a single, relatively large dose. Chronic poisoning, in contrast, is characterized by a gradual and unnoticed development, a gradual onset of symptoms, and a slowly increasing weakness. The illness may appear in any one or more of a number of ways; for example, it may involve the nervous system, the blood and blood-forming tissues, or the gastrointestinal (digestive) tract. Chronic poisonings, often with frequent remissions and recurrences, as a rule develop from protracted or repeated exposures to small doses. The danger of chronic poisoning is high with some compounds, low with others; a few compounds never cause it. Identification of the causes of chronic poisonings may be exceedingly difficult.

Factors affecting toxicity

In both acute and chronic poisoning, the size of the dose required to elicit a toxic response depends on many factors: (1) the physical form; whether the substance is a gas, a liquid, or a solid; (2) certain chemical properties; especially solubility in water or in oil; (3) the chemical nature of the substance; *e.g.*, whether it is a hydrocarbon, an alcohol, or an amine; (4) the route of exposure; *e.g.*, the skin, eye, lung, gastrointestinal tract, or by way of the

veins or the muscles; (5) the species exposed; (6) the characteristics of the individual; *e.g.*, age, sex, state of health, susceptibility, heredity; (7) the kind of biological response; *e.g.*, local or systemic, characteristic or not, whether tolerance develops, whether there is interaction with other agents.

The susceptibility of some individuals and the toxic potentialities of many drugs combine to make drug-induced illness so common that 1 out of 20 acute medical hospital admissions for acute disease is for a poisoning. In addition, in some hospitals there are as many admissions for acute alcoholic intoxication. One in ten hospitalized persons is treated for drug-induced disease during hospitalization. Although most of the 1,600,000 incidents of ingesting possible poisons that occur annually in the United States are of little moment, more than 80,000 are serious enough to appear in the records of the National Clearinghouse for Poison Control Centers. Drugs are the substances most frequently involved, with cleaning and polishing agents, pesticides, cosmetics, solvents, various botanicals, and miscellaneous agents listed in order of frequency of ingestion. The Poisoning Treatment Centre of the Royal Infirmary of Edinburgh found that most (80 percent) of the acutely poisoned persons treated there suffered from "self-poisoning," defined as "a conscious, often impulsive manipulative act, undertaken to secure redress of an intolerable situation." The remaining 20 percent were distributed evenly between accidental poisonings and determined suicidal attempts; only a few cases were instances of homicidal poisonings. Fifty percent or more of the Edinburgh patients had psychiatric problems that were severe enough to require subsequent treatment.

Poisonings, now considered a common emergency, are most frequent in the young adult, 18–25 years of age; there is an equal distribution between males and females. Fatal poisonings often follow overdoses of the nonbarbiturate sedatives, the hypnotics, the tranquillizers and antidepressants (sedatives allay excitement; hypnotics induce sleep; tranquillizers have a quieting and calming effect), the widely prescribed barbiturates, aspirin and other salicylates, and the drugs taken for cultural or social reasons —ethyl alcohol and, recently, morphine and other opium derivatives. Carbon monoxide and other gases account for at least half of all fatal poisonings.

Recognition that an acute illness is a poisoning may come by testimony or by circumstantial evidence; for example, a suicide note or a partially empty container; in general, specific diagnostic features are rare. In the absence of head injury, poisoning is the commonest cause of unconsciousness from adolescence to middle age. An acute poisoning is always an emergency.

When the poisoned person is brought to the medical centre, the physician immediately assesses the status of the patient, the degree of impairment of consciousness, of respiratory failure, of circulatory failure, or of other injuries. Respiration and circulation of the blood must be maintained so that other emergency measures can be effective. Measures to prevent further absorption of swallowed poisons traditionally include gastric lavage (washing out the stomach), inducing vomiting, or administration of an adsorbing substance, such as activated charcoal.

Lavage is dangerous after the ingestion of nonirritating, deceptively bland hydrocarbons (*e.g.*, kerosene) or of corrosive agents (*e.g.*, concentrated acids or lye). Certain recent studies have questioned the practice of stomach lavage (1) as potentially dangerous in unconscious persons, and (2) as a strenuous and traumatic experience of variable but usually limited efficacy. Instead, inducing of vomiting by administration of syrup of ipecac is recommended.

Life-saving treatment depends primarily on intensive supportive therapy; *i.e.*, on expert nursing and on vigilance in maintaining vital functions. From the mid-1940s to the present, deaths from ingesting potentially lethal quantities of barbiturates have decreased from 25 percent to less than 1 percent as a result of combatting shock, maintaining respiration, and placing patients under the expert and conservative care of nurses and physicians in teams and special facilities analogous to the intensive care units for heart patients. The role of antidotes is often exaggerated, there being specific antidotes available in less than 2 percent of all poisonings.

The most effective measures for reducing incidence of poisoning are those of prevention: the keeping of medicines under lock and key, the physician's giving thought to limiting the prescribing and stocking of sedatives and to watching for side effects of drugs and drug combinations; and the segregation and guarding of dangerous pesticides and household chemicals.

Poison-control centres

Beginning in 1953, under the leadership of the American Academy of Pediatrics, certain hospitals in the United States began collecting and furnishing information on poisonings through units called poison-control centres. Over 500 such centres are now registered with the National Clearinghouse for Poison Control Centers established in 1957 by the Public Health Service. In many cities, the telephone directory lists a poison-control centre. Government-sponsored poison information services are maintained in several cities in Great Britain and elsewhere in Europe. These services and the National Clearinghouse serve as central repositories for information on new and old products and for case reports of poisonings from local centres and from other government agencies. Statements of product composition and toxicity and of treatment are distributed in the United States to poison-control centres.

Sophisticated drug-information services are being developed in certain teaching hospitals under the supervision of drug specialists; *e.g.*, clinical pharmacists. These services furnish information and advice to physicians not only on poisonings and on adverse drug reactions but also on any aspect of drug use in patient care.

Chronic poisoning. Chronic poisoning frequently develops gradually after long-continued exposures to relatively small, repeated doses; there is characteristically a symptom-free latent period. This type of poisoning represents cumulative effects, usually from repeated undetectable injuries that are not completely repaired in the intervals between exposures. In exceptional circumstances these injuries may follow a single dose; *e.g.*, of a radioactive element, of a cancer-causing substance, or of a toxic metal. Even rarer are chronic effects that suddenly appear when an effective acute dose is given in divided portions; an example is the paralytic response to tri-*o*-cresyl phosphate in chicks when one-thirtieth of the acute paralytic dose is given daily for a month. Cumulative effects are often but not necessarily associated with a buildup of the poisonous substances in the body. The body burden of lead is increased in chronic lead poisoning; chronic carbon monoxide poisoning, on the other hand, does not bring accumulation of carbon monoxide; and there may be scarcely any concentration of beryllium in the tissues in chronic beryllium disease with widespread pulmonary granulomatosis—formation in the lungs of the type of connective tissue that develops in the process of wound healing and scar formation. Furthermore, accumulation is not synonymous with illness; witness the slow but demonstrable increase in skeletal fluoride during the first five or six decades of life in persons with low fluoride intakes who show no functional evidence of any fluoride effect. Certain delayed, long-lasting effects of exposures to substances foreign to the body, such as sensitizations or immune responses, are not usually classed as chronic poisoning.

The group of symptoms in a chronic poisoning may be entirely different from that in an acute poisoning by the same agent. Acute and chronic mercury poisoning furnish classic examples. Fatal, acute inorganic mercury poisoning, such as follows the ingestion of mercury bichloride (corrosive sublimate) runs a catastrophic course of nausea, vomiting, and diarrhea, leading to a shocklike, circulatory collapse, either with death in a few hours or, for survivors of this phase, to death in a few days from kidney failure. Chronic mercurialism sometimes takes the form of progressive kidney disease, but the affected person may instead show profound personality changes, sometimes other mental effects (drowsiness, confusion)

or predominantly nervous-system signs (tremour, pain, or fever) or may develop muscle weakness and tenderness or may show abnormalities of the heart and circulatory system.

Types of treatment

Treatment cannot be generalized. For some important poisonings, specific antidotes of variable efficacy are available. Chronic lead poisoning in adults yields, on occasion, to treatment with ethylenediaminetetraacetic acid (EDTA); in children afflicted with the preventable effects of lead on the brain (one-half of all chronic cases) EDTA and dimercaprol (BAL) may be life-saving. Unfortunately, of the 75 percent who survive, a third suffer permanent brain damage with mental retardation and sometimes blindness.

In most chronic poisonings, treatment is supportive rather than specific. Removal from further exposure, good nursing care, and the remarkable restorative potential of the body favour recovery. Irreversible alterations in structures or in functions of susceptible parts of the body may cause crippling, even when the poison no longer is present and undamaged tissues function normally.

FATE OF POISON

A poisonous effect, acute or chronic, whether discerned as an impairment of function or as a derangement of structure, usually can be traced to the presence of the toxic agent in the affected area. The degree and kind of injury is often dependent (1) on the amount as well as (2) on the form in which the agent reaches the tissue, (3) on what happens to the agent there, and (4) on how long it remains. One of the keys to understanding how poisons injure lies in discovering what the body does to the poison (variously described as its fate, its metabolism, or its biotransformation).

Introduced into the stomach or gut, toxic substances may be absorbed or they may be destroyed by digestive action (as is rattlesnake toxin, a mixture of proteins). Insoluble substances, when swallowed, may pass through the body virtually unchanged and be lost in the feces.

The speed with which a poison swallowed, inhaled, spread on the skin, or injected enters the body tissues depends on (1) its solubility in fat or water, (2) its degree of irritancy, (3) its chemical reactivity, (4) the toxicologic reactivity, and other factors. In the bloodstream, portions are carried in greater or lesser amounts to all parts of the body. Gaseous substances stable in the body are usually exhaled. Some substances are tightly bound by (enter into stable combination with) blood constituents, especially proteins, and are slowly released. Particulates are apt to be engulfed by scavenger cells, the macrophages, and carried to the special trapping cells of the spleen, liver, bone marrow, or lymph nodes. Insoluble particulates may thereby create special hazards if their properties are dangerous (*e.g.*, if they are radioactive) and if mobilization (release from the tissues) is slow. Toxic substances with high fat solubility pose special problems. Membranes contain fatty layers across which fat-soluble substances are readily transported during initial absorption as well as during reabsorption from newly formed urine in the kidney tubule. Reabsorption tends to retard excretion and to build up the body load, obviously increasing the hazard of poisoning. Water-soluble substances generally pass out in the urine; some, like penicillin, are swiftly excreted (half is gone in 30 minutes). In contrast, substances of high fat solubility, such as alcohol, tend to remain longer; in the case of alcohol, until it is oxidized.

The body has some chemical defenses. Metabolic machinery in the cytoplasm (cell substance outside the nucleus) alters foreign molecules by various methods, including oxidizing, reducing, or hydrolyzing them (breaking them down by adding of water). All of these chemical changes tend to form products that are less fat-soluble, more water-soluble, frequently more acidic or basic than the original molecule, and therefore less resorbable and much more readily excreted in the urine. In these ways, such processes detoxify the foreign molecule and have long been known as detoxication mechanisms. Natural foodstuffs may contain traces of biologically active substances that, if retained unchanged, may be poisonous. The various detoxifying processes therefore have survival value.

The products of the chemical reactions within the body are occasionally more poisonous than the original molecules. A famous example is the transformation of fluoroacetate (inherently not toxic) into the extremely toxic fluorocitrate, which blocks a key enzyme, aconitase, of an important energy-producing mechanism (the citric-acid cycle) of the mitochondria (the energy-producing structural elements of cytoplasm).

Toxic substances may be stored in the body, sometimes briefly (carbon monoxide as carboxyhemoglobin for hours), sometimes for years (uranium dioxide particles in lymph nodes of the lung, strontium-90 in the skeleton, DDT in fatty tissue, silver in the skin).

Manner of excretion

Excretion occurs mainly by way of the urine or the feces, or both. Some solvents (certain chlorinated hydrocarbons, for example) are exhaled in detectable traces over periods of weeks. Traces of lead, mercury, and arsenic are deposited in hair. Sweat contains detectable quantities of certain poisons. Excretion in milk, though uncommon, is known; epidemics of sometimes fatal milk sickness have afflicted those drinking milk from cows feeding on white snakeroot (*Eupatorium rugosum*).

MECHANISM OF ACTION

Although most poisons act at several sites in the body, simultaneously or in sequence, not infrequently, the injury to one organ or system is predominantly responsible for serious illness or death. Cyanide blocks a basic energy-providing mechanism (cytochrome oxidase) in most cells. Many substances in overdose injure the liver; the list includes such diverse materials as carbon tetrachloride, *Amanita phalloides* (death cap mushroom) toxins, and elemental phosphorus. Carbon tetrachloride also destroys kidney tissue, as do mercury and arsenic.

Substances that break down red blood cells (hemolysis), such as primaquine and a number of other drugs, are more likely to cause trouble to individuals anemic because of hereditary deficiency in the enzyme glucose-6-phosphate dehydrogenase. Hemolysis can secondarily produce acute kidney failure by blockage of the tubules. Carbon monoxide acts primarily in the blood, combining with hemoglobin to form carboxyhemoglobin in a combination 200 times tighter than with oxygen, thereby effectively preventing the transport of vitally essential oxygen to the brain, heart, muscles and other tissues. Squill can fatally affect the heart by overexciting it. Recently, sudden deaths from inhaling a variety of substances (gasoline and fluorocarbon propellants, for example) have been traced to another kind of heart injury, a loss of coordination of muscle contraction so that individual muscle fibres beat at random; the heart fails to propel blood, and death swiftly ensues. Intoxications prominently involving the central nervous system are commonplace: drunkenness with alcohol, the "high" of marihuana, the ecstasy or the terror of lysergic acid diethylamide (LSD), convulsions from cocaine, poisoning with strychnine, organic phosphate, or dieldrin, destruction of brain tissue by methyl mercury or by severe exposure to carbon monoxide. Injuries to bone tissues are rare but not unknown; the bone cancers of the radium-dial painters of World War I are a frequently cited example.

The dislocations and distortions induced by poisons into the normal workings of organs and tissues can be useful in revealing secrets of the life processes. The inhibitory role of recurrent nerve branches in the spinal cord, for example, was made clear by the effect of strychnine in blocking their normal function. The neurohormone (acetylcholine) was discovered when the enzyme that normally inactivates acetylcholine in seconds was inhibited by a primitive ordeal poison (eserine).

Despite a multitude of experiments and the availability of immensely powerful new tools, only a few mechanisms of toxic action have been described in detail. The limit is often ignorance of the normal action that is disturbed by the poison. From the volumes of experimental work, a

few examples of the varied natures of toxic actions have been selected and are listed (Table 4), more to illustrate the diversity than to claim detailed knowledge, and certainly not to suggest that the mechanisms of all or even of many are fully understood.

Table 4: Various Toxic Reactions

site	nature of toxic effect	example of poison
Cell membrane	destruction	lye; phenol; heat
	altered permeability	mercury; squill, carbon tetrachloride, hydrocarbons
	formation prevented	penicillin in susceptible bacteria
Blood	interference with circulation by clotting	some snake venoms
	interference with oxygen access in lung	chlorine, nitrous oxides, ozone
	interference with oxygen transport	carbon monoxide, nitrites
	clotting prevented	warfarin
Nerve	transmission blocked along fibre	cocaine
	release of neurohormone blocked	botulinus toxin
	access to receptor blocked	strychnine, atropine, curare
	enzyme inactivating neurohormone inhibited	organic phosphates
Cell function	growth interference	colchicine, nitrogen mustard
	DNA structure or function altered	nitrosamine, radiation, actinomycin D
	energy-supplying cycle blocked	fluoroacetate
	antimetabolite	methotrexate

Synergism

Interactions between similar or dissimilar compounds can be important. Two organic phosphates, for example, may be synergistic (each may enhance the effect of the other) in their insecticidal action. For this reason, when several chemically-related pesticides are used as mixtures, the amounts permitted as food residues are set for the class and not for individual pesticides.

Another kind of interaction has been identified. Some pesticides (DDT, for example) and some drugs are able to stimulate the production by the liver cells of enzymes that account for much of the detoxication of foreign molecules (including drugs). While the traces of pesticides in foods as consumed are insufficient to stimulate these enzymes, occupational exposures either in manufacture or use may possibly do so. By this type of interaction, certain drugs (*e.g.*, barbiturates) are known to increase the rate of destruction of other drugs, thereby necessitating increase in dosage to maintain therapeutic efficacy.

ATTAINMENT OF SAFETY AGAINST POISONING

The mission of the toxicologist is most importantly safety —ensuring that injury will not result from chemical exposures whether of consumer or housewife, factory worker or chemist, farmer, astronaut, or submariner. The toxicologist's intent on safety is joined by the conservationist, who seeks to re-establish and maintain a pristine quality of the human environment. In these efforts, as in all phases of the practical mission of the toxicologist, judgments repeatedly and inevitably balance the benefit against cost or risk. At some sites, trace pollution of the environment must be accepted or an industry closed. Higher yields and better quality must more than repay the farmer the cost of pesticides. The use of chemicals in food production and processing increases the amount and variety of foodstuff in the supermarket; some items unavoidably contain additives, in traces judged to be innocuous. Modern drugs add to the general health but frequently induce side effects, toxic effects that may endanger life.

Testing new drugs and other chemicals

The role of the toxicologist includes the search for antidotes and for means of treating toxic injuries, but a prime task in the development of new drugs, food additives, pesticides, and household or other chemicals is to identify and eliminate unacceptably toxic substances. Toxicity evaluation begins by experiments, generally in the mouse or rat, to classify the substance with respect to its lethality. Effects of repeated doses—*e.g.*, daily for a week or a month or more—are next observed, usually in at least two species, to learn (1) the sites and kinds of any characteristic injuries from exaggerated doses, (2) the pattern of toxic responses, and (3) the size of a tolerated or asymptomatic dose. For food additives or drugs to be administered over protracted periods, further studies are stipulated: customarily (1) lifetime (*i.e.*, two-year) studies of rats, or mice, and (2) one- or two-year studies of dogs or monkeys or both. Additional research may include (1) investigation of the metabolic fate of the substance in at least one species and preferably several; (2) tissue analyses to discover any sites of accumulation; (3) dosing pregnant animals to detect fetal injuries or abnormalities; (4) studies on actively dividing cells to detect chromosome damage or mutations; and (5) detailed and frequently protracted observations designed to detect cancer-producing potential. All such data are brought to focus in the selection of a tolerance value for a pesticide in a raw agricultural product, or an acceptable daily intake (ADI) for an intentional food additive, to provide margins of safety in human use. Candidate new drugs that have shown, through such animal tests, evidence of reasonable safety, are then evaluated in progressively larger groups of human subjects. In phase 1 of such studies in human beings, observations often include the fate as well as the effects of minimal doses on a limited number of normal subjects. Phase 2 explores the efficacy of the new drug as a curative, prophylactic, or diagnostic agent in small groups of selected ill patients. In phase 3, the usefulness and drawbacks (*i.e.*, the medicinal value and the side effects) are recorded on larger groups of patients, when specialists employ the investigational new drug under conditions of practice. In the United States, satisfactory performance throughout normally leads to a new drug application (NDA) and permission from the Food and Drug Administration to market the product. Costs of toxicity testing, including chronic tests on a single compound, are high.

Industrial exposures typically occur when workmen inhale dusty or contaminated air. To provide standards for safe working conditions, the toxicologist compares the concentrations of toxicants in factory air and the consequent illness or health with the effects in experimental animals breathing controlled concentrations of the same toxicants in laboratory studies. All available evidence enters into the choice of a Threshold Limit Value (TLV), a concentration to which "it is believed that nearly all workers may be repeatedly exposed, day after day, without adverse effect."

BIBLIOGRAPHY. F.A. PATTY (ed.), *Industrial Hygiene and Toxicology*, vol. 2, *Toxicology*, 2nd ed. (1963), systematic coverage by classes, by compounds, and by elements of important occupational exposures; R.H. WHITTAKER and P.P. FEENEY, "Allelochemics: Chemical Interactions Between Species," *Science*, 171:757–770 (1971), a review of allomones by function and by chemical nature; M.W. MILLER and G.G. BERG (eds.), *Chemical Fallout: Current Research on Persistent Pesticides* (1969), includes biologic and ecologic details; UNITED STATES DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, FOOD AND DRUG ADMINISTRATION, *Appraisal of the Safety of Chemicals in Foods, Drugs and Cosmetics* (1959), descriptions of basic methods of toxicological evaluation; UNITED STATES DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, *Report of Secretary's Commission on Pesticides and Their Relationship to Environmental Health*, 2 pt. (1969), compilations of factual material, with critical summaries on the status of problems, use, benefit, and risk; R.T. WILLIAMS, *Detoxication Mechanisms*, 2nd ed. rev. (1959), an encyclopaedic compilation of biochemical information; NATIONAL RESEARCH COUNCIL, FOOD PROTECTION COMMITTEE, *Evaluating the Safety of Food Chemicals* (1970), an up-to-date, authoritative statement of purposes and values of current procedures; *Use of Human Subjects in Safety Evaluation of Food Chemicals* (1967), a broad and searching inquiry into a major problem of safety evaluation, discussing experimental, legal, ethical, and philosophical aspects; *Toxicants Occurring Naturally in Foods* (1966), summaries of information on classes of toxic substances naturally present in foods; J.M. KINGSBURY, *Poisonous Plants of the United States and Canada* (1964), descriptions of poisonous plants, algae, and fungi; S.M. HERSH, *Chemical and Biological Warfare: America's Hidden Arsenal* (1968), the history, uses, and some of the implications of chemical and biological warfare.

(H.C.Ho.)

Poker

Poker is a card game or, more properly, a family of card games. Although played throughout the world, it has its greatest popularity in North America and at one time was known as the national card game of the United States.

A Poker hand consists of five cards. Players try to build combinations of two or more cards of a kind, five-card sequences, or five cards of the same suit. The more unusual hands rank higher than the more common.

Poker is essentially a game of money management and is practically never played except for purposes of gambling. Each deal is a separate game in which there is a pot, which is the aggregate of all bets and antes made on that deal. Players try either to win the pot or to get out as cheaply as possible if they cannot win it. There are one or more rounds of betting. At the conclusion of the last round, the players who have remained show their hands. All money in the pot, including the bets made by players who dropped out after an earlier round, is taken by the player with the best hand or, in some forms of the game, divided among several hands. A player may bluff by betting that he has the best hand when in fact he knows he does not. He can win by bluffing if players with superior hands do not match, or call, his bet.

History. The earliest known game of the Poker family evolved in Europe and was mentioned in literature in 1526. While Poker is essentially a five-card game, this was a three-card game known as *primero* in Spain, *primiera* in Italy, and *la prime* in France. The game involved betting, and the valued combinations were three of a kind, pairs, and three of the same suit, called a flux (flush).

European beginnings

By 1700 the betting and bluffing aspects of these games had been incorporated in five-card games such as Brag in England, Pochen in Germany, and Poque in France. In these early games the importance of bluffing, or betting with a poor hand, to get other hands to drop out of the pot may be seen from the fact that *pochen* is a German verb meaning "to bluff." It further appears that the French name Poque is a corruption of the German and that the name Poker is a corruption of the French.

The game was carried to America by the French colonists of the Louisiana territory and spread up the Mississippi to the rest of the country. Poker itself was first mentioned in 1829 in the memoirs of Joe Cowell, a touring English actor. By 1834 it was adapted to the 52-card deck and made its first appearance in books on card games in 1858, but it was treated as an old game in those references. Poker undoubtedly received its first great impetus during the Civil War, when the game was much played among soldiers on both sides.

In Brag, Pochen, and Poque there was just one round of betting. The showdown—exposing the hands to determine the winner—took place after this one round. The rank of hands was four of a kind (high); full house (three of one kind and two of another); three of one kind; two pairs; one pair; and no pair (high card). Cards ranked ace (high), king, queen, jack, ten, and so on down to deuce. There was, and is, no distinction among suits.

Straights (five cards in sequence, regardless of suit), flushes (five cards of the same suit, regardless of sequence), and additional rounds of betting were added later. Two principal forms of the game developed: (1) Closed (Straight or Draw Poker), in which each player's hand is dealt face down and (2) Open, or Stud, Poker, in which each player is dealt one or two cards face down and the rest of his hand one card at a time face up.

The spread of Poker from the United States back to Europe started when Col. Jacob Schenck, U.S. ambassador to Great Britain, introduced it to members of the court of Queen Victoria around 1870. Schenck actually prepared a set of rules for the Queen, who had expressed interest in the game. This was the first known work devoted exclusively to Poker. The game did not reach the rest of Europe until brought there during World War I by the American Expeditionary Force.

The first book on Poker

For many years Poker was considered to be a game for men only, but by the mid-20th century it was often played by women as well. Draw and Stud Poker remain classic men's games, however.

GENERAL PRINCIPLES OF POKER

Poker can be played by any number from 2 to 14, but 6 or 7 make the best game. It is almost always played with the standard 52-card pack. All suits are of equal value, and cards rank A (high)–K–Q–J–10–9–8–7–6–5–4–3–2, although the ace may also be considered low in forming a straight with the 5–4–3–2.

Rank of hands. The traditional ranking is as follows: (1) straight flush (five cards of the same suit and in sequence; the highest, A–K–Q–J–10, is called a royal flush); (2) four of a kind (as 6–6–6–6 and any fifth card); (3) full house (three of one kind and a pair, such as K–K–K–8–8); (4) flush (five cards of the same suit, regardless of sequence); (5) straight (five cards in succession, regardless of suit; the highest is A–K–Q–J–10, and the lowest is 5–4–3–2–A); (6) three of a kind (as 10–10–10, and any two other cards not a pair); (7) two pairs (as J–J–9–9 and any fifth card); (8) one pair (as 8–8 and any three unmatched cards); and (9) high card, or no pair (hand having no pair, ranked by its highest card or cards).

Determining the winner among hands of the same rank is a problem often encountered. Among straight flushes, flushes, straights, or hands containing no pair, the one containing the highest card wins. If the high cards are identical, the second highest card decides, and so on. Among hands of four or three of a kind, the higher matched set wins, as 9–9–9–9 beats 7–7–7–7. With full houses, the higher three of a kind wins. With two pairs, the highest pair decides, or if these are identical, the other pair; if these two are identical, the odd card is the deciding factor. Among hands of one pair, the higher pair wins, or if identical, the highest odd card. Identical hands have identical value.

Wild cards. A wild card is one that may be used to represent any card its holder desires. There are any number of wild card games, including Joker, in which the joker is wild and is added to the deck as a 53rd card; Bug, in which the joker is called the bug, and though not completely wild, may be used as a fifth ace or to form a straight or a flush; Deuces Wild, in which all deuces are wild cards; or any other wild cards the players care to designate. When there are any wild cards in a game, the highest hand becomes five of a kind, although in some games a royal flush is allowed to beat any five of a kind.

Canadian Poker. In Canadian Five-Card Stud, a hand of four cards in sequence and open at both ends (a four-straight, such as 6–5–4–3 or 9–8–7–6 but not A–K–Q–J) beats a pair, and four cards of the same suit (a four-flush) beats a four-straight. In some forms only the four-flush is counted.

European Poker. In the European form of Draw Poker, some of the low cards (usually the 2s, 3s, and 4s) are taken out of the pack. Flushes rank above full houses, and the ace may be used as low card at the bottom of a straight, as in 8–7–6–5–A.

Betting. In each deal there are one or more betting intervals. In each interval one player, as designated by the rules of the variant being played, has the privilege or obligation of opening, or making the first bet. This player and each player in turn after him must place in the pot a number of chips (representing money, for which Poker is almost invariably played). After the opening, each bettor must make his total contribution to the pot at least equal the total contribution of any player before him. When a player does this, he is said to have "called," or "seen," the bet. A player who bets more than the previous bettor is said to raise. If a player declines to bet or raise, he discards his hand (drops or passes) and may no longer compete for the pot.

Before the deal, each player may be required to make a contribution to the pot, called an ante. In some variants a player is permitted to check, or stay in without betting, provided no other player has made a bet in that betting interval.

"Showdown"

At the end of the last betting interval there is a "showdown," in which each active player shows his full hand and the highest-ranking hand wins the pot. In practice, especially in informal games, players simply announce

the values of their hands and only show their hands as confirmation. On demand of any player, however, the full hand must be shown, and any false or mistaken announcement is not valid: "the cards speak for themselves."

In Poker legends there were "no-limit" or "sky's-the-limit" games, but in practice some limit is placed on what one may bet in any game. There are three popular methods: fixed limit, pot limit, and table stakes. In the first, no one may bet or raise by more than a limit previously and arbitrarily established by mutual agreement or the house. A limit is also usually placed on the number of raises that may be made in any betting interval. In pot-limit games a player may bet or raise by no more than the amount in the pot at the time the bet or raise is made. In the table stakes method, which most closely approximates the legendary no-limit game, each player's limit is the number of chips he has on the table at the beginning of the deal.

PRINCIPAL FORMS OF POKER

Poker has two main branches: Closed (Straight or Draw) Poker, in which each player's full hand remains concealed until the showdown, and Open (Stud) Poker, in which some but not all of a player's cards are dealt or turned face up.

Straight Poker. In Straight Poker each player is dealt five cards, face down. There is one betting interval, followed by a showdown. This was the original form of Poker, but it has practically passed out of existence.

Draw Poker. In Draw Poker, there is usually an ante, after which each player is dealt five cards face down. There is a betting interval. Then each active player, in turn, beginning at dealer's left, may discard one or more of his original cards (usually no more than three) and receive replacements for them, dealt from the undealt portion of the pack. A player who declines to draw cards is said to stand pat. After the draw there is a final betting interval, followed by the showdown.

Jackpots and Blind Opening

Draw Poker has two main varieties: Jackpots, the variant most played in the United States and Canada, and Blind Opening, played almost to the exclusion of other variants in England, Europe, and the British Commonwealth except Canada and favoured in men's clubs in the U.S. and Latin America. The basic difference between the two is in their betting procedures. In Jackpots there is an ante before each deal. There is usually a requirement that a player must hold a pair of jacks or a better hand to open the betting. If no one opens, players ante again and there is a new deal. In Blind Opening there may or may not be an ante, but before any cards are dealt, the player to dealer's left must make an opening bet, called the blind, usually only one chip or unit. In some games the player to the blind opener's left must then "straddle," or bet double the amount of the first bet.

Stud Poker. In the basic five-card form of Stud Poker, each player receives one card face down, called his hole card, then one card face up. The deal is then interrupted for a betting interval (in some games a betting interval also follows the dealing of the hole cards). There follow three more rounds of dealing, each consisting of giving one face-up card to each active player, with a betting interval after each round. In each betting interval the first bettor is the player with the highest ranking Poker combination among his face-up cards; if two or more players have the same combination, the one nearest the dealer's left bets first. The showdown, in which the hole cards are shown, follows the fourth and last betting interval.

Stud Poker usually is played with pot limit. The Canadians have enlivened it by playing with a fairly large ante, a moderate limit increasing each round, and with special hands (see above *Canadian Poker*). This encourages a player who is dealt something like the 7–6 of a suit to stay for a third card. Then if his next card fits the first two he has a strong hand.

Seven-Card Stud. In Seven-Card Stud, a popular variant of Stud Poker, each player is dealt two hole cards and a face-up card, followed by a betting interval; then three more face-up cards and one final face-down card, each followed by a betting interval. For the showdown each player selects the best five of his seven cards to be his Poker hand. There are also six-card and eight-card variants of this game, in each of which a player ultimately selects five of his cards.

A game with two winners

High-Low Poker. In High-Low the highest-ranking hand and the lowest-ranking hand divide the pot. If two or more hands tie for high or low, they divide their half of the pot equally. Nearly any form of Poker may be played High-Low.

In High-Low Seven-Card Stud, each player may select any five of his cards as his candidate for high hand and any five as his candidate for low hand in attempting to win the entire pot. In some games each player must declare for high, low, or both before the showdown and cannot win unless he fulfills his declaration.

Low Poker or Lowball. This is Draw Poker and is played similarly to Jackpots, except that the lowest-ranking hand wins the pot. In California, where the game is most popular, straights and flushes never count and the ace is always the lowest ranking card. The lowest possible hand is 5–4–3–2–A and is called a bicycle or wheel.

Dealer's Choice. In informal Poker games, each successive dealer may dictate the variant of Poker that will be played. He may also designate certain cards to be wild. Sometimes it is agreed that the dealer may select or invent any variant he wishes, subject to only two restrictions: the dealer may not require any player to ante more than any other player; and, if the game requires a minimum to open (as Jackpots) and is passed out, the same dealer deals again.

RULES OF POKER

No code of Poker rules has been universally adopted. Nearly every club or serious game has certain house rules that may differ in some respects from those of other clubs and groups. The usual practice is to adopt a published code and add the house rules to it. A code prepared by Oswald Jacoby in 1940, and another appearing in *The Official Rules of Card Games* published by the United States Playing Card Company since 1945, are the ones most often adopted in the United States. Poker rules, unlike those of other card games, provide only rectification; there are no penalties for irregularities. Certain departures, such as having the wrong number of cards in hand or cards exposed in dealing, are necessarily treated differently in Draw Poker and Stud Poker.

PLAYING THE HAND

No game rewards skillful play or penalizes bad play more than Poker. Although it is not so complicated as some other games, the player has better control over the result, since he may drop bad hands. Consequently, a good player is less likely to lose to inferior players.

Knowledge of probabilities is important. Chances are about 1 in 2 of receiving a hand of one pair or better, 1 in 5 for a pair of jacks or better, rising to 1 in 590 for a full house or better, and much higher for four of a kind or a straight flush. Odds against improving the original hand in Draw Poker range from 2½ to 1 to improve a pair by drawing three to 1,080 to 1 to make a pair and an ace into four of a kind by drawing two. In Five-Card Stud, after 12 cards have been seen, odds against pairing an ace are 6 to 1 if none of the 12 were aces and 10 to 1 if one was. Chances of receiving a strong hand are better in games where cards are wild or where more than five cards are dealt to each player. Many published works on Poker contain detailed tables of odds and statistics applicable to a variety of situations.

Betting decisions are based on the ratio of expected profit to the amount bet and on chances of winning. If a one chip bet can win ten chips (10 to 1 odds) and chances of winning appear to be 1 in 6, this is a good bet.

Psychology in Poker

It is important to know the mathematics of the game, but it is even more important to know the character of the other players. For example, in playing Draw Poker, A opens. B calls, as do two other players. A draws three cards. B draws one card to a four-flush and makes his flush. A checks without looking at his draw. B bets the

limit. The next two players drop out. A raises the limit. The mathematical odds against his making a full house or four of a kind are over 70 to 1, and these are the only hands in this situation that can beat a flush (A had to have a pair to open, has obviously drawn to this pair, and therefore could not get a flush or a straight flush). Is he silly enough to be raising with just a pair or three of a kind? B decides that he is not. Then A either has made a winning hand or he is bluffing. B must decide, based on his knowledge of A as a player.

The fine points of psychological play, adjustment of strategy to the various forms of the game, and knowledge of when and how to bluff are acquired best through experience.

BIBLIOGRAPHY. See *Oswald Jacoby on Poker* (1947) for the game as played in the U.S.; M. ELLINGER, *Poker* (1950), for the game as played in England. Other works include: A. DOWLING, *Great American Pastime: Notes on Poker, the Game and the Players* (1970); T. REESE and A. WATKINS, *Poker: Game of Skill* (1962); and J.L. CASTLE, *How Not to Lose at Poker* (1970).

(O.J.)

Poland

The location of Poland has been a major force in the molding of the character of its contemporary peoples, their culture, and their society. The country lies between approximately 49° and 55° north latitude and 14° and 24° east longitude. It is in the zone of transition between the continental lowland block of eastern Europe and the mountainous peninsula of western Europe. It also lies midway between the northern and southern extremities of the continent, in the belt of cool, temperate climates, and is at the same distance from the Equator as England and southern Canada. The area of Poland, 120,725 square miles (312,677 square kilometres), is greater than that of Italy, while its population, which exceeded 34,000,000 by the mid-1970s, is comparable to that of Spain. In area and population, Poland (officially, the Polish People's Republic) ranks second among the Communist states of eastern Europe, after the Soviet Union itself.

On the north the country borders the Baltic Sea; the southern boundary runs along the peaks of the Carpathian and Sudety mountain ranges; and the eastern and western boundaries follow, respectively, parts of the courses of the Bug and Odra–Nysa (Oder–Neisse) river systems. Historically, the nation has proved vulnerable to movements and incursions of peoples across its exposed eastern and western flanks, for its northern plains are part of one of the great corridors of the Eurasian landmass. Poland has but three contemporary neighbours. The mainly mountainous frontier with Czechoslovakia runs for 814 miles (1,310 kilometres) in the south; to the east and northeast, the frontier with the Soviet Union is 773 miles (1,244 kilometres) long; and to the west is a 286-mile (460-kilometre) border with the German Democratic Republic. The Baltic Sea border is 326 miles (524 kilometres) long, from the mouth of the Odra (Oder) River in the west to the Gulf of Gdańsk (Zatoka Gdańska) in the east.

A thousand-year heritage

Poland is a country with a thousand years of culture and historical tradition behind it. Although afflicted by hostile incursions, wars, economic difficulty, and the total loss of national independence for well over a century, Poland has nevertheless contributed much of lasting value to world culture. The astronomer Nicolaus Copernicus was one of the creators of modern scientific thought, and Marie Curie (*née* Skłodowska), the discoverer of radioactive elements, provides a more contemporary parallel. The music of Frédéric Chopin has stirred the world. So have the exploits of Poland's military and political leaders: King John III Sobieski defended the whole of Europe against the threat of Turkish invasion, and Tadeusz Kościuszko and Kazimierz Pułaski became heroes in the youthful United States as well as in their native land.

Not until the end of World War I in 1918 did a republic of the west European type emerge in Poland, and this proved to be economically weak and burdened with several large national minorities. During World War II, a fifth of the population, some 6,000,000 people (including almost the entire Jewish population of more than 3,000,000), perished under the Nazi occupation, and the nation suffered great physical devastation. A new Polish state was formed in 1945; the country's boundaries reverted to the original territories of ethnic Poland in the Middle Ages, and, as a result of a substantial migration of Poles and Germans, the ethnic composition of the nation since then has been almost entirely homogeneous. The need for reconstruction faced the nation at all levels; the ensuing 25 years were to witness the transformation of the Polish economy to one in which industry played a much larger role than it had in the past. Economic development was irregular, however, and the Poland of the 1970s continued to be plagued—as many Polish experts pointed out—by the unsatisfactory performance of agriculture, the weak development of transportation, the relatively poor quality of many services, and the damage that rapid industrial growth had caused to the environment. Progress was nevertheless being made in the difficult task of catching up with richer societies.

This article is concerned with the contemporary nation. For history, see POLAND, HISTORY OF. For information on related topics, see CARPATHIAN MOUNTAINS; ODER RIVER; VISTULA RIVER; and WARSAW.

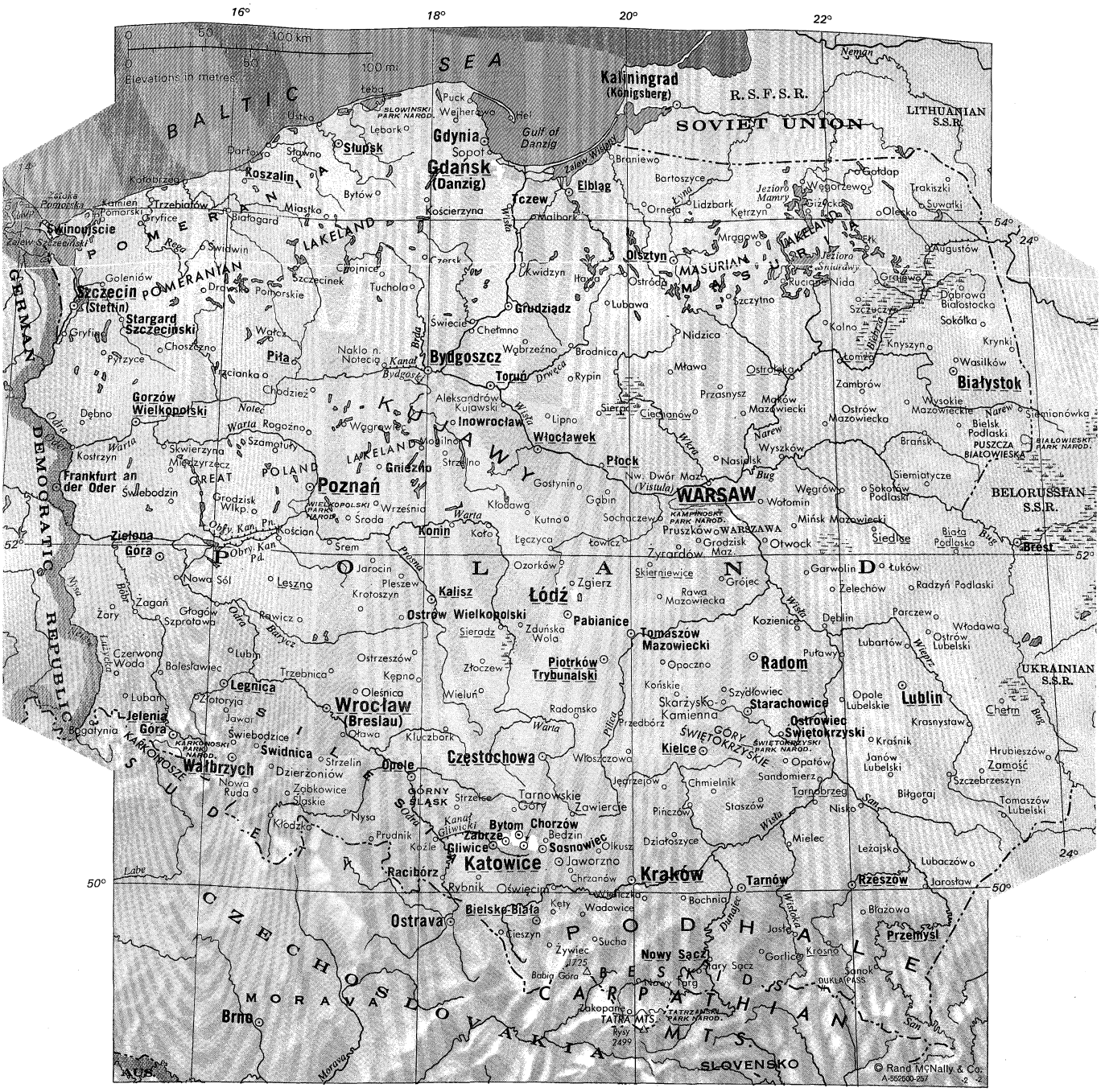
THE LAND

Relief. Poland has an almost entirely lowland character. The average elevation of the whole country is only 568 feet (173 metres); a mere 3 percent of the land lies above 1,650 feet (500 metres), and more than 70 percent lies below 650 feet (200 metres). Over three-quarters of the national territory owes its contemporary surface appearance to the actions of Ice Age glaciers, which advanced and receded over the northern part of the country a number of times during the Pleistocene Epoch. The great and often monotonous expanses of the Polish lowlands, part of the Central European Plain, are composed of geologically recent deposits that lie over a vast structural basin. In the southern portion of the country, by contrast, older and more diverse geological formations are exposed, and the topography is dominated by the mountainous arc of the Carpathians, dating from the Tertiary mountain-building period of about 25,000,000 years ago. Around the northern rim of the Carpathians lies a series of structural basins, separating the mountain belt proper from a much older structural mass, or foreland, that appears in the relief patterns of the region as the Bohemian Massif (Czeszka Masyw Górski), the Sudety, and the Little Poland Uplands (Wyżyna Małopolska). The upthrust of these mountains was a part of the much greater mountain-building movements that gave rise to the Carpathians and, farther west, the Alps.

Carpathian Mountains

This structure helps to explain the division of the country into a series of east–west zones. To the north lie the swamps and dunes of the Baltic coast; south of these is a belt of chaotic terrain with thousands of lakes, the southern boundary of which marks the limit of the last ice sheet. The third zone can be characterized as the Middle Polish Lowlands, whose minimal relief was created by streams issuing from the retreating glaciers. This zone is the Polish heartland, the site of agriculture in places where loess has been deposited over the relatively infertile fluvioglacial deposits. It is here also that Warsaw (Warszawa), the capital, and the cities of Wrocław and Poznań are found. The fourth zone is made up of the older mountains and highlands to the south; though limited in extent, it offers spectacular scenery. The Little Poland Uplands protect the Little Poland Lowlands, in which Kraków lies, from the colder air of the north. In the extreme south are the Carpathians and their foothills.

Climate. Varying types of air masses collide over Poland, a process that strongly influences the character of both weather and climate. The major elements involved are oceanic air masses from the west, cold polar air from the North Atlantic or from Russia, and warmer, subtropical air from the region of the Azores. All the year round, a series of barometric depressions moves eastward along the polar front, dividing the subtropical from the


16°
18°
20°
22°
50
0
100 km
100 mi
Elevations in metres
SEA
BALTIC
Kaliningrad (Königsberg)
R.S.F.S.R.
SOVIET UNION
LITHUANIAN S.S.R.
Gdynia
Gdańsk (Danzig)
Gulf of Danzig
Słupsk
Elbląg
Koszalin
POMERANIA
LAKELAND
Szczecin (Stettin)
Stargard Szczeciński
Bydgoszcz
Toruń
Grudziądz
Inowrocław
Włocławek
Gorzów Wielkopolski
Piła
Poznań
Gniezno
Konin
Kalisz
Łódź
Pabianice
Ostrów Wielkopolski
Tomaszów Mazowiecki
Piotrków Trybunalski
Radom
Płock
WARSAW
Frankfurt an der Oder
GERMAN DEMOCRATIC REPUBLIC
POLAND
Zielona Góra
Leszno
Legnica
Wrocław (Breslau)
Jelenia Góra
Świdnica
Wałbrzych
Częstochowa
Opole
Gliwice
Zabrze
Bytom
Sosnowiec
Katowice
Racibórz
Kraków
Tarnów
Rzeszów
Bielsko-Biała
Ostrava
Nowy Sącz
Przemyśl
Lublin
Starachowice
Ostrowiec Świętokrzyski
Kielce
Siedlce
Białystok
Ostrołęka
Mława
Olsztyn
Ełk
Suwałki
Augustów
Sokółka
Brześć
BELORUSSIAN S.S.R.
UKRAINIAN S.S.R.
CZECHOSLOVAKIA
Brno
SLOVENSKO
CARPATHIAN MTS.
Babia Góra 1725
Rysy 2499
TATRA MTS.
DUKLA PASS
Odra
Wisła
Warta
Bug
Narew
Dunajec
Dniestr
Prut
Labe
Morava
Dyje
Bóbr
Nysa
Kwidzyn
Malbork
Tczew
Chełmno
Grudziądz
Świecie
Brodnica
Nidzica
Olsztynek
Ostróda
Szczytno
Ełk
Giżycko
Węgorzewo
Gołdap
Olecko
Ruciane-Nida
Pisz
Mrągowo
Kętrzyn
Lidzbark
Orneta
Braniewo
Bartoszyce
Frombork
Elbląg
Nowy Dwór Gdański
Tczew
Kościerzyna
Chojnice
Człuchów
Tuchola
Świdwin
Białogard
Kołobrzeg
Trzebiatów
Gryfice
Gryfino
Police
Goleniów
Choszczno
Pyrzyce
Drawsko Pomorskie
Szczecinek
Wałcz
Piła
Nakło n. Notecią
Inowrocław
Mogilno
Strzelno
Gniezno
Września
Środa
Kościan
Leszno
Rawicz
Krotoszyn
Ostrzeszów
Kępno
Wieluń
Oleśnica
Oława
Kluczbork
Strzelin
Dzierżoniów
Świebodzice
Jawor
Bolesławiec
Lubań
Złotoryja
Lubin
Głogów
Żary
Nowa Sól
Żagań
Szprotawa
Zielona Góra
Sulechów
Międzyrzecz
Skwierzyna
Kostrzyn
Gorzów Wielkopolski
Dębno
Myślibórz
Barlinek
Gryfice
Kamień Pomorski
Wolin
Świnoujście
Zalew Szczeciński
Police
Stargard Szczeciński
Pyrzyce
Chojna
Myślibórz
Gryfino
Nowogard
Goleniów
Szczecinek
Drawsko Pomorskie
Tuczno
Wałcz
Czarne
Człuchów
Chojnice
Starogard Gdański
Tczew
Malbork
Nowy Dwór Gdański
Elbląg
Frombork
Braniewo
Bartoszyce
Lidzbark Warmiński
Orneta
Olsztynek
Ostróda
Iława
Kwidzyn
Grudziądz
Świecie
Chełmno
Brodnica
Nidzica
Mława
Ostrołęka
Przasnysz
Maków Mazowiecki
Ostrów Mazowiecka
Sierpc
Ciechanów
Lipno
Włocławek
Płock
Nowy Dwór Mazowiecki
Wyszków
Nasielsk
Pułtusk
Węgrów
Sokołów Podlaski
Siedlce
Mińsk Mazowiecki
Wołomin
Otwock
Grodzisk Mazowiecki
Żyrardów
Skierniewice
Rawa Mazowiecka
Gostynin
Kutno
Łęczyca
Łowicz
Sochaczew
Pruszków
Warszawa
Żyrardów
Grójec
Garwolin
Łuków
Zelechów
Radzyń Podlaski
Biała Podlaska
Międzyrzec Podlaski
Parczew
Włodawa
Ostrów Lubelski
Lubartów
Puławy
Kozienice
Dęblin
Zwoleń
Skarżysko-Kamienna
Końskie
Szydłowiec
Opoczno
Tomaszów Mazowiecki
Zduńska Wola
Sieradz
Ozorków
Zgierz
Łódź
Pabianice
Ostrów Wielkopolski
Kalisz
Krotoszyn
Pleszew
Jarocin
Ostrzeszów
Kępno
Wieluń
Radomsko
Złoczew
Piotrków Trybunalski
Bełchatów
Pajęczno
Kluczbork
Oława
Strzelin
Oleśnica
Wrocław (Breslau)
Dzierżoniów
Świdnica
Jawor
Świebodzice
Bolesławiec
Lubań
Złotoryja
Lubin
Głogów
Żary
Nowa Sól
Żagań
Szprotawa
Zielona Góra
Sulechów
Międzyrzecz
Skwierzyna
Kostrzyn
Gorzów Wielkopolski
Dębno
Myślibórz
Barlinek
Gryfice
Kamień Pomorski
Wolin
Świnoujście
Zalew Szczeciński
Police
Stargard Szczeciński
Pyrzyce
Chojna

colder air, bringing to Poland, as to other parts of northern Europe, cloudy, wet days. In winter, polar-continental air from eastern Europe often becomes dominant, bringing crisp, frosty weather, with still colder Arctic air following in its wake. The warm, dry, subtropical-continental air from the southeast often brings pleasant days in late summer and autumn.

These elements combine to give the overall climate of Poland a transitional—and highly variable—character, between oceanic and continental types. Six seasons may

Six seasons

MAP INDEX (continued)

Danzig see Gdańsk
Darłowo.....54·26n 16·23e
Dęblin.....51·35n 21·50e
Dębno.....52·45n 14·40e
Drawsko Pomorskie.....53·32n 15·48e
Działoszyce.....50·22n 20·21e
Dzierżoniów.....50·44n 16·39e
Elbląg.....54·10n 19·25e
Ełk.....53·50n 22·22e
Gąbin.....52·25n 19·44e
Garwolin.....51·54n 21·37e
Gdańsk (Danzig).54·23n 18·40e
Gdynia.....54·32n 18·33e
Giżycko.....54·03n 21·47e
Gliwice.....50·17n 18·40e
Głogów.....51·40n 16·05e
Gniezno.....52·31n 17·37e
Gołdap.....54·19n 22·19e
Goleniów.....53·36n 14·50e
Gorlice.....49·40n 21·10e
Gorzów Wielkopolski....52·44n 15·15e
Gostynin.....52·26n 19·29e
Grajewo.....53·39n 22·27e
Grodzisk Mazowiecki.....52·07n 20·37e
Grodzisk Wielkopolski....52·14n 16·22e
Grójec.....51·52n 20·52e
Grudziądz.....53·29n 18·45e
Gryfice.....53·56n 15·12e
Gryfino.....53·12n 14·30e
Hel.....54·37n 18·48e
Hrubieszów.....50·49n 23·55e
Iława.....53·37n 19·33e
Inowrocław.....52·48n 18·15e
Janów Lubelski..50·43n 22·24e
Jarocin.....51·59n 17·31e
Jarosław.....50·02n 22·42e
Jasło.....49·45n 21·29e
Jawor.....51·03n 16·11e
Jaworzno.....50·13n 19·15e
Jędrzejów.....50·39n 20·18e
Jelenia Góra.....50·55n 15·46e
Kalisz.....51·46n 18·06e
Kamień Pomorski.....53·58n 14·46e
Katowice.....50·16n 19·00e
Kępno.....51·17n 17·59e
Kętrzyn.....54·06n 21·23e
Kęty.....49·53n 19·13e
Kielce.....50·52n 20·37e
Kłodawa.....52·16n 18·55e
Kłodzko.....50·27n 16·59e
Kluczbork.....50·59n 18·13e
Knyszyn.....53·19n 22·55e
Kolno.....53·25n 21·56e
Koło.....52·12n 18·38e
Kołobrzeg.....54·12n 15·33e
Konin.....52·13n 18·16e
Końskie.....51·12n 20·26e
Kościan.....52·06n 16·38e
Kościerzyna.....54·08n 18·00e
Kostrzyn.....52·37n 14·39e
Koszalin.....54·12n 16·09e
Kozienice.....51·35n 21·33e
Koźle.....50·20n 18·08e
Kraków.....50·03n 19·58e
Kraśnik.....50·56n 22·13e
Krasnystaw.....50·59n 23·10e
Krosno.....49·42n 21·46e
Krotoszyn.....51·42n 17·26e
Krynki.....53·15n 23·47e
Kutno.....52·15n 19·23e
Kwidzyn.....53·45n 18·56e
Łeba.....54·47n 17·33e
Lębork.....54·33n 17·44e
Łęczyca.....52·04n 19·13e
Legnica.....51·13n 16·09e
Leszno.....51·51n 16·35e
Leżajsk.....50·16n 22·24e
Lidzbark.....54·09n 20·35e
Lipno.....52·51n 19·10e
Łódź.....51·46n 19·30e
Łomża.....53·11n 22·05e
Łowicz.....52·07n 19·56e
Lubaczów.....50·10n 23·07e
Lubań.....51·08n 15·18e
Lubartów.....51·28n 22·46e
Lubawa.....53·30n 19·45e
Lubin.....51·24n 16·13e
Lublin.....51·15n 22·35e
Łuków.....51·56n 22·23e
Maków Mazowiecki.....52·52n 21·06e
Malbork.....54·02n 19·01e
Miastko.....54·01n 17·00e
Międzyrzecz.....52·28n 15·35e
Mielec.....50·18n 21·25e
Mińsk Mazowiecki.....52·11n 21·34e
Mława.....53·06n 20·23e
Mogilno.....52·40n 17·58e
Mrągowo.....53·52n 21·19e
Nakło nad Notecią.....53·08n 17·35e
Nasielsk.....52·36n 20·48e
Nidzica.....53·22n 20·26e
Nisko.....50·30n 22·07e
Nowa Ruda.....50·35n 16·31e
Nowa Sól.....51·48n 15·44e
Nowy Dwór Mazowiecki.....52·26n 20·43e
Nowy Sącz.....49·38n 20·42e
Nowy Targ.....49·29n 20·02e
Nysa.....50·29n 17·20e
Oława.....50·57n 17·17e
Olecko.....54·03n 22·30e
Oleśnica.....51·13n 17·23e
Olkusz.....50·17n 19·35e
Olsztyn.....53·48n 20·29e
Opatów.....50·49n 21·26e
Opoczno.....51·23n 20·17e
Opole.....50·41n 17·55e
Opole Lubelskie..51·09n 21·58e
Orneta.....54·08n 20·08e
Ostróda.....53·43n 19·59e
Ostrołęka.....53·06n 21·34e
Ostrowiec Świętokrzyski...50·57n 21·23e
Ostrów Lubelski.....51·30n 22·52e
Ostrów Mazowiecka....52·49n 21·54e
Ostrów Wielkopolski....51·39n 17·49e
Ostrzeszów.....51·25n 17·57e
Oświęcim.....50·03n 19·12e
Otwock.....52·07n 21·16e
Ozorków.....51·58n 19·19e
Pabianice.....51·40n 19·22e
Parczew.....51·39n 22·54e
Piła.....53·10n 16·44e
Pińczów.....50·32n 20·35e
Piotrków Trybunalski.....51·25n 19·42e
Pleszew.....51·54n 17·48e
Płock.....52·33n 19·43e
Poznań.....52·25n 16·55e
Prudnik.....50·19n 17·34e
Pruszków.....52·11n 20·48e
Przasnysz.....51·06n 19·53e
Przemyśl.....49·47n 22·47e
Puck.....54·44n 18·27e
Puławy.....51·25n 21·57e
Pyrzyce.....53·10n 14·55e
Racibórz.....50·06n 18·13e
Radom.....51·25n 21·10e
Radomsko.....51·05n 19·25e
Radzyń Podlaski.....51·48n 22·38e
Rawa Mazowiecka....51·46n 20·16e
Rawicz.....51·37n 16·52e
Rogoźno.....52·46n 17·00e
Ruciane-Nida....53·38n 21·32e
Rybnik.....50·06n 18·32e
Rypin.....53·05n 19·25e
Rzeszów.....50·03n 22·00e
Sandomierz.....50·41n 21·45e
Sanok.....49·34n 22·13e
Siedlce.....52·11n 22·16e
Siemiatycze.....52·26n 22·53e
Siemionówka....52·53n 43·50e
Sieradz.....51·36n 18·45e
Sierpc.....52·52n 19·41e
Skarżysko-Kamienna.....51·08n 20·53e
Skierniewice....51·58n 20·08e
Skwierzyna.....52·35n 15·30e
Sławno.....54·22n 16·40e
Słupsk.....54·28n 17·01e
Sochaczew.....52·14n 20·14e
Sokółka.....53·25n 23·31e
Sokołów Podlaski.....52·25n 22·15e
Sopot.....54·28n 18·34e
Sosnowiec.....50·18n 19·08e
Śrem.....52·08n 17·01e
Środa.....52·14n 17·17e
Starachowice....51·03n 21·04e
Stargard Szczeciński....53·20n 15·02e
Stary Sącz.....49·34n 20·38e
Staszów.....50·34n 21·40e
Stettin, see Szczecin
Strzelce.....50·31n 18·19e
Strzelin.....50·47n 17·03e
Strzelno.....52·38n 18·11e
Sucha.....49·44n 19·36e
Suwałki.....54·07n 22·56e
Świdnica.....50·51n 16·29e
Świdwin.....53·47n 15·47e
Świebodzice.....50·52n 16·19e
Świebodzin.....52·15n 15·32e
Świecie.....53·25n 18·28e
Świnoujście.....53·53n 14·14e
Szamotuły.....52·37n 16·35e
Szczebrzeszyn...50·41n 22·58e
Szczecin (Stettin).....53·24n 14·32e
Szczecinek.....53·43n 16·42e
Szczuczyn.....53·34n 22·18e
Szczytno.....53·34n 21·00e
Szprotawa.....51·34n 15·33e
Szydłowiec.....51·14n 20·51e
Tarnobrzeg.....50·35n 21·40e
Tarnów.....50·01n 21·00e
Tarnowskie Góry.50·27n 18·52e
Tczew.....54·06n 18·47e
Tomaszów Lubelski.....50·28n 23·25e
Tomaszów Mazowiecki.....51·32n 20·01e
Toruń.....53·02n 18·35e
Trakiszki.....54·16n 23·07e
Trzcianka.....53·03n 16·28e
Trzebiatów.....54·04n 15·14e
Trzebnica.....51·20n 17·01e
Tuchola.....53·36n 17·50e
Ustka.....54·35n 16·50e
Wąbrzeźno.....53·17n 18·57e
Wadowice.....49·53n 19·30e
Wągrowiec.....52·49n 17·11e
Wałbrzych.....50·46n 16·17e
Wałcz.....53·17n 16·28e
Warsaw.....52·15n 21·00e
Wasilków.....53·12n 23·12e
Węgorzewo.....54·14n 21·44e
Węgrów.....52·25n 22·01e
Wejherowo.....54·37n 18·15e
Wieliczka.....49·59n 20·04e
Wieluń.....51·14n 18·34e
Włocławek.....52·39n 19·02e
Włodawa.....51·34n 23·32e
Włoszczowa.....50·52n 19·59e
Wołomin.....52·21n 21·14e
Wrocław (Breslau).....51·06n 17·00e
Września.....52·20n 17·34e
Wysokie Mazowieckie....52·56n 22·32e
Wyszków.....52·36n 21·28e
Ząbkowice Śląskie.....50·36n 16·53e
Zabrze.....50·18n 18·46e
Żagań.....51·37n 15·19e
Zakopane.....49·19n 19·57e
Zambrów.....53·00n 22·15e
Zamość.....50·44n 23·15e
Żary.....51·38n 15·09e
Zawiercie.....50·30n 19·25e
Zduńska Wola....51·36n 18·57e
Żelechów.....51·49n 21·54e
Zgierz.....51·52n 19·25e
Zielona Góra.....51·56n 15·31e
Złoczew.....51·25n 18·36e
Złotoryja.....51·08n 15·57e
Żyrardów.....52·04n 20·25e
Żywiec.....49·41n 19·12e

Physical features and points of interest

Babia Góra, *mountain*.....49·34n 19·30e
Baltic Sea.....55·00n 16·00e
Barycz, *river*.....51·42n 16·15e
Beskids, *mountains*.....49·25n 20·15e
Białowieska Park Narodowy, *national park*....52·42n 23·52e
Biebrza, *river*....53·37n 22·56e
Bóbr, *river*.....52·04n 15·04e
Brda, *river*.....53·07n 18·08e
Bug, *river*.....52·31n 21·05e
Bydgoski, Kanał, *canal*.....53·08n 17·50e
Carpathian Mountains.....49·20n 21·00e
Danzig, Gulf of...54·40n 19·15e
Drwęca, *river*....53·00n 18·42e
Dukla Pass.....49·28n 21·40e
Dunajec, *river*....50·15n 20·44e
Gdańsk, see Danzig
Gliwicki, Kanał, *canal*.....50·25n 18·20e
Górny Śląsk (Upper Silesia), *historical region*.50·30n 18·20e
Great Poland Lakeland, *physical region*..52·30n 16·55e
Kampinoski Park Narodowy, *national park*....52·20n 20·35e
Karkonoski Park Narodowy, *national park*....50·45n 15·35e
Karkonosze, *mountains*.....50·45n 15·45e
Kujawy, *physical region*.....52·45n 18·30e
Łyna, *river*.....54·15n 21·00e
Mamry, Jezioro, *lake*.....54·10n 21·28e
Masuria, *historical region*.53·40n 21·10e
Masurian Lakeland, *physical region*..53·45n 21·15e
Narew, *river*.....52·26n 20·42e
Neisse (Nysa Łużycka), *river*..52·04n 14·46e
Noteć, *river*.....52·44n 15·26e
Nysa Łużycka, see Neisse
Oder (Odra), *river*.....53·33n 14·38e
Odra, see Oder
Podhale, *physical region*.....49·50n 21·00e
Pomerania, *historic region*...53·50n 15·20e
Pomeranian Bay, see Pomorska, Zatoka
Pomeranian Lakeland, *physical region*..53·50n 16·30e
Pomorska, Zatoka, *bay*.....54·00n 14·20e
Prosna, *river*.....52·10n 17·39e
Rega, *river*.....54·10n 15·18e
Rysy, *peak*.....49·12n 20·04e
San, *river*.....50·45n 21·51e
Silesia, *historic region*...50·58n 16·53e
Slowinski Park Narodowy, *national park*....54·40n 17·25e
Śniardwy, Jezioro, *lake*....53·46n 21·59e
Sudety, *mountains*.....50·30n 16·00e
Świętokrzyskie, Góry, *mountains*.....50·57n 21·02e
Świętokrzyskie Park Narodowy, *national park*....50·55n 21·00e
Szczeciński, Zalew, *lagoon*...53·46n 14·14e
Tatra Mountains.49·14n 20·00e
Tatrzanski Park Narodowy, *national park*....49·15n 20·00e
Upper Silesia, see Górny Śląsk
Vistula see Wisła, *river*
Warta, *river*.....52·35n 14·39e
Wielkopolski Park Narodowy, *national park*....52·15n 16·50e
Wieprz, *river*.....51·34n 21·49e
Wisła (Vistula), *river*.....54·22n 18·55e
Wiślany, Zalew, *lagoon*.....54·27n 19·40e
Wisłoka, *river*....49·55n 21·26e
Wkra, *river*.....52·27n 20·44e

be clearly distinguished: a snowy winter of one to three months; an early spring of one or two months, characterized by alternating wintry and springlike conditions; a predominantly sunny spring; a warm summer with plenty of both rain and sunshine; a sunny, warm autumn; and a foggy, humid period signifying the approach of winter. Sunshine reaches its maximum over the Baltic in summer and the Carpathians in winter, and mean annual temperatures range from 46° F (8° C) in the southwestern lowlands to 43° F (6° C) in the colder northeast. The climate of the mountains is determined by altitude. The highest peaks have mean annual temperatures below 32° F (0° C), but the absolute minima (as low as —44° F [—42° C]) are found in the basins. In the distribution of mean monthly temperatures may be seen a decrease from west to east (in January, from 30° F to 23° F [—1° C to —5° C] except in summer, when the trend is from south to north (e.g., from 66° F [19° C] at the foot of the Carpathians to 63° F [17° C] on the Baltic coast). There are, however, local variations caused by topography, which is also influential in the case of annual precipitation totals.

The annual average precipitation is about 24 inches (600 millimetres), but in the mountains the figure approaches 31 to 47 inches (800 to 1,200 millimetres), dropping to about 18 inches (450 millimetres) in the Middle Polish Lowlands. In winter, snow makes up about half the total in the plains and almost all of it in the mountains.

Rivers, lakes, and soils. No less than 99.7 percent of Poland drains to the Baltic, more than half via the Wisła (Vistula) River and a third via the Odra. Polish rivers experience two periods of high water each year: in spring, the lowland rivers are swollen by melted snow, the effect intensified both by ice dams (which block the rivers for one to three months) and by the fact that the thaw first strikes the upper reaches of the northward-flowing rivers; the summer rains bring a second maximum around the beginning of July.

There are about 9,300 Polish lakes with areas of more than 2½ acres (one hectare), and their total area is about 1,200 square miles (3,200 square kilometres), or 1 percent of the national territory. The majority, however, are found in the northern glaciated belt, where they occupy more than a tenth of the surface area.

Polish soils are varied and without clearly marked regional types. The greatest area is covered by podzol and pseudopodzol types, followed by the less widely distributed brown-earth soils, which are richer in nutrients. In the south are patches of the fertile loess-based soils, especially in the Little Poland Lowlands. The rendzinas, formed on limestone rocks, are a unique type. Alluvial soils filling the river valleys and peaty swamp soils found in the lake area and in poorly drained valleys are also of importance.

Plant and animal life. The vegetation of Poland has developed only over the 10,000 years that have elapsed since the last Ice Age, and, as a result, it is not rich, consisting of some 2,250 species of seed plants, 630 mosses, 200 liverworts, 1,200 lichens, and 1,500 fungi. Holarctic elements (*i.e.*, those appertaining to the temperate belt of the Northern Hemisphere) are dominant among the seed plants.

Trees

The northeastern limits of certain trees—notably beech, fir, and the variety of oak known as pedunculate—run through Polish territory. There are few endemic species: the Polish larch (*Larix polonica*) and the Ojców birch (*Betula oycoviensis*) are two examples. Some relics of tundra vegetation have been preserved in the peat bogs and mountains. About 27 percent of the country is wooded; fundamentally, Poland lies in the zone of mixed forests, but in the southeast a fragment of the forest–steppe vegetation zone intrudes, while in the northeast there are portions of the east European subtaiga, with spruce as a characteristic component. In the mountains, the vegetation, like the climate, is distinctly altitudinal. Fir and beech woods give way to the spruce of the upper woods, which in turn fade into subalpine, alpine, and snowline vegetation. The areas with the best preserved vegetation cover are now under protection in Poland's 11 national parks and more than 500 nature reserves.

Poland's animal life belongs to the zoogeographic province known as the European–West Siberian, itself part of the Palaearctic subregion, and is closely linked with the vegetation cover. The vertebrate fauna includes nearly 400 species, of which 83 are mammals and 212 native birds. Deer and wild pig roam the woods; the elk is found in the coniferous forests of the northeast; and such steppe rodents as the brindled gopher are found in the south. Brown bear and wildcat live in the mountain woods, and the chamois and marmot are found at the highest levels. The great Puszcza Białowieska (Białowieża Forest, a national park contiguous with the Belovezhskaya Pushcha in Belorussian S.S.R.) provides shelter for about 300 wisent, or European bison, which once roamed the continent in vast numbers.

Geographic regions. The natural landscape of Poland can be divided broadly into three groups: the lowlands, the highlands, and the mountains. Further subdivision is possible by allowing for the effect of specific locations; the eastern extremes of Poland display characteristics common to eastern Europe, but the rest of the country is linked by structure, climate, and the character of its vegetation to western Europe.

The coastal plain. The Baltic Coastal Plain stretches across northern Poland from East Germany to the Soviet border, forming a low-lying region built of various sediments. The scarcely indented Baltic coastline was formed by wave action after the retreat of the ice sheet and the raising of sea levels. The Pomeranian Bay (Zatoka Pomorska) in the west and the Gulf of Gdańsk in the east are the two major inlets. In the southern portion of the former, two islands block off the Szczecin Lagoon (Zalew Szczeciński), into which the Odra discharges its waters. In the Gulf of Gdańsk, the Wisła forms a large delta. Sandbars, on which the winds have created large dunes, line much of the coast, separating the coastal lakes and lagoons from the sea. The main urban centres are the ports of Szczecin (German Stettin) on the lower Odra, and Gdańsk (Danzig) and Gdynia in the east. The central portion of the Baltic Coastal Plain is scantily peopled—there are only small fishing ports, of which Kołobrzeg is the most important—and the landscape has a desolate beauty. The regional capital is the administrative centre of Koszalin.

Coastal settlements

The lake region. The belt immediately to the south of the coastal plain presents, with its many lakes, a varied, hilly landscape of glacial origin. Wide north–south river valleys divide the region into three parts: the Pomeranian Lakeland between the lower Odra, the ancient river valley of the Warta and Noteć, and the lower Wisła; the Masurian Lakeland, east of the lower Wisła; and the Great Poland Lakeland, lowest of the three, which lies south of the Warta–Noteć river valley, between the middle Odra and the middle Wisła. The larger settlements and the main communications routes of this zone lie in and along the river valleys; the remainder of the area is mostly wooded and thinly populated. Only the eastern portion of the Great Poland Lakeland has a developed agriculture. Poznań is the main urban centre of the region; Bydgoszcz and Toruń, twin towns in a basinlike broadening of the Wisła Valley, are also important, while Olsztyn functions as the main centre of the Masurian Lakeland.

The Middle Polish Lowlands. The extensive Middle Polish Lowlands contain isolated relief features shaped by the oldest glaciations, but their character is generally flat and monotonous. The postglacial lakes have long since been filled in, and glacial outwash masks the weakly developed meltwater valley channels. The basins of the main rivers make possible a subdivision: the Silesian Lowlands lie in the upper Odra; the southern Great Poland Lowlands lie in the middle Warta basin; and the Mazovian and Podlasian lowlands lie in the middle Wisła basin. Lower Silesia and Great Poland are important agricultural areas, but large industrial centres are found in many parts of the Middle Polish Lowlands. Warsaw, the capital, situated on the middle Wisła, is most prominent, followed by Łódź, on the Wisła–Warta watershed, and Wrocław (German Breslau), on the Odra. Białystok dominates the eastern section. Rich deposits of copper have

been discovered in the ancient rocks underlying the lowlands, and mining has started north of Legnica.

The Little Poland Uplands. These uplands, south of the Middle Polish Lowlands, extend from east to west, but they are folded transversely. In the west is the Silesian–Cracovian upthrust, with rich deposits of coal. A second upthrust is formed by the ancient rocks of the Holy Cross Mountains (Świętokrzyskie Góry), which reach a maximum height of 2,008 feet (612 metres). Between these two regions lies the Nida Basin, with an average height of 650 to 1,000 feet. East of the Holy Cross Mountains, the uplands are cut by the valley of the Wisła, beyond which lie the Lublin Uplands. Climatically, this region is somewhat cooler (by 1° or 2° C) and more humid than the surrounding lowlands and basins. In the south occur patches of loess on which fertile brown- and black-earth soils have developed. The older geological regions contain valuable minerals: in the Silesian–Cracovian uplands there are—apart from coal—iron, zinc, and lead deposits, while the Holy Cross Mountains supply more ores and raw materials for building. These mineral resources have made possible the rise of Poland's most important industrial region, and the landscape of Upper Silesia is highly urbanized. Ten closely linked cities in Katowice province have populations ranging between 100,000 and 350,000 and form the core of the Upper Silesian metropolitan conurbation with a total population of almost 3,500,000, Katowice being the largest centre. Closely linked with this region is that around Kraków, which contains the Nowa Huta steelworks, Poland's largest, based on Silesian coal and on iron ore imported from the Soviet Union. The old Polish industrial region was the area around the Holy Cross Mountains, where the not very rich local iron ore was exploited in the 19th century; Kielce is the urban centre. Lublin is the main centre of the uplands that bear its name, under which substantial deposits of coal were discovered in the early 1970s.

The industrial region

The air and water of the Upper Silesian industrial district are seriously polluted. Efforts at control on the part of the authorities have met with but limited success.

The Sudety. The Sudety (German and conventional English Sudeten) and the Sudety Foreland, part of the larger Bohemian Massif, have a long and complex geological history. They owe their present rugged form, however, to earth movements that accompanied the Carpathian uplift, and the highest portion, the Giant Mountains (Karkonosze; German Riesengebirge, Czech Krkonoše), reach a height of 5,256 feet (1,602 metres) above sea level. The region contains rich mineral deposits, notably coking coal, which has occasioned the growth of an industrial centre around Wałbrzych. The whole region is densely populated and crossed by railway lines; resorts and spas are found here and there in more secluded regions. The Sudety Foreland, similar in geological structure to the Sudety but lower in relief, is separated by a large fault from the larger mass. The region contains many granite quarries.

The Carpathians. The southernmost and also most scenic portion of Poland embraces the Carpathian Mountains and their associated basins, created in the geologically recent mountain-building Tertiary Period. Within the Polish frontiers lie the Oświęcim (German Auschwitz) and Sandomierz basins; a portion of the Outer Western Carpathians (Beskidy); and the Orawa–Podhale Basin and the Tatry (Tatra Mountains), associated with the Central Western Carpathians. The sub-Carpathian basins contain deposits of salt, sulfur, natural gas, and some petroleum. There is a large rural population and many towns of medium size. The nearby mountain regions are readily accessible and also densely peopled. The highest peak of the Outer Western Carpathians, Babia Góra, reaches 5,659 feet (1,725 metres); the small but high Tatry, the highest portion of the Polish Carpathians, with a maximum elevation of 8,199 feet (2,499 metres), has a unique character. Zakopane, the largest tourist and resort centre in Poland, lies at their feet; indeed, the entire Carpathian region attracts visitors from the industrial cities to the north and the lowlands beyond.

Rural settlement. In contemporary rural Poland, survivals of a medieval settlement pattern can be found adjacent to forms originating in the 19th and 20th centuries. The first group includes compact villages with an average of 20 to 100 dwellings, associated with a nodal road system; this type is especially typical of the west and north of the country. Compact, elongated villages also occur in the northeast, having originated in economic reforms of the 16th century. In the Carpathians, the Sandomierz Basin, and the Lublin Uplands, an elongated form of settlement along river valleys is common, with roads running at right angles to the main line of buildings and linking them to the fields.

Older villages

A newer, more dispersed type of village is found only where natural conditions are less favourable—in major river valleys, in swamps and forests, and in some parts of the mountains. A final type of settlement is of more complex structure, generally compact but with some dispersion. It results from the transformation of old villages after the concentration of landholdings in long-cultivated regions and is found in the central parts of the country and also in the Masurian and Pomeranian lakelands.

Urban settlement. The earliest forms of urban settlement in Poland were linked with the disappearance of tribal organization and the development of the feudal system. Settlements of artisans and merchants were already developed in the early Middle Ages, clustering around centres of royal authority or large feudal estates, either of which could provide a stronghold against attack. In the 13th and 14th centuries, settlements of this type received legal rights and privileges and at the same time acquired a standard urban layout, with a central square and streets at right angles. In time, a settlement might be joined by neighbouring or rival towns, and the town grew, changing its character in later centuries according to changing styles. In spite of the changes, however, medieval town centres—with Gothic churches, town halls, and fragments of walls and defensive gates—were preserved until well into the 20th century in many Polish towns and cities, including Warsaw, Kraków, and Toruń. In the Middle Ages there were many towns, of much the same size and located at fairly equal distances from one another, in the southwestern part of the country. After the development of better transportation and the growth of industrial centres, however, many of them lost their function and became villages. The growth of capitalism, and the industrialization associated with it, broke the pattern and caused much greater variation in the distribution and size of towns. By the 1970s, Poland had 30 towns with populations of more than 100,000; well over 100 towns with populations between 20,000 and 100,000; and about 750 smaller towns. The number of urban dwellers has more than doubled in the 20th century, approaching 56 percent of the national population by 1976.

Effects of industrialization on settlement

The majority of the Polish cities have complex functions: Warsaw, for example, is a large centre of diversified industrial activity, but it takes its character from its function as the capital and from its large concentration of academic and cultural institutions. Port functions are of fundamental importance for Gdańsk, Gdynia, and Szczecin, while in the towns of the Upper Silesian industrial region, and in Wałbrzych and Konin, the extractive metallurgical and heavy engineering industries are dominant. In Częstochowa and Kraków metallurgy is important (although Kraków is also second only to Warsaw as a cultural centre), Poznań concentrates on engineering, and the textile industry predominates in Łódź and Białystok. In smaller urban centres, service functions—notably tourism and recreation—and local transportation are important elements in the economy. Almost all towns, however, have at least some manufacturing or extractive industry.

THE PEOPLE

Composition. All but 2 percent of the population of Poland is of Polish nationality. By the 1970s, the most numerous minorities were Belorussians (165,000), located in the eastern part of the country, and Ukrainians (180,000), scattered in the northern provinces. There

Poland, Area and Population

	area		population	
	sq mi	sq km	1970 estimate	1976 estimate
Provinces (*województwa*)				
Biała Podlaska	2,066	5,353	280,000	281,000
Białystok	3,880	10,053	597,000	622,000
Bielsko-Biała	1,430	3,703	735,000	789,000
Bydgoszcz	3,996	10,352	940,000	1,006,000
Chełm	1,492	3,867	220,000	223,000
Ciechanów	2,456	6,362	396,000	399,000
Częstochowa	2,390	6,189	713,000	731,000
Elbląg	2,356	6,103	411,000	426,000
Gdańsk	2,853	7,388	1,138,000	1,276,000
Gorzów Wielkopolski	3,281	8,498	412,000	439,000
Jelenia Góra	1,691	4,380	475,000	489,000
Kalisz	2,514	6,512	626,000	648,000
Katowice	2,567	6,649	3,245,000	3,535,000
Kielce	3,556	9,210	1,014,000	1,044,000
Konin	1,984	5,139	416,000	427,000
Koszalin	3,271	8,471	404,000	440,000
Kraków	1,256	3,254	1,045,000	1,137,000
Krosnio	2,201	5,701	405,000	426,000
Legnica	1,558	4,036	367,000	424,000
Leszno	1,603	4,153	334,000	345,000
Łódź	587	1,520	1,023,000	1,093,000
Łomża	2,590	6,709	325,000	320,000
Lublin	2,622	6,792	842,000	896,000
Nowy Sącz	2,153	5,576	579,000	604,000
Olsztyn	4,760	12,329	628,000	671,000
Opole	3,295	8,535	929,000	977,000
Ostrołęka	2,499	6,472	359,000	362,000
Piła	3,168	8,205	403,000	421,000
Piotrków Trybunalski	2,417	6,261	578,000	584,000
Płock	1,975	5,114	467,000	484,000
Poznań	3,147	8,152	1,103,000	1,187,000
Przemyśl	1,710	4,430	366,000	376,000
Radom	2,816	7,294	664,000	682,000
Rzeszów	1,698	4,399	579,000	618,000
Siedlce	3,280	8,495	601,000	602,000
Sieradz	1,880	4,870	388,000	387,000
Skierniewice	1,528	3,957	384,000	390,000
Słupsk	2,878	7,453	335,000	361,000
Suwałki	4,050	10,490	401,000	417,000
Szczecin	3,853	9,980	789,000	867,000
Tarnobrzeg	2,425	6,282	520,000	539,000
Tarnów	1,603	4,151	561,000	585,000
Toruń	2,064	5,345	557,000	594,000
Wałbrzych	1,609	4,167	702,000	717,000
Warszawa	1,464	3,794	1,998,000	2,192,000
Włocławek	1,700	4,404	396,000	405,000
Wrocław	2,428	6,289	970,000	1,038,000
Zamość	2,697	6,986	475,000	471,000
Zielona Góra	3,418	8,853	553,000	586,000
Total Poland	120,725	312,677	32,658,000	34,528,000*

Source: Official government figures. *Detail does not add to total given because of rounding.

were some tens of thousands of Jews, mostly concentrated in urban areas; they represented but a tiny remnant of the former Jewish population of more than 3,000,000, most of whom perished during the Nazi occupation (1939–44). About 10,000 Lithuanians live near the border of the Lithuanian Soviet Socialist Republic. After the great migratory movements that took place toward the end of, and immediately after, World War II, there remained only a small number of Germans in Poland. Their exact number is a matter of dispute between the Polish and West German authorities. A similar situation occurs in some villages on the Czechoslovak border. Finally, entirely non-indigenous inhabitants of Poland include about 20,000 Russians, about 10,000 Greek and Macedonian immigrants, and about 12,000 Gypsies.

Ethnic origins. The Polish people originated in several West Slavic tribes inhabiting the basins of the Odra and Wisła rivers. What used to be tribal territories have retained to this day a certain diversity in dialects spoken, and indeed in folk culture generally, although these ancient traces are disappearing as a result of internal migration, urbanization, and industrialization. Five large tribal territories may be distinguished: (1) Great Poland (Wielkopolska), lying in the basin of the middle and lower Warta as far as the middle Odra and the cradle of the Polish nation, whose origin lay in the Polanian tribe (Polanie; Latin name Poloni for the tribe, Polonia Major for their territory); (2) Silesia (Śląsk), the homeland of the Silesians (Ślęzanie) and a few smaller groups inhabiting the basin of the upper and middle Odra; (3) Little Poland (Małopolska, Latin Polonia Minor), in the basin of the upper Wisła, covering the tribal territories of the Vistulans (Wiślanie); (4) Mazovia (Mazowsze), around the middle Wisła and its tributaries; and (5) Pomerania (Pomorze), stretching in a wide coastal belt from the mouth of the Odra to the Wisła, separated by the Warta–Noteć valley from Greater Poland.

Tribal territories

As a result of the expansion of settlement by these ethnic groups, further territories were annexed. East of the lower Wisła, in a coastal and lakeland belt, lived a non-Slav people—the Prussians—belonging to the Baltic tribes and related to the Lithuanians and Latvians. On these lands in the 13th century arose the German state of the Knights of the Teutonic Order, having as its purpose the conversion of the pagan Prussians. The Knights brought about the extermination or Germanization of the Prussians and conflict with Poland. After long and bloody wars, a part of this region was joined to Poland, and the remnant became the Duchy of Prussia, enfeoffed to the Polish kings. The frontier lands of Prussia were colonized by Polish peoples from Mazovia, but they became converted to Protestantism, the prevailing religion in Prussia, in contrast to the situation in predominantly Catholic Poland. In this way there arose a group of Prussian Mazurians, who retained their language until the 20th century but gradually became Germanized. Some of them still inhabit the Masurian Lakeland.

The Prussian Mazurians

The eastern part of contemporary Poland, centred around the basin of the Neman and the upper Narew and partially the Bug River, does not form a part of ethnic Poland; it formerly belonged to the Grand Duchy of Lithuania and was inhabited by Belorussian tribes. These lands were settled later by colonists from Mazovia but still contain a substantial Belorussian element.

The western parts of Silesia, Great Poland, and Pomerania gradually became Germanized from medieval times onward, this being carried out intensively in the Prussian kingdom from the middle of the 18th century, especially after the fall of the former Polish state. It was not until the end of World War II that a reversal of this trend took place. The Potsdam Conference (1945) approved the evacuation of the remainder of the German population to west of the Odra–Nysa (Oder–Neisse) Line, and it was replaced by Poles repatriated from the Soviet Union and other countries, as well as by a spontaneous migration from the central provinces. The western region thus developed over the ensuing years a completely new society, most of whose peoples are not directly descended from the original Polish inhabitants. The western and northern third of the country contain a fast-growing population that is more than a quarter of the national total.

Anthropologically, the Polish population is mixed, transitional between the north and south as well as the east and west of Europe. Poles are generally blonder and have lighter coloured eyes than do the peoples of southern Europe, but they are darker than the Scandinavians to the north.

Language. Linguistically Poles, together with Czechs, Slovaks, and Sorbs (Wends, or Lusatians), belong to the West Slavic group, and the literary Polish language developed on the basis of Great and Little Polish dialects. The first records of Polish place and personal names appear in 12th-century church documents; numerous Polish texts of hymns and sermons, psalters, and lawcourt records have been preserved from the 14th and 15th centuries. The literary language blossomed in the 16th century and was linked with the spread of Humanism, the invention of printing, and the conflict of the Reformation. A second period of the development of classical and scientific literature took place in the second half of the 18th century. Despite the fall of the Polish state, Polish culture continued to develop under foreign domination, and the language and the classical literature were, in fact, factors uniting the whole nation (see also below *Cultural life and institutions*). Despite the standardization of the language—by universal education, literature, and, more recently, the mass media—country dialects, reflecting former tribal patterns, remain of significance in the con-

Development of the written language

temporary population. They include Great Polish, Silesian, Little Polish (further subdivided into a mountain dialect), and Mazovian; the very distinctive Kashubian is commonly classified as a separate language.

Dominance of Catholicism

Religion. The overwhelming majority of the Polish population is, and traditionally has been, Roman Catholic. Other Christian denominations and Judaism have limited influence. Among Catholics, the greatest numbers —including Belorussians, Ukrainians, and Russians— adhere to churches with a measure of episcopal independence. An increasing percentage of the population does not practice any religion, or describes itself as areligious, but their numbers are difficult to ascertain as official statistics do not record religious affiliation. In November 1976 the Polish Catholic bishops issued a pastoral letter that challenged the claims of the government to religious toleration and charged that the church was being subjected to attack in an official conspiracy to enforce atheism.

Social characteristics. The social structure of the population underwent great changes after World War II. After 1944, the class of landowners, developed from the former nobility, disappeared with the dissolution of the landed estates, as did the urban bourgeoisie as a result of the nationalization of industry and trade. Social and political emphasis shifted to the working class. Many of the peasants, very numerous before the war, moved to the towns, so that those engaged in agriculture (together with their families) formed only 27 percent of the population by 1975. The character of the intelligentsia has also changed. Before the war it consisted mainly of property owners, whereas in contemporary Poland its membership consists largely of persons with peasant or working class backgrounds, this process being facilitated by educational opportunity. The number of white-collar workers in Poland has more than tripled since prewar days. Finally, as a result of the survival of private businesses, a lower middle class element is still significant in the population, although its size has been considerably decreased.

These trends are seen by Polish analysts as among the major components of the demographic and social effect on postwar economic development. Such effects, however, are produced not only by alteration of economic or social classes but also by the increasing participation of women in the economic life of the country. In addition to the increased supply of women available to the labour pool, produced by the postwar baby boom (between 1950 and 1970, the participation of women between the ages of 15 and 59 in the Polish labour force rose from 67 to 77 percent), the women themselves are increasingly well-educated, career-oriented, and motivated by family responsibilities.

The beneficial effects of the social trends mentioned above inevitably have side effects that can be expected to further exacerbate problems already created by these momentous changes. One of the most serious of these is the aging of the agricultural labour force. Most of the population migrating to the cities for the industrial and white-collar jobs there have been individuals and couples of procreating age; the result has been a rise in the proportion of agricultural workers over the age of 45 from 36 percent in 1950 to 54 in 1970. Before this trend can be reversed, or at least abated, the gap between the social conditions of employment in the cities and rural areas must narrow appreciably.

Aging agricultural workers

Another side effect of these social changes appears in the fertility differentials among women of childbearing age. As elsewhere in the world, fertility appears to vary according to level of education; for example, married women in Poland between the ages of 20 and 49 with higher education bear, on the average, 1.71 children, while those with a secondary education bear 2.01 children, and those with an incomplete elementary education bear 3.4 children. These patterns suggest that, in the future, the growth of Poland's population can be expected to slow as levels of education rise, women's social and economic roles alter, and information about birth control is more broadly disseminated.

Polish socio-economic planners hope to arrange the economic life of the country so as to afford maximum scope to members of Polish society throughout the whole of their economic life—for example, by providing an apartment to every household by 1990 and by raising the proportion of the labour force in the services sector and in manufacturing, particularly of consumer goods, to about 40 percent. Since these goals imply a decline in primary sectors of the economy (agriculture, mining, heavy industry) and since labour is already in short supply in certain sectors, planners are concentrating on improvements in mechanization of these primary activities, on modernization of existing plant, and on raising productivity of workers in these sectors.

Demographic trends. The population of Poland underwent great changes as a result of World War II. In 1939, 34,800,000 people lived within the then boundaries of Poland—a number reduced to 23,900,000 within the new frontiers in 1946. The decrease was caused primarily by war losses, calculated at 6,000,000 persons; only 644,000 of these persons were killed as a result of direct war activity, the remainder falling before the extermination policies of the Nazis. Postwar migrations and the associated change of boundaries were also important; from 1945 to 1949, the repatriation of Poles from the Soviet Union involved about 1,500,000 people; of Germans to the West, about 2,000,000 people (not counting the millions of people evacuated during wartime); and of Ukrainians, Belorussians, and Lithuanians to the Soviet Union, about 500,000 people. In the same period, the return of former emigrants and combatants from western Europe also helped swell the population.

Following these mass population movements, the 1950 census showed a total population of 25,000,000, the 1960 census 29,800,000, and the 1970 census 32,600,000. At the end of 1975 it was estimated to be 34,200,000. The total population thus increased by 10,300,000 over 30 years but still was less than it had been in 1939. The increase was not the same in all parts of the country and was not, in terms of time, steady; there was, for example, a rapid increase in the lands annexed to Poland as a consequence of the Potsdam Conference. This was attributable not only to the influx from outside but also to the large natural increase among the generally young people who were overwhelmingly in the majority among the newcomers. At the same time, there has been a nationwide movement of population from the country to the towns, linked with the rebuilding of the latter and the associated development of industry. Urbanization is especially marked in the west and north. The national urban population increased from 39 percent to about 56 percent between 1950 and 1975.

Population movements

The rate of annual natural increase in the first decade after the war was among the highest in Europe, approaching 20 percent, and resulted from a decrease in mortality as well as from an increase in births. After reaching a maximum in 1955, the rate of natural increase began to fall. This was caused by a decline in the birth rate occasioned by the spread of birth control and by the coming of age of the diminished age group born during the war. The rate of natural increase fell to 10 percent by 1965 and as low as 8.5 percent in the early 1970s but rose again slightly after 1971. The difference between the rates of natural growth in towns and the countryside is slowly declining, though the rate in the towns is still lower than that in the country. The large number of births in 1948–56 has created special demands on national education and on the labour market.

War losses and the abnormal course of natural population movements have caused other anomalies to occur in the sex and age structure, notably in the preponderance of women (especially widows) in certain older age groups. The age structure also shows great regional differences. The traditional heartland of Poland has witnessed a steady increase in percentage of older persons; in the west and north the percentage of older people is smaller, and that of the young greater, than elsewhere. The high natural increase of the postwar years has given the Poland of the mid-1970s a population that is relatively much younger than that of other northern and western Euro-

pean nations, with a pronounced excess of persons between 15 and 30. Demographic forecasts indicate that by 1980 the population of Poland will exceed 37,000,000 and that the age structure will undergo an important change through the decrease in the number of young people and an increase in the older population. (J.A.K./A.H.D.)

THE NATIONAL ECONOMY

The Poland of the 1970s is eastern Europe's largest producer of agricultural crops after the Soviet Union, as well as the world's second largest grower of potatoes and rye and its fifth most important source of sugar beets. In industry it ranks among the world's top 15 countries. It is Europe's second largest source of bituminous coal and zinc. Although Polish industry is not so well developed as that of Czechoslovakia or East Germany, the country is eastern Europe's largest producer of several industrial products (after the Soviet Union), notably steel, cement, tractors, railway freight cars, and trucks. It is not, however, a very important foreign trader; although its population was the sixth largest in Europe at the start of the 1970s, its exports were only in 10th place. After 1951 Poland's national income grew at a rate that was some 30 percent faster than the average for the world as a whole. Available statistical data do not permit detailed comparisons, but it is estimated that, in spite of this growth, national income per person is still well below that of the most developed east European and most advanced non-Communist countries.

Growth of national income

Resources. *Mineral resources.* Poland is relatively well endowed with natural resources. Its principal mineral asset is bituminous coal, and by 1975 it was the world's fourth largest producer. Almost all the output—which increased by one-third over the 1960s—is derived from the rich Upper Silesian coalfield; the remainder comes from Nowa Ruda and Wałbrzych, in Lower Silesia. In the 1970s new seams were discovered and a pilot mine was opened near Zamość. Brown coal and lignite are derived from the Konin–Turek area and from Turoszów in the southwest; development of huge deposits in the Bełchatów region also began in the 1970s. Other fuel resources include small amounts of oil—found between Bochnia and Sanok and at Krosno in the south—and large deposits of natural gas in the vicinity of Lubaczów, Nowa Sól, Trzebnica, Ostrów Wielkopolski, and Zamość. Output of crude oil more than doubled over the 1960s, but it accounts for only a fraction of total requirements. Natural gas production is substantial.

Sulfur is Poland's second most important mineral. Reserves, thought to be among the largest in the world, are located on both sides of the Vistula near Tarnobrzeg. Other important nonmetallic minerals include salt, kaolin, limestone, chalk, gypsum, and marble. Scattered rock salt reserves are estimated to total some 43,000,000,000 tons. Potassium and magnesium salts are found in some areas bordering the Baltic, around Kłodawa, and in the Sudety. Reserves are thought to total some 500,000,000,000 tons, but commercial exploitation has not yet begun. Kaolin is present in the Bolesławiec district of Lower Silesia, at Czerwona Woda, and in the Świdnica region.

Metals

Among metallic minerals, zinc is the most important. Reserves of zinc and lead ores, believed to be among the largest in the world, are located around Bytom and in the Zawiercie–Olkusz region, and by the mid-1970s Poland was Europe's second largest producer of zinc ore. There are also large quantities of copper, the two main producing areas being near Legnica and Głogów and near Bolesławiec and Złotoryja. Iron ore is mined in small quantities around Częstochowa and Kraków and near Kielce and Łęczyca, and there are also deposits at Suwałki, near the Soviet border. Nickel is found near Ząbkowice, and such other metals as vanadium, cobalt, and silver are also produced.

Other resources. The quality of the soil varies, but there are large fertile areas, including Lower Silesia and the Little Poland Lowlands, the Kujawy, the Wisła delta, and the Lublin area. The soil is somewhat poorer in large parts of central and northern Poland. About 27 percent of the country's total land area is wooded, and some 80 percent of this is occupied by coniferous timbers, with pine, larch, and spruce the most important for resource purposes. The bulk of the country's hydroelectricity comes from the Carpathians, the Sudety region, and the Brda and Wisła rivers; potential was limited in the 1970s, but more hydroelectric power stations were planned.

Sources of national income. *Agriculture.* By the mid-1970s, agriculture produced less than 10 percent of the national income and employed less than 30 percent of the economically active population. Polish agriculture is unique in eastern Europe in that private farms still account for the greater part—80 percent, in fact—of its total output. The 6,700 state farms hold only 22 percent of the land and are responsible for only about 19 percent of agricultural production; cooperatives (*i.e.*, collective farms) cultivate less than 2 percent of the land.

In the early 1950s Polish agriculture was adversely affected by inadequate investment, but the 1960s saw a gradual growth in the share of total investment spending devoted to this sector. Farm production went up by a quarter over the 1960s; this growth—which would have been faster but for the unusually adverse weather conditions of 1969—was mainly in the form of an increase in yields, made possible by the greater use of fertilizers, other agricultural chemicals, and machinery. In spite of this, yields tended to be well below the figures reached in East Germany and in Czechoslovakia, where farming is the most advanced in the Communist bloc. Polish agriculture is unable to satisfy internal demand for a number of important products—especially cereals and fodder crops—and large imports are necessary. The bulk of the country's 21,000,000 acres (8,600,000 hectares) of forest is state owned; most of the substantial log production consists of coniferous timber.

Manufacturing. By the mid-1970s, industry was responsible for just over half of Poland's net material product and employed more than 5,000,000 persons, or about 30 percent of the country's economically active population. As a result of heavy capital spending in the 1960s (more than 40 percent of all investment expenditures over the decade went to industry), growth in this sector has been very rapid. Industrial production went up by more than 100 percent during the period, bringing it to a level more than seven times higher than that of 1950. Principal emphasis was on the capital goods industry, which increased its output by 130 percent; production of consumer goods, however, advanced by only 75 percent. By the mid-1970s electrical engineering, machinery building, and transport equipment accounted for more than 30 percent of gross industrial production, compared with 15 percent in 1960. The food industry held second place (about 16 percent), although it grew only about 40 percent over the 1960s. The chemical, textile, fuel, basic metals, and metal-fabricating industries are also significant.

The role of heavy industry

Though there has been a shift of agricultural labour into industry, a large part of the gain in industrial production over the 1960s was the result of higher productivity, which raised output per person by 53 percent. Poland is eastern Europe's principal producer of a number of industrial goods, including steel, mineral fertilizers, sulfuric acid, cement, tractors, railway freight cars, and motor trucks.

By the mid-1970s there were some 174,000 industrial enterprises in Poland, of which 119,000 were private handicraft establishments. Of the remainder, cooperative firms accounted for about two-thirds, though half of these employed fewer than 11 workers. Large state-owned enterprises (with more than 1,000 employees) accounted for less than 2 percent of the total number of enterprises but provided employment for almost all of the industrial labour force. Private, nonhandicraft industrial establishments employed an average of only three workers per firm.

Energy. Over the 1960s and '70s, electricity output increased rapidly to about 97,000,000,000 kilowatt-hours in 1975, most of them generated by public supply plants. All but a small portion of this total was derived from bituminous coal and lignite; the share of liquid fuels was

less than 4 percent and that of hydroelectricity only about 2 percent. There are long-term plans for building a giant nuclear-power station in northern Poland. By the early 1970s natural gas had largely replaced manufactured gas. Crude oil production contributes less than 5 percent to the annual consumption of petroleum, most of which has to be imported.

The financial sector. The banking system is state owned and is controlled by the National Bank of Poland (Narodowy Bank Polski). This body fulfills all the usual functions of a central bank, including issuing notes; it also controls credit, exercises financial control over state enterprises, regulates the money supply, and executes the government's financial policy. Other important banks include the Commercial Bank of Warsaw (Bank Handlowy w Warszawie), which carries out foreign trade transactions; the Food Management Bank (Bank Gospodarki Żywnościowej), which distributes investment funds for agriculture and food production; and the Polish General Savings Bank (Bank Polska Kasa Opieki). There are no stock exchanges. The rate of exchange for certain transactions is very much higher than the official foreign exchange value.

Foreign trade. Although Poland has about a third of eastern Europe's population, it accounts for only slightly more than a quarter of its foreign trade. Over the 1960s Polish exports grew by almost 140 percent, while imports advanced by 115 percent. Imports exceeded exports by a considerable margin over the decade, but the last few years saw some reduction in the deficit. By the mid-1970s Poland's foreign trade with the Communist bloc had declined to less than half of total trade, while trade with the developed non-Communist countries was responsible for well over a third and was growing. The country's principal trading partner (though declining in importance) is the Soviet Union, which provides almost 25 percent of all Polish imports and takes about 30 percent of the exports. East Germany, West Germany, and Czechoslovakia, in that order, are the other important trading partners, while all the Western countries account for about a fifth of the total trade.

Balance of payments

By the mid-1970s more than 40 percent of Poland's exports consisted of machinery and equipment, with complete plants, power and electrical machinery, and vessels the most important items. Many of these go to Communist and developing countries. Fuels and raw materials account for a further quarter of the total, with coal, coke, iron, and steel figuring prominently. Consumer goods and food make up the remainder. Almost half of Poland's imports are machinery, 5 percent fuels. Poland is a net exporter of consumer goods, food, and fuels.

Management of the economy. *The private sector and the role of the government.* The state plays a dominant role in the economy, although its involvement is not so extensive as it is in other countries of the Communist bloc. The bulk of industry is state owned and controlled, with nationalized and cooperative enterprises accounting for all but a tiny fraction of gross industrial production and employment of the industrial labour force. Banking, foreign trade, insurance, and much internal commerce are also controlled by the state. Private enterprise, however, plays an important role in handicrafts and accounts for about 20 percent of the net output of the building trades and 80 percent of final agricultural production. In consequence, the state sector of the economy produces 86 percent of the national income—a figure high by Western standards but below that reached in other Communist countries. Since the state lays down detailed economic policies and regulates most prices, however, private enterprises are under strict control and operate within closely defined limits.

Taxation. Tax receipts amounted to 85 percent of total state revenue by the early 1970s. More than two-thirds of this was accounted for by turnover taxes—a fixed percentage of the sales value levied on consumer goods—while a further 20 percent was derived from profits and income taxes paid by enterprises. Taxes collected from the population—mainly by way of a payroll tax—made up the remainder.

Trade unions. Polish trade unions are, on the whole, organized on industrial lines. By the 1970s there were almost two dozen unions, with a total membership exceeding 11,000,000 and including more than 90 percent of the labour force (excluding those engaged in private farming). The unions do not seem to have given any support to such actions as the Baltic workers' demands for lower prices and higher wages during serious disturbances late in 1970 and are regarded by Western experts as an extension of the state apparatus.

Contemporary economic policies. Although the government did not persist with its early attempts to collectivize agriculture, it followed, over the 1950s and '60s, the policy of detailed regulation and centralization then prevailing in the Communist bloc. This led to a spectacular growth in industry, but only at the cost of serious structural difficulties, high production costs, low productivity, and an irrational pattern of production and consumption. These problems became increasingly obvious during the 1960s and forced the government to evolve a program of economic reform.

Economic difficulties

The main features of this program, which came into effect in early 1971, were somewhat greater managerial independence and stronger emphasis on profit, a system of credit and investment finance that took into account the performance of enterprises, more material incentives, and a revision of prices. The price revision (which involved a rise in food prices and a cut in consumer goods prices) was followed by widespread disturbances and strikes in some Baltic towns. These led to changes in the national leadership, increases in wages for the lower paid, a two-year freeze on food prices, and a pledge that the 1971–75 plan would be revised. Further disturbances in June 1976, when another attempt was made to raise food prices, eventually obliged the authorities to withdraw the proposed increases.

Economic problems and prospects. Many Western experts believed that the new government would retain the essential features of the reform package. This process, however, although likely to go some way toward solving the problems described above, cannot be expected to bring about any noticeable short-term improvement in the economic situation. On the contrary, since the reform program involves the pursuit of new, untried economic policies aimed at increasing competition, channelling investment funds into more profitable areas, weeding out inefficient producers, and increasing individual effort, it may well lead to some transitional difficulties. These could include unemployment, as enterprises become more conscious of cost and profit; the decline of certain inefficient industrial sectors; and other problems that could have an unfavourable effect on short-term growth in several industrial areas. In the long run, however, a policy of reform is likely to bring the benefits of greater industrial specialization—whether within a Comecon framework or not—greater efficiency, a better appreciation of the needs of the domestic and foreign markets, and a faster growth in sales to (and purchases from) the advanced non-Communist world. The export of fuel, especially coal, and the import of factories and advanced technology from western Europe, Japan, and the U.S. increased in the early and mid-1970s. (E.I.U./A.H.D.)

TRANSPORTATION

The communications system in Poland developed in the 19th and early 20th centuries, in a period when the country was divided among Russia, Germany, and Austria. The three areas thus developed in different economic and political conditions, and the main railway lines were oriented to the main towns of those nations. This characteristic was especially marked in the section that was under German rule, where the main junction was Berlin. In Austria-Hungary there were two such centres, Vienna and Budapest. The density of the railway network in the Russian sector was much less than in the Austrian, let alone the German sector, although the main junction of the western part of Russia was at Warsaw. In 1918 independent Poland thus took over a railway system that was inadequate for its needs. Among the most important rail-

way lines built after that date were those linking Warsaw with Poznań and Kraków and a coal trunk line linking Upper Silesia with the newly built seaport of Gdynia.

Railways. After the devastation of World War II the railway system was reconstructed and rebuilt. The main junctions, in particular, were enlarged, as were the reloading stations on the Soviet border, where the track gauge changes. Electrification of the most used lines also took place, and on other routes diesel engines gradually replaced steam. The electrified routes focus primarily on the Upper Silesian industrial region and include the line to Poznań via Wrocław; the coal trunk line to the ports of Gdańsk and Gdynia; lines to Warsaw; and lines to the Soviet border via Kraków and to the Czechoslovak border in the direction of Prague. By 1975 a quarter of the system had been electrified, and work was in progress to electrify more lines. A second railway line was planned between Upper Silesia and the Baltic ports.

Electrification

The railways are administered by the Polish State Railways (Polskie Koleje Państwowe). The average track density is 13.7 miles per 100 square miles (8.5 kilometres per 100 square kilometres), but there are great regional differences; Lublin province, for example, has less than half the average. Narrow-gauge railways, about 11 percent of the total in the mid-1970s, are being replaced.

In the amount of freight transported, Polish railways occupy second place in Europe (after the Soviet Union); in passenger transport, Polish railways occupy third place, after the Soviet Union and West Germany. Because of the position of the country, Polish railway lines are important in the carriage of transit freight between Comecon countries, notably between the Soviet Union and East Germany and between Czechoslovakia and the Polish ports.

Highways. The highway system originally showed disproportions similar to those of the railways; that is, the densest network was on lands belonging to Germany and the least dense on lands belonging to Russia. An attempt to remedy this situation was made between 1918 and 1938 and again, but more intensively, after 1945. Between 1945 and 1970 about 23,000 miles (37,000 kilometres) of new roads were built and the surfaces of about 35,000 miles were modernized, making possible the establishment of bus routes to almost all village communities. Modern multi-lane highways adapted to mass traffic requirements are being built in Warsaw and between that city and Kraków and Poznań. In 1945 the Polish state inherited the *Autobahn* running from the East German border to Wrocław.

Autobahn

About half of the Polish highways have improved surfaces, and the roads are more evenly distributed than the railways are. Road transport is not centralized to the same degree as rail transport since not only the National Road Transport Service (Państwowa Komunikacja Samochodowa, or PKS) but also various smaller economic bodies and branch transport concerns—notably in building, forestry, and mining—are involved in highway transport. Although roads in the mid-1970s handled about four times as much freight as railways, PKS had just under 10 percent of the total. PKS conveyed two-thirds of all passengers, but if the measure of miles per passenger is used, then PKS and the railways contribute equally to the transport service. Poland is behind most European countries in numbers of cars, though it had more than 1,000,000 in 1976; the number has been doubling every five years since 1960.

Inland waterways and merchant shipping. The middle course of the main Polish river, the Wisła (Vistula), is still not regulated, and thus the Wisła is of less importance as a waterway than the smaller Odra, linked by the modern Kanał Gliwicki to the Upper Silesian industrial region and carrying coal to the port of Szczecin. The Odra basin is also linked to the lower Wisła by the Kanał Bydgoski. Inland navigation is of little importance in Poland, however, less than 1 percent of Polish freight being carried on rivers and canals.

Shipping is well developed, and there are three large seaports—Szczecin (the leader), Gdynia, and Gdańsk—as well as smaller fishing and coastal navigation ports. The Polish merchant fleet has nearly 700 ships, with a deadweight tonnage of more than 4,000,000. The export of coal and coke accounts for more than 40 percent of the total shipping. The passenger liner "Stefan Batory" is internationally known for service on the Gdynia–Montreal route.

Air transport. Domestic and international air transport is provided by a state enterprise, LOT (from Polskie Linie Lotnicze). There are some two dozen international routes, centred on the airport at Warsaw.

ADMINISTRATION AND SOCIAL CONDITIONS

The Polish constitution of 1952 (as amended) reflects the changes that occurred after World War II following the collapse of the prewar government, the formation of radical underground political organizations during the German occupation, the liberation of the country by the army of the Soviet Union, and the taking of power by left-wing forces following the parliamentary elections of 1947. Until 1952 the legal basis of the system was the constitution of 1921, partially modified in 1947.

The national constitutional framework. The chief bodies reflecting the forms of authority laid down in the constitution are the one-house parliament (the Sejm) and the Council of State, which is chosen by the Sejm and fulfills the traditional functions of the head of state.

The Sejm consists of 460 members elected for four-year terms. Its sessions are short, and votes tend to be unanimous. The Council of State not only passes legislation during the long intervals between sessions of the Sejm but also has the power to issue decrees with the force of law. In the judicial sphere, it has the power of binding interpretation of laws. Chosen from members of the Sejm, the council consists of a chairman, four deputies, a secretary, and 11 members. It represents the country abroad, nominates and recalls ambassadors, receives the credentials of representatives of foreign countries, and ratifies international treaties.

The role of the Sejm

The Council of Ministers, which is the supreme executive and administrative organ of the Sejm, prepares the budget and formulates the economic plan and supervises the execution of both. It exercises general management of national defense and of the armed forces. The Supreme Audit Office, independent of the Council of Ministers, oversees the economic, financial, and administrative activities of central and regional state bodies.

The supreme *de facto* power in Polish government, however, is the 14-member Political Bureau (Politburo) of the Central Committee of the Polish United Workers' (Communist) Party (or PUWP; see below), headed by the first secretary of the Central Committee. The party was not mentioned in the constitution until that document was amended in 1976 to note that the Polish People's Republic is a socialist state and that the PUWP is the leading political force in the construction of socialism.

Regional administration. The regional administrative system (since the reform of 1973–75) is on two levels. The largest units are the *województwa* (provinces), of which there are 49 (including three self-governing cities). Before the reform there had been only 22 *województwa;* Western observers believe that the breaking up of the large provinces had the effect of weakening the resistance of local PUWP bosses to directives from Warsaw.

Each province is responsible for the submission of proposals for five- and one-year plans for the local economy and land use to the Central Planning Commission in Warsaw. It is also responsible for achieving the targets that are eventually approved for these periods by the Sejm. However, in order to reduce the problems of coordinating the proposals of the larger number of provinces and to avoid the difficulties that are likely to arise as a result of the boundaries of the three city provinces lying well within the actual boundaries of their hinterlands, eight *makroregiony* (macro-regions) have also been established. These units have a consultative role only and are not able to compel changes in provincial proposals. Their influence in the planning process may also be reduced because their boundaries cut through func-

tional zones in some places, thus throwing doubt upon the extent to which they are able to foster the required coordination of the proposals of neighbouring provinces.

The former *powiaty* (singular *powiat*) were abolished in 1975 and the *gminy* (rural and urban districts; singular *gmina*) enlarged. These bodies, of which there were 2,327 in the late 1970s, are responsible for the provision of some cultural and social services and for local land-use planning. The four largest cities are also divided into districts, called *dzielnice*, which have similar responsibilities.

At province and *gmina* levels, local government is controlled by people's councils, whose functions locally and composition are similar to those of the Sejm nationally.

Political parties. Three political parties are active in Poland. The most important is the Polish United Workers' (or Communist) Party (PUWP; Polska Zjednoczona Partia Robotnicza, or PZPR), formed in 1948 by the union of the Polish Socialist Party (Polska Partia Socjalistyczna) and the Polish Workers' Party (Polska Partia Robotnicza). The PUWP had more than 2,400,000 members by the mid-1970s. A much smaller role is played by the United Peasant Party (Zjednoczone Stronnictwo Ludowe), with more than 425,000 members, and the Democratic Party (Stronnictwo Demokratyczne), with 92,000 members. The latter two are in no sense opposition parties.

All three parties cooperate within the framework of the National Unity Front, which embraces all political parties, mass organizations, etc. The front handles elections; it prepares a single electoral platform, selects the candidates, and submits a single list of candidates to the electorate. The balloting is so arranged that the PUWP is guaranteed a majority in the Sejm. The political composition of the Sejm changes little; after a general election in 1976 it was identical in composition to its two predecessors.

Justice. The administration of justice is in the hands of the Supreme Court, common courts (for *województwa*), and special courts for social insurance and the army. The judges of the Supreme Court are chosen for five years by the Council of State, and the other judges are nominated. The courts investigate and solve matters in the first instance with the cooperation of independent investigative officers. There is no right of judicial review, and the courts do not question the validity of government decrees and ordinances. The Supreme Court consists of four chambers: civil, penal, work and social security, and army.

Prosecutors

The public prosecutor's office consists of the general and *województwo* offices. These prosecuting bodies attend to law and order and the protection of public property; they start penal proceedings, supervise investigations, support prosecutions before the court, and have supervision over the execution of punishment. They are subject only to the attorney general, who is responsible for their proceedings before the Council of State.

The armed forces. The armed forces are based on general army service. Basic army service is started by men in a fit state of health from the age of 19 and by volunteers from the age of 18. Active service lasts from two to three years, after which the soldier passes into the reserve. There are, however, other ways of fulfilling active service. Students of higher schools of learning, for example, have their army training during their period of study. The armed forces are under the Ministry of National Defense (Ministerstwo Obrony Narodowej), which consists of the general staff, the political headquarters, the main quartermaster's department, and a number of lesser bodies. The ministry has command over the three army regions (Silesia, Pomerania, and Warsaw) as well as over the commanding staff of the naval forces, the Operational Air Force, the Armies of the Air Defense of the Country, the Armies of Frontier Defense, and the Armies of Internal Defense.

The armed forces consist of operational armies that are subject, in the case of war, to the command of the Warsaw Pact forces and of armies for the defense of the national territory. Officers have at least a secondary education. Officer candidates are trained in more than a dozen schools, and professional officers can acquire a higher education at one of the five army colleges or at civilian higher schools of learning.

Education. Education is based on the following principles: schools of all types and on all levels are free; the system of schooling is standard; and attendance throughout the 10-year primary school course and at supplementary study until the age of 18 is obligatory.

The first grade of schools is nursery schools, for children between the ages of three and six. They serve 45 percent of the children in this age group. The 10-year primary schools, for ages seven to 14, number about 18,000, with around 4,500,000 children in attendance. Secondary schools can be divided into four-year grammar schools, art schools, five-year professional and technical schools, two-year agricultural schools, and three-year professional schools (with the possibility of continuing studies in a three-year technical school).

Secondary schools

Basically, all schools are subject to the Ministry of Education, but medical schools and colleges are subject to the Ministry of Health, army colleges to the Ministry of National Defense, and higher schools of art to the Ministry of Art and Culture. The highest academic institution is the Polish Academy of Sciences (Polska Akademia Nauk), with numerous research institutes. The academy represents Polish learning abroad.

Health services. Health care in Poland is largely handled by the social health service. The central body is the Ministry of Health and Social Care, which has authority over the health departments of *województwa* People's Councils. The health service is divided into clinics, hospitals, and sanatoriums, ambulance services, and the national sanitary inspection. By 1975 there were 66 hospital beds and 17 doctors for every 10,000 inhabitants, a ratio comparable to that in the United States, Switzerland, both the German states, and Belgium, and higher than that in France, The Netherlands, Sweden, Norway, Canada, and Japan.

Health insurance

Social insurance for the health service covers free treatment for all workers and the members of their families, as well as for pensioners, invalids, students, and certain others. Those insured pay 30 percent of the cost for medicines, and invalids and pensioners, as well as those suffering from certain social diseases (notably tuberculosis and the venereal diseases), receive medicines free of charge. Apart from the health service there is also a social service whose purpose is to ensure a suitable means of support for old people and invalids. There are almost 500 homes for social care, catering to pensioners, the chronically sick, and the mentally retarded; these contain more than 50,000 patients.

Housing. Municipal and housing authorities faced a mammoth task following World War II, which obliterated about a third of the national housing and sent a flood of people from rural to urban areas. The simultaneous development of industry limited the funds available for housing, so that the situation in this area has been and remains unsatisfactory. There is a shortage of living space in large urban centres and some overpopulation of the facilities available. Housing in towns is in the hands of the People's Councils and cooperative building societies; in the country, private building is dominant.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. The culture of Poland is fed by a great variety of traditions. The source of the basic folk culture is Slavic. In the early Middle Ages (specifically, in 966), however, Poland received, along with Christianity, elements of the evolving Western world view, particularly from the Roman Catholic priesthood. Czech, German, French, and Italian influences were felt. In the Humanist period the nobility and, in part, the rising mercantile classes took much from the civilizations of antiquity. The Polish-Lithuanian political union, by contacts with the Mongols and with the Turks, brought east European and Oriental influences into the heart of Poland. Various branches of culture held sway in successive centuries: in the 16th century social and political thought predominated; in the 18th century learning and education

came to the fore, to be succeeded in the 19th century by a flowering of literature and music. Until the 20th century, cultural life was mainly focussed on and supported by the nobility and, latterly, by the merchant bourgeoisie.

Fundamental changes in Poland's cultural life really occurred only after 1945, when great social changes, together with a quite different revolution wrought by the introduction of mass media, made culture available to the masses. This process contributed to the rapid decline of folk culture and its regional variations.

The ethnic basis of folk culture. The traditional folk culture of Poland is disappearing rapidly. In the Little Poland region, however, there is still a marked distinction between the Lachy (extra-Carpathian peoples) and the Górale, or mountain people (from *góra*, "mountain"). The Górale of the Podhale area, in particular, have a most distinctive folk culture, with characteristic costumes, decoration, architecture, and music. The Górale living in the Silesian part of the Carpathians form a similarly distinctive group. Also in Silesia, the folk tradition of the ancestral Polish population around Opole is very distinctive. In Great Poland, the Sieradzanie from the middle Warta, the Kujawiacy from the upper Noteć, and the Krajniacy of the Pomeranian frontier all have their own folk culture. In Mazovia, the inhabitants of the Łowicz area are noted for their costumes and decorative skills, and the Kurpie from the middle Narew, together with the inhabitants of the Opoczno area, to the north of the Holy Cross Mountains, are also significant. In Pomerania the Kashubians are the most distinct folk group, followed by the Kociewacy from the lower Wisła, and the Borowiacy from the Tuchola area. Among the Polish population of former East Prussia there is a sharp cultural distinction between the Catholic Warmiacy, from the Olsztyn area, and the Protestant Mazurians, who are very close to the neighbouring Kurpie.

Literature. The oldest traces of Polish folklore are preserved in the medieval Latin works of the chroniclers, who handed down numerous lay and religious legends. From the Renaissance period, many texts of Polish folk plays survive. The systematic collection of songs and other works, which were to prove a stimulus to much of Polish culture, was not started until the end of the 18th and the beginning of the 19th century.

Early literature

The first half of the 19th century witnessed the flowering of the three greatest Polish poets (Adam Mickiewicz, Juliusz Słowacki, and Zygmunt Krasiński), while in the second half, and at the beginning of the 20th century, the great Polish prose writers—including the Nobel Prize winners Henryk Sienkiewicz and Władysław Reymont—were active.

Music. The dance rhythms of the *krakowiak*, *mazur* (*mazurek*, *kujawiak*, *oberek*), and *polonaise* are particularly characteristic of Polish folk music. The music of the Górale of the Podhale has a different character: in contrast to the monodic, or elegiac, music of other regions, it is polyphonic, blending a number of discrete melodies. The oldest remnants of church music stem from the Middle Ages, and the most outstanding works—notably the compositions of Wacław of Szamotuły and Mikołaj Gomółka—date from the 16th century. The development of modern musical forms—symphony, opera, and chamber music—may be traced from the end of the 18th century. Folk music was a major source of musical content (as in the opera *Krakowiacy i Górale* [*The Cracovians and the Highlanders*]), and, inspired by it, the most famous of Polish composers, Frédéric Chopin, composed much of his work for the piano. In the second half of the 19th century, Stanisław Moniuszko created the popular operas *Halka*, *Straszny Dwór*, and *Hrabina* and also was a leading exponent of the solo song. At the end of the 19th century and at the beginning of the 20th, famous Polish performers came onto the world stage, the pianists Ignacy Paderewski and Artur Rubinstein heading a procession of violinists, conductors, and singers. The most outstanding composer of the first half of the 20th century was Karol Szymanowski, and contemporary Polish composers are known far beyond the national boundaries, having won many prizes in international competitions. The leading contemporary symphony orchestras—the National Philharmonic of Warsaw and the Great Symphonic Orchestra of the Polish Radio and Television, from Katowice—have also achieved international successes.

Folk music continues to be propagated by Mazowsze and Śląsk, two representative song and dance groups, and other musical forms by such chamber groups as the Capella Bydgostiensis and the Musicae Antiquae Collegium Varsoviense. Many festivals and international music competitions are held in Poland, the Chopin piano competition, which takes place every five years, being probably the best known. Poland also has a rich variety of musical societies and magazines devoted to music.

International music competitions

The visual arts. The oldest examples of the visual arts in Poland are survivals from the prehistoric period, chiefly pottery. From the Middle Ages many fine examples of Romanesque and Gothic architecture, both secular and religious, have been preserved, together with outstanding sculptures, among which the wooden altar triptych of Veit Stoss, in the Church of the Virgin Mary in Kraków, is the most famous. The architecture and sculpture of the Renaissance and Baroque periods were formed under Italian influence but nevertheless developed individual Polish forms, as, for example, in the characteristic Renaissance attics. The best preserved urban architecture of the late Middle Ages and Renaissance is that of the old town in Kraków, together with Wawel Castle. The classicism of the end of the 18th century and the beginning of the 19th left its most valuable monuments in Warsaw, as can be seen in the palace in Łazienki Park.

Painting attained its greatest development in the second half of the 19th century, representing styles developing at that time in western Europe, but again with specific national characteristics. The greatest fame was achieved by the creator of monumental historic pictures Jan Matejko, by Henryk Siemiradzki, and by a number of landscape and genre painters.

Theatre and motion pictures. A Polish religious theatre existed in the Middle Ages, with initially mystery and later morality plays. In the 16th century the court theatre, with the first native dramas, appeared. In the 17th century and at the beginning of the 18th, opera and ballet troupes, mostly foreign, performed in the courts of the kings and the magnates. Only in the reign of the last king, Stanisław II August Poniatowski, did the first public Polish national theatre come into being, first in Warsaw and later in other towns. The most outstanding figure of this period of the theatre was the actor, director, and author Wojciech Bogusławski. After the downfall of the Polish state, the theatre played an enormous part in the survival of the Polish language and spirit. Among the actors and actresses of this period, Helena Modjeska achieved major international status. In the interwar period, acting and staging in Poland reached a very high level. During the 1960s, the Polish Laboratory Theatre (the official title is Teatr 13 Rzędów, "Theatre of 13 Rows") in Wrocław gained an international reputation for its experimental work in the training of actors and in methods of theatrical production. Under the direction of Jerzy Grotowski, it toured western Europe and the United States, where many critics have acclaimed its artistic theories and its performances as representative of the most important new direction in theatre since the work of Konstantin Stanislavsky at the Moscow Art Theatre in the early decades of the 20th century.

The Polish Laboratory Theatre

The beginnings of Polish cinema date from before World War I. In 1909 the first film-producing company, Sfinks (Sphinx), was formed, and this, with others, was active in the interwar period. Film production at this period was not, however, marked by a high artistic and technical level, despite many ambitious attempts. After World War II, Polish films had a series of international successes, and particularly in the 1950s and early '60s when Andrzej Wajda and Andrzej Munk came to the forefront as directors. By the early 1970s, however, the quality of films had diminished, and the number of both cinemas and audiences decreased, a trend also observable in other countries.

Mass media of communication. The regular broadcasting of a radio program for general reception was begun by the Polish radio cooperative in 1926, and by September 1939 there were 10 radio stations and around 1,000,000 subscribers. After World War II Polish radio began anew. Experiments with television were begun in 1948, and regular broadcasting started in 1956, first in Warsaw and then in Łódź. By the 1970s the National Committee of Polish Radio and Television had at its disposal almost 20 radio stations and more than half a dozen television centres. On a ratio per thousand people, however, Poland still held a low place among European countries in respect of subscribers to radio and television.

The press

The beginnings of the Polish press date from the 16th century, when irregularly printed information sheets and leaflets began to appear. In 1661 the first permanent periodical, the *Merkuriusz Polski Ordynaryjny* ("Current Polish Mercury"), was published in Kraków. The growth of the periodical press continued through the period of foreign domination, and before World War II there were more than 2,500 publications, some of them in the languages of minority groups, although average circulation was small. The most readable daily, the *Ilustrowany Kurier Codzienny*, had a circulation of only 150,000. By 1975 the leading dailies were *Trybuna Ludu* (organ of the Polish United Workers' Party), *Życie Warszawy* (independent), *Dziennik Ludowy* (organ of the Supreme Executive of the United Peasant Party), and *Kurier Polski* (organ of the Democratic Party). Most of the dailies, and political and social writings generally, are issued under the Prasa imprint, and distribution is in the hands of Ruch, a nationwide group.

PROSPECTS

Poland is a proud country with roots deep in the past. Its cultural life, in particular, has been very rich, drawing on a thousand years of tradition and sustaining the national spirit when Polish frontiers and indeed the Polish state collapsed as a result of the country's unfavourable historical and geographical position. Since the blows suffered during World War II, the government has attempted to build a new state without losing sight of tradition. By the 1970s much had been accomplished in economic growth, especially in heavy industry, and in reform of education and cultural life. But, as Polish experts acknowledged, much remained to be done in both economic and social life to raise standards to those of western Europe. (J.A.K./A.H.D.)

BIBLIOGRAPHY. The quinquennial *Poland, a Handbook* is a comprehensive outline of the country and its society issued by the government of Poland; an extensive list of Polish-language publications available in English translation, *Polish Books in English, 1945–1971* (1974), compiled by JANINA HOSKINS, was issued by the U.S. Library of Congress; the basic Polish source is KAROL J. ESTREICHER, *Bibliografia Polska* (39 vol., 1870); the most extensive reference work is the *Wielka Encyklopedia Powszechna PWN*, 13 vol. (1962–70), published by the Polish Academy of Sciences; R.H. OSBORNE, *East Central Europe*, ch. 8 (1967), a concise geographical summary of facts about the land and its people; INSTITUTE OF BRITISH GEOGRAPHERS, *Problems of Applied Geography: Proceedings of the Anglo-Polish Seminar, Keele, Sept. 9–20, 1962* (1964), contains important contributions on contemporary industrialization and urbanization in Poland and on their influence on land use and agriculture; J. BOGUCKI, *The Fine Arts in Poland Today* (Eng. trans. 1967); J.Z. LOZINSKI and A. MILOBEDSKI, *Guide to Architecture in Poland* (Eng. trans. 1967), includes an introductory outline that concisely describes the development of architecture in the territory that constitutes Poland today, with maps of selected architectural monuments and a list of all the buildings located; J.C. FISHER (ed.), *City and Regional Planning in Poland* (1966), a comprehensive record of city and regional planning in Poland from theories and methods to specific case studies; *Polish Statistical Yearbook*, up-to-date statistics published annually.

Poland, History of

Poland is a nation whose often tragic history is yet coloured by past eminence and a vital role played in European history. Emerging from a loose union of Slavic tribes in the 9th century, Poland had become by the late Middle Ages a virtual empire with the highest level of civilization in eastern Europe; it was a bulwark of the Roman Catholic faith there. Polish political theory, culture, and science had a vitalizing effect on the mainstream of European thought, and ideals such as religious toleration were pioneered in Poland at a time when most of Europe was engaged in bitter persecution. Although it virtually disappeared as a state after the three partitions of the late 18th century, Poland was re-established after World War I. Its invasion by Germany in 1939 was the event that marked the outbreak of World War II.

This article is divided into the following sections:

I. Poland during the Middle Ages, 966–1492
- The founding of the Piast Kingdom, 966–1025
- Unity and disintegration, 1025–1241
- The Mongol invasions and the re-establishment of the kingdom, 1241–1382
- The Jagiellon dynasty, 1382–1492

II. The Polish Commonwealth, 1492–1795
- The "Golden Age" of the Polish-Lithuanian Empire, 1492–1572
- The "Silver Age" of the Royal Republic, 1572–1648
- The Iron Era period of wars and disintegration, 1648–97
- The Saxonian era and the Russian protectorate, 1697–1763
- Reform and partition during the reign of Stanisław II Augustus Poniatowski, 1764–95

III. Poland under partition, 1795–1914
- Foreign rule and the Duchy of Warsaw, 1795–1815
- The Kingdom of Poland and the other Polish lands, 1815–31
- Russian reprisals and the January Insurrection, 1831–64
- The founding of modern Poland, 1864–1914

IV. Modern Poland, since 1914
- The founding of the Second Republic, 1914–21
- Poland between the wars, 1921–39
- Poland during World War II (1939–45)
- Poland after World War II

I. Poland during the Middle Ages, 966–1492

THE FOUNDING OF THE PIAST KINGDOM, 966–1025

Early origins. The Poles belong to the North Indo-European language group comprising the Slavs, the Balts, and the Germans. The primordial homeland of the Slavs can be located in the immediate vicinity of the most ancient settlements of the Germans and Balts, the area occupied by the "Poles" lying in the basins of the Oder and Vistula rivers. The Polish nation originated from a union of such West Slavic tribes as the Polanians (Polanie), Vistulans (Wiślanie), Silesians (Ślęzanie), East Pomeranians (Pomorzanie), and Mazovians (Mazowianie). These "Polish" tribes shared a basically common culture and language and were much more closely related to each other than, for example, the Germanic tribes who founded the Franconian or German kingdom about 800–1024. Chronicles and geographical descriptions of Carolingian, Saxonian, Arabian, and Byzantine origin have described in some detail these Polish tribes, their names, and the regions where they lived. The Polanie tribe (open-country dwellers) lived in the fertile Polish heartland crossed by the Warta River, and this was the group after which the later Polish nation was named. Besides the Polanie tribe, there were approximately 20 Polish tribes that formed small states between the years 800 and 960. The individual tribes owed allegiance to regional leaders, chosen from an assembly of all free Polish males as commanders in case of war. These were the beginnings from which the dynasties of hereditary kings, princes, or dukes slowly evolved. Each tribe commanded several well-fortified castles or fortresses, the larger castles developing into principalities or "castle districts." There were usually fortified city-settlements (*suburbia*) near all the larger castles even in very early times. Inhabited by artisans, workmen, and merchants, these cities also served as refuges on occasion. As early as 900–1025, along the Baltic Sea coast, there were large trading centres with between 6,000 and 10,000 inhabitants; these included Wolin (Wollin) and Szczecin (Stet-

The Polanie tribe

tin), on the mouth of the Oder River, and Truso (later Elbląg) and Gdańsk (Danzig), founded before 990 on the mouth of the Vistula River. The ancient foreign trade routes linking this area with distant regions stimulated the founding and development of small states. The most significant trade route from East to West connected Ratisbon (now Regensburg)–Vienna and Cracow with Kiev and Byzantium (now Istanbul). A well-developed monetary exchange system made it possible for the individual dukes to raise tariffs and tributes, in this manner exerting their sovereign rights on the monetary economy. In all probability, mining of precious metals and iron ores must have taken place in Upper Silesia and in the Polish Central Highlands at that time.

Because of its advantageous location with regard to foreign trade, the region grew into a larger state under the dynasty of the Wiślanie (or Vistulans). One of the dynasty's dukes was described as very powerful as early as 875, and, in all probability, his rule already extended over the Lędzianie (who lived either on the upper Vistula or the upper Warta River) and perhaps even over some of the East Slavonic tribes on the Bug and San rivers. Around 875 to 880 his state and Silesia came under the sovereign rule of the Great Moravian Empire, led by Prince Svatopluk (ruled 870–894). Under his reign, Christianity was introduced in both its Latin and Greek forms through the disciples of "the apostles of the Slavs," Cyril (died 869) and Methodius (died 885). Destroying the Great Moravian Empire, the Magyars conquered and settled the Hungarian lands from about 896 to 907. This restored the independence of southern Poland for a short time until it came under the control of the Bohemian dukes Boleslav I (died 972) and Boleslav II (ruled from about 973 to 999).

Introduction of Christianity

Both the Latin and Eastern Orthodox faiths were already well established in the area by this time, and the first bishopric of Cracow was established in 969. Even though nothing is known about a local dynasty, a territorial association was formed a few decades later, consisting of the tribes of southern Poland, commonly known as Little Poland (Małopolska). Little Poland, with the exception of a few border regions, covered the same area as the bishopric of Cracow.

The Piast dynasty. The second and politically more important founding of a state took place under the Polanie dynasty. The descendants of "Piast," the legendary founder of the dynasty—Siemowit (Ziemowit), Leszek (Lestko), and Siemomysł (Ziemomysł)—united the lands around Poznań, Kruszwica, Gniezno, and Kalisz, and this region later became known as Great Poland (Wielkopolska).

Mieszko I, c. 963–992. Siemomysł's son, Mieszko I, was the first of the Piast dukes to have his biography recorded in detail in the chronicles of that time. Through his desire to conquer Western Pomerania along the Oder River to the trading cities of Szczecin and Wolin, Mieszko I came into contact with the German kingdom, whose ruler Otto I the Great (the Holy Roman emperor) had conquered and subjugated Bohemia in 950 and all the Polabic tribes (between the Elbe and the Oder rivers) between 936 and 955. This extended the sovereignty of the German king to the Oder River, and, for the time being, Mieszko acknowledged Otto's supremacy and entered into a close trade and cultural relationship with the Duchy of Saxony, Otto's homeland. To counter the German threat, Mieszko I entered into an alliance with the Bohemian duke Boleslav II, marrying his sister, Dubravka (Dąbrówka), in 965. Baptized in 966, he promoted the general spread of Christianity in his lands, although a few small Christian communities had already been in existence. In consequence, the year 966 is considered to be the year that marks the founding of Poland. The bishopric of Poznań was established in 968 and placed under the sovereignty of the archbishopric of Magdeburg. During the Great Slav Rebellion in 983, Germany lost its territory east of the Elbe and Saale rivers. In tacit agreement with Germany, Mieszko I conquered the entire coastal region of the Baltic Sea, including Szczecin, during 967 to 990. After dissolving his treaty with Bohemia, he further added Silesia and Little Poland to his other conquests. This established Mieszko as the ruler of an empire covering about 96,500 square miles (or 250,000 square kilometres) and populated by 1,250,000 people. Poland was now larger and more powerful than the Scandinavian countries, Hungary, or Bohemia. Of decisive importance for the later development of internal unity was the fact that Mieszko submitted to the formal supremacy of the Holy See in Rome around 990. He took this step in order to avoid being subjected to the authority of the German Church.

Bolesław I the Brave, 992–1025. The eldest son of Mieszko I, Bolesław, further enlarged his empire, going beyond the ethnic boundaries of the Polish tribes. With Bolesław's help, the Holy Roman emperor Otto III set up his own Polish metropolitanate (church province), with the bishoprics of Cracow, Wrocław, and Kołobrzeg and the archbishopric of Gniezno, at the Congress of Gniezno (1000). The first bishop of Prague, Adalbert (Voytěch), became the national patron saint of Poland in that year, following his martyrdom in 977 during his missionary work with the heathen Prussians in Samland in east Prussia. When Otto III died, Bolesław I conquered Bohemia, Moravia, the German borderlands in Lusatia (Lausitz), west of the Oder River, and probably also Slovakia between 1003 and 1004. But the new king of Germany, Henry II (ruled 1002–24), reconquered Bohemia and also led three campaigns against Bolesław I, without success. In 1018 Bolesław turned east, occupying Kiev for a short period and regaining the borderland at the Bug and San rivers, which had belonged to Poland before 981. Bolesław I was crowned king (presumably with the consent of the Holy See in Rome) in 1024–25, and the new kingdom became gradually known under the name of Poland (Polonia) even during his reign. After his death the Polish monarchy came to be the secular symbol of the country's unity, in spite of all the partitions and divisions that the future held in store.

UNITY AND DISINTEGRATION, 1025–1241

Attempts to expand Polish domination. *Mieszko II Lambert, 1025–34.* One of the younger sons of Bolesław I, Mieszko II Lambert was well educated but did not possess the strategic and statesmanlike talents of his father. After victorious campaigns against the barbarian Polabs in 1030, he attacked the German king Conrad II, who, in alliance with the Russian grand prince Yaroslav of Kiev, attacked Poland (1031), further supported by rebellions within the country. Eventually, about 1033, Mieszko was forced to recognize German supremacy. His death in 1034 thus came at the height of a crisis. The boom in foreign trade that had been a factor in the establishment of the Piast empire had come to a close; the regional overlords were in a state of rebellion, and the barbarian insurrections deeply affected the barely established organization of the church. Chaos followed for the next few years (1034–39), the Bohemian duke Bretislav I occupying Silesia, totally devastating Great Poland, and resettling many of its inhabitants in Bohemia. The coastal regions from Szczecin to Gdańsk seceded from the Polish crown for about a century. The bishopric of Kołobrzeg—and with it the entire Christian movement in Western Pomerania—foundered again.

Defeat of Mieszko II

Casimir I the Restorer, 1039–58. A former monk, Casimir (Kazimierz) I slowly restored Poland with the help of the German king Henry III. Both Great Poland and Little Poland recognized Casimir I as their duke in 1039–40. Mazovia followed in 1047 and Silesia recognized him in 1050, but Casimir had to pay tribute to Bohemia for Silesia.

Bolesław II the Bold, 1058–79. Casimir's successor, Bolesław II, continued the power politics of Bolesław I, leading many successful campaigns into Hungary (1060–77) and Kiev (1068–77). He also maintained the independence of Poland from Germany by skillfully taking advantage of a controversy between the German king Henry IV and Pope Gregory VII. With the help of the Pope he also renewed the crown of Poland. But after Henry IV had regained power (1077), a party of Polish

nobles rebelled in his favour against Bolesław. For reasons unknown, Bolesław had decreed the execution of Bishop Stanisław of Cracow in 1078, and this led to open rebellion. Bolesław was forced to flee to Hungary, where he died in exile in 1081. His younger brother Władysław I Herman (1079–1102) sided with the imperial party and in 1088–89 married Judith, Henry's sister, and relinquished all claims to the Polish crown.

Exile of Bolesław II

Bolesław III the Wry-Mouthed, 1102–38. Władysław's younger son, Bolesław III, again strengthened the position of the Polish monarchy, expelling his elder brothers and his half brother Zbigniew (1107). Two years later he drove back a campaign of the German king Henry V into Silesia. In difficult campaigns, he reconquered all of the Western Pomeranian lands during 1113–22 and incorporated Eastern Pomerania and Gdańsk into Poland. Bolesław then ordered the final and complete Christianization of Western Pomerania and forced Polish rule on the land of Lubusz (Lebus), on the Oder, thus restoring the previous boundaries of the empire of Mieszko I. But the German emperor Lothair finally pushed the German frontier to the banks of the Oder and the western shores of the Neisse (Nysa) River. Bolesław was forced to swear allegiance to Lothair in return for Western Pomerania and the island of Rügen in 1135.

The period of division. Endeavouring to reconcile the separatist demands of the provinces with the needs of national unity, Bolesław III divided Poland among his sons, with the result that Poland ceased to be a united state for two centuries. The eldest son, who was already prince of Silesia and also owned Pomorze (Pomerania), was to have seniority over his brothers and rule over "the land of the senior," with its capital in Cracow; this territory stretched from western Little Poland to central Poland and on to the Baltic shore (eastern Pomerania and Gdańsk). His brothers ruled over Great Poland with Poznań and western Kujavia; Mazovia (in the basin of the middle Vistula) and eastern Kujavia (within the Vistula bend); and Sandomierz (east of Little Poland).

The seniority system. A "seniority constitution" was also introduced with this arrangement, but the plan immediately led to strife. The first "senior," Władysław II the Exile (1138–46), was forced into exile and fled to Germany. His brother Bolesław IV the Curly, hereditary duke of Mazovia, asserted himself as senior between 1146 and 1173, but this led to conflict with the German emperor Frederick I Barbarossa, who intervened twice (1163 and 1172) to secure Silesia for the sons of Władysław II. Władysław's sons divided it among themselves and founded the lineage of the Silesian Piasts; yet another son became the third senior. Of these sons, Mieszko III the Old (1173–77) was exiled from Cracow during a rebellion in 1177 but succeeded in retaining his hereditary land, Great Poland. His youngest brother, Casimir II the Just (1177–94), succeeded as senior after him and, in 1186, also acquired Mazovia, thus becoming the ancestor of the Mazovian line of the Piasts. After the Congress of Łęczyca in 1180, the land of the senior—Little Poland with its capital, the city of Cracow—became a hereditary land like the other lands. Both Pope Alexander III and the emperor Frederick I recognized this decision, but one of its great disadvantages was that the internal unity of Poland would be constantly exposed to foreign hereditary claims as a result. Leszek the White was the last duke who claimed the honour of senior with some success, but he was assassinated together with other Piast dukes in 1227.

The reigns of Mieszko III the Old and Świętopełk I

Territorial losses. The internal disintegration of Poland led to the secession of the lands along the Baltic Sea and other border regions. Bogusław I (who ruled about 1156–87), the son of Warcisław I, was forced to yield Western Pomerania to the Danish crown. Even the reigning governor of Eastern Pomerania and Gdańsk, Świętopełk I (died 1266), took advantage of the weakness of the Piasts. Instrumental in the assassination of Leszek the White, he asserted his country's independence and styled himself sovereign duke of Eastern Pomerania, though regarded as a usurper by the Piasts.

The Teutonic Order. The establishment of the Order of the Teutonic Knights, a German crusading organization that had lost a sphere for its activities in Palestine, was of dangerous consequence for Poland. His lands harried by heathen Lithuanians, Jatvingians, and Prussians, Duke Konrad I of Mazovia (died 1247) called for help from the Teutonic Order in 1225–26, leaving them the lands of Chełmno (Kulm) east of the lower Vistula River as a territorial base for combat and control of the Prussians in 1228–30. The grand master of the Teutonic Knights, Hermann von Salza, received sovereign rights over the "masterless heathen country" of the Prussians in the Golden Bull of Rimini (1226) from the emperor Frederick II and in the Privilegium of Rieti (1234) from Pope Gregory IX. His Order then tried to create an independent state, in spite of Konrad I's claim to sovereignty over the land. In the event, the Order established a system of local government in Chełmno and in Prussia. The subjugation and Christianization of Prussia was completed by 1283. The Teutonic Order soon became the principal military power in Europe, for it was in possession of about 20 provinces (bailiwicks) in Germany and Italy alone and received reinforcements from all the knightly bodies in Europe for its wars.

Supremacy of the Teutonic Order

Internal developments. In Poland the era of knighthood, from about 1025 to 1241, saw the emergence of a Latin church organization and, in architecture, the development of the Romanesque style. By now the Polish aristocracy owned large landed estates and also occupied the most important administrative positions at court and in the provinces. With the division of the country into six to nine provinces and the continuous partition of the country itself from 1138 on, similar posts were also established within each individual duchy. The division into castle districts, each under a feudal baron, remained within the individual sections of the country. Fortified by mighty ramparts, the cities were under the patronage of the chief castles. The internal unity of the Polish nation remained intact in spite of the many partitions and, in accordance with Polish law, the dukes exercised no personal power but a collective sovereignty. From about 1058 Cracow was considered the common capital for about five centuries. The most important protector of the unity of Poland was the metropolitanate of Gniezno. The church supremacy of Magdeburg was once again postulated in 1133, but it was definitely rejected and Gniezno won recognition as its own church province. The archbishopric of Gniezno held the largest estates in Poland (1136), controlling 149 cities and villages. The oldest monastery of Poland was established about 1044 in Tyniec (near Cracow) by the Benedictines. The Order of the Cistercians led the second wave of founding monasteries, mainly from Morimond (Burgundia) and Altenberg (near Cologne).

THE MONGOL INVASIONS AND THE RE-ESTABLISHMENT OF THE KINGDOM, 1241–1382

The Mongol invasion (1241–42), in Poland's period of division, turned the tide of Poland's history. The Mongol leader Batu Khan first conquered Kiev and most of the Russian principalities before defeating the Polish army at Chmielnik and the Silesian knights of the Teutonic Order at Legnica (Liegnitz; 1241). This frustrated the attempt of the two Silesian dukes Henry I (1201–38) and Henry II (1238–41) to unite all the Polish lands. The Mongols retired in 1242 in order to elect a new great khan in Mongolia but still remained masters, for another 50 years, of Red Ruthenia (Red Russia), east of Little Poland; from this base their repeated raids devastated Poland. Meanwhile, Great Poland and especially Mazovia suffered from raids by the heathen Lithuanians (1246–1307) and Jatvingians (1248–82). The Mazovian border region around Drohiczyn on the Bug River was lost to Mongol Red Ruthenia around 1241–52, and the land of Lubus with its river passage across the Oder River to the Askanians of Brandenburg, about 1249–52.

The Mongol raids

Attempts at reunification. When Bishop Stanisław was canonized in 1253, Poland acquired a new national patron saint whose cult spread rapidly in the church province of Gniezno. As part of the community of the Holy See, Polish church members contributed the special levy,

Peter's Penny, to Rome. The Synod of Łęczyca (1285) further promoted unity by demanding the advancement of the Polish language. The Bohemian king Vaclav II (1278–1305) attempted a new dynastic reunification of Poland. Establishing his sovereignty over most of the Silesian duchies around 1289–91, he also gained supremacy over the Duchy of Cracow in 1291–92.

Przemysł II. His opponent was Przemysł II (1279–96), who had reunited Great Poland (including Gniezno) in 1279 and Eastern Pomerania in 1294 and who became the first Piast duke to regain direct rule over more than just a sector region. His coronation by the archbishop in 1295 in Gniezno re-established the kingdom of Poland. Przemysł was assassinated in 1296, however, presumably at the instigation of the Margrave of Brandenburg.

Re-establishment of the kingdom

Vaclav II. Vaclav (Wacław) II (1300–05) of Bohemia, who succeeded him as king, ruled over about two-thirds of all Polish territory. Vaclav II himself considered these lands as almost an extension of Bohemia and, consequently, as part of the Holy Roman Empire. His early death in 1305 and the assassination of his son Vaclav (Wacław) III the following year left both the Polish and the Bohemian thrones vacant.

Władysław I the Short, 1314–33. The opportunity had now arrived for the minor Piast prince Władysław the Short to establish himself. Władysław cleverly availed himself of the support of Hungary and the traditional protection of the Holy See. He occupied Little Poland in 1305 and sought the help of the Teutonic Order to regain Eastern Pomerania in 1308. The Teutonic Knights, however, bought the interests to these lands from the Brandenburg rulers and established themselves as the legal authority, incorporating Eastern Pomerania into Prussia. By 1314 Władysław had secured Great Poland, and in 1320 he was crowned king of Poland as Władysław I with the consent of the Roman Curia. This act restored the Polish Crown through a national dynasty, and for the next 475 years the crown was to remain the symbol of unity for Poland. The Piast dukes welcomed immigrant Germans and Flemings into Polish territories, and resettlement took place according to the municipal law of Magdeburg, based on German law, which was applied first to the great old cities of Wrocław (1241), Poznań (1253), and Cracow (1257).

Struggle against the Teutonic Order. The new king Władysław I, faced the opposition of a coalition between Bohemia and the Teutonic Order. The new Bohemian king, John of Luxemburg, laid claim to the Polish crown, and almost all of the Piast dukes in Silesia had recognized him as their feudal lord by 1327–29. Only the little principality of Wieluń (1321–29) was loyal to Władysław I. He instituted lawsuits against the Teutonic Order for the sovereignty of Eastern Pomerania but without success. The Teutonic Knights and the Bohemians terribly devastated Great Poland, especially in the war of 1328–32.

Alliance with Hungary and Lithuania. At the Battle of Płowce (1331), however, Władysław drove them back and, in alliance with Hungary and Lithuania, secured Poland. Ties with Hungary were strengthened by the marriage of Władysław's daughter Elizabeth (Elżbieta) to the new Angevin king of Hungary, Charles I, in 1320, and the Polish crown prince Casimir (Kazimierz) was married to the Lithuanian princess Anna (Aldona) in 1325, thus creating a pact with Lithuania.

Casimir III the Great, 1333–70. Władysław's son Casimir was a highly talented statesman and administrator. During the congresses of Trenčin and Vyšehrad in Upper Hungary (Slovakia), he persuaded John I of Luxemburg to relinquish his claims to the Polish crown. In return, after an unsuccessful campaign in 1348 and in face of the Silesian Piast dukes' support of the Bohemian crown, Casimir was forced to renounce finally all demands on Silesia, the country's most densely populated and wealthiest region. In the Treaty of Kalisz in 1343, Casimir relinquished all claims to Eastern Pomerania in favour of the Teutonic Order—for the time being. Instead, he coerced the Piast dukes of Mazovia to recognize him as their overlord and seized parts of Mazovia. He also regained some of the lands in the west that had been annexed by Brandenburg. Between 1340 and 1352 Casimir won Red Ruthenia with the cities of Przemyśl, Lwów (Lvov), and Halicz; and in 1366 he extended his rule over the Ruthenian principalities of Chełm, Betz, Włodzimierz, and Podolia. Poland had now become a power, and at the Congress of Cracow, Casimir functioned as a mediator between Emperor Charles IV of Bohemia and Louis (Ludwig) I of Hungary. Casimir died in 1370 without a male heir, but in his will he appointed his grandson Casimir to be prince of Pomerania and awarded him parts of the Polish territory to tie Western Pomerania to Poland.

Treaty of Kalisz

Louis I of Hungary, 1370–82. Poland nevertheless remained a unified kingdom under Casimir's successor, his nephew Louis I of Hungary. The only losses were some border areas annexed by Brandenburg and Lithuania. Louis, however, regarded Poland as his secondary kingdom, and a specially appointed governor took over the administration of Halicz Ruthenia. The Polish aristocracy, a deciding factor in the succession to the throne of Poland in 1370 and 1382, obtained from Louis the grant of special privileges at Košice (Kaschau; 1374), which formed the basis of a Polish "Magna Carta."

Internal developments and reform. *Construction of cities and buildings.* The years from 1320 to 1382 marked the founding of new cities. During the reign of Casimir the Great, as many as 47 new cities were founded in Little Poland and 32 in Great Poland. Casimir also established local courts of justice for the cities of Little Poland (1356) and Great Poland (1365), in order to separate his country from the jurisdiction of Magdeburg and the need for its appellate court. A great builder, Casimir was said to have found a country of wood and left a country of stone: he ordered the construction of 53 castles and fortifications for 27 cities, and during his reign numerous Gothic town halls and merchant houses were erected.

Role of the nobility. Even though a new social class of burghers and merchants evolved in the cities of Poland, the aristocracy remained the most influential group in politics. The first confederation of nobles in Polish history was founded in Great Poland in 1352–56. Since the previous official positions in the separate palatinates of Poland had long before become the benefices of the aristocracy, Casimir and Louis appointed a governor (*starosta capitaneus*) for each section of the country, after the example set by Vaclav II (1291). The office of *starosta* remained the most important branch of executive power during the next 470 years. In the Statute of Wiślica (about 1346) Casimir codified the law that was in force in Little Poland. This *jus Polonicum*—the law of the nobility (as opposed to the *jus Teutonicum*, which was the law for the burgher-peasant population of Poland)—was later extended and applied throughout Poland, including Red Ruthenia and Podolia.

The church. The Catholic Church influenced the character of public life in Poland. Poland had lost Silesia in 1348, but the Silesian bishopric of Wrocław remained under the church authority of Gniezno for another 400 years. The Dominican, Franciscan, and Augustine mendicant preaching orders had already built 80 monasteries in the metropolitanate of Gniezno by 1382. Yet, in spite of the power of the Catholic Church, the ideal of religious tolerance became so entrenched in the consciousness of the Polish people that Casimir the Great found it advisable to establish it as one of his basic policies. In addition to Gniezno, the metropolitanate of Halicz-Lwów was established; and a separate Eastern Orthodox metropolitanate in Halicz (later Lwów), and an Armenian bishopric (later archbishopric) of Lwów were founded. Casimir's kingdom was the only continental European country where several Christian creeds were able to live side by side for any length of time two centuries before the Reformation.

Religious tolerance

Protection of the Jews. Casimir the Great also provided, with the Statute of Wiślica (1346), protection for the Jews from pogroms and ritual murders at the hands of the Christians. When the Black Death raged through the German lands between 1348 and 1352, over 300 Jewish communities were either destroyed or expelled from their

homelands. The survivors migrated to Poland. Ever since, the majority of Polish Jews has consisted of Ashkenazim (from Germany). Retaining their own Middle High German language—later evolved into Yiddish—most of the Jews rejected any language assimilation until their near extermination by the Nazis during World War II. The Poland of Casimir thus developed into a cosmopolitan empire: Poles, Germans, Ruthenians, Flemings, Walachians, Jews, Armenians, and even Tatars and Karaites living together. There were then about a dozen official languages of administration and worship. Religious and political tolerance was reflected in the founding in 1364 of the University of Cracow, after Prague the second oldest university in Central or Eastern Europe, whose brilliant faculty of law prepared for the later juridical thinking during the Renaissance in Europe.

Founding of the University of Cracow

THE JAGIELLON DYNASTY, 1382–1492

Jadwiga and Władysław II Jagiełło. The union between Poland and Hungary was broken with the death of Louis because the Hungarian aristocracy chose his eldest daughter, Maria, to be their queen, whereas the Polish aristocracy gave preference to his ten-year-old daughter Jadwiga (1384). This was the first step in the direction of developing an elected Polish monarch. The Polish aristocracy favoured a dynastic alliance with Lithuania, especially in order to regain Red Ruthenia and Podolia, which had become estranged from Poland under Louis's regime.

Union of Poland and Lithuania. The grand duke of Lithuania (from 1377 to 1401), Jagiełło (Iogaillas), was still a heathen; but after adoption by the widow queen Elizabeth in the Union of Krewo (1385), he was married to the crown princess Jadwiga. Jagiełło thus accepted the Catholic faith and became the recognized "Master and King of Poland" under the assumed name of Władysław II Jagiełło in 1386. This procedure created a Polish–Lithuanian union, a union that continued under the Jagiełło dynasty during the following 183 years. But the effect was that Poland was able to incorporate, during 1385–86, a state about four times as large as the original Polish realm. The bishopric of Wilno was established in 1387 to convert the Lithuanian people to the Catholic faith. Another area belonged to the Eastern Orthodox Church. The polyethnic, multireligious, and legal-constitutional structure of the Jagiełło empire was more complicated than ever before. The authority of Władysław II also included the old Russian capital Kiev, which had come under Lithuanian suzerainty in 1366 and was now reduced to the status of a residential border town, set up as a buffer against the Mongols and Tatars. To the southeast of these lands stretched more than 58,000 square miles (or about 150,000 square kilometres) of almost uninhabited steppe country, which belonged to Lithuania in name only. The annexation of Lithuania secured the eastern boundaries of Poland against the attacks of the Mongol and Tatar tribes as well as against those from the great principality of Moscow, then under Mongol and Tatar control. It meant also, however, that Poland had to defend Lithuania against the incursions of the Teutonic Order. With the help of the Lithuanians (1387), Poland was able to renew the incorporation of Red Ruthenia and Podolia and to assert its authority over the principality of Moldavia (1387–1497), until then a minor country under Hungarian rule. This meant access to the direct trading route to the Orient, and Poland now became a partner in the rich Oriental commerce that flowed via the cities of Kaffa (Feodosia, in the Crimea) and Akkerman (now Belgorod-Dnestrovsky).

Władysław II and Vytautas. Władysław II and Jadwiga reigned together until Jadwiga's death in 1399. Władysław then ruled as the sole king of Poland (1399–1434). In Lithuania, Władysław found it expedient to acknowledge his cousin Vytautas (Witold; 1392–1430), first as governor, then as duke (1401), and finally as grand duke of Lithuania (1413). But he himself kept supreme command. To free his hands to attack the Tatars and annex Moscow, Vytautas surrendered the Duchy of Samogitia (Żmudź), which linked the Teutonic Order lands of Prussia and of Courland, to the Teutonic Order in 1398 in return for the cessation of all hostilities between them. He suffered a grave defeat in 1399, however, on the Vorskla River against the Mongols of the Golden Horde, which ended Lithuania's eastward expansion. Vytautas also supported rebellions in the still pagan Samogitia (1401 and 1409), which prompted the Teutonic grand master Ulrich von Jungingen to declare war against Władysław II in 1409. The Teutonic Knights suffered a decisive defeat in the Battle of Grunwald (Tannenberg) in 1410, which broke the power of the Teutonic Order. The first Treaty of Toruń (1411) returned Samogitia to Lithuania; the return of Eastern Pomerania to Poland was discussed at the Council of Constance (1414–18) and was a factor in two later wars (1421–22 and 1431–35), but the Order remained in possession of the territory. When the Protestant Hussite rebellion broke out in 1419, Władysław II favoured the Hussites, even promoting his relative Sigismund (Zygmunt) as the Hussite candidate for the Bohemian throne in 1422. Because this brought him under suspicion of being a heretic and friend of the heathens, Władysław felt obliged to issue the strong Edict of Wieluń in 1424, against Hussite trends among the Polish nobility, and he crushed a pro-Hussite confederation of the lower Polish nobility in 1438–39. Władysław had obtained 16 mainly German mountain cities in the Spisz (south of the High Tatra Mountains) in 1412, thus giving Poland a share in the mining activities of upper Hungary, the richest in Europe at that time.

Defeat by the Mongols

Edict of Wieluń

Władysław III, 1434–44. Władysław II was succeeded on his death by his ten-year-old son, Władysław III, the chief position in the realm being held by the most eminent of the Cracow magnates, Zbigniew of Oleśnica, bishop of Cracow and one of the previous king's chief advisers. Władysław III renewed the Polish–Hungarian union with the help of the Hungarian commander in chief János Hunyadi (1443–44). Under the influence of the strongly Catholic Zbigniew of Oleśnica, he organized a crusade in order to save Byzantium, which was then threatened by the Ottoman Empire. Władysław died when only 20 in the catastrophic defeat of Varna (1444) on the coast of the Black Sea. The Ottoman troops conquered Byzantium in 1453 and marched immediately against Hungary. The threat was delayed for some decades, however when Hunyadi defeated the Ottomans in 1456. But the Hungarian imperial administration of Hunyadi terminated the Polish–Hungarian union.

Casimir IV Jagiellończyk, 1447–92. The younger brother of Władysław III, Casimir, had been ruling as the grand duke of Lithuania since 1440, but it was only after a long interregnum in Poland that he was finally recognized as the king. He signed a treaty with Moscow that established the Lithuanian boundary about 90 miles (150 kilometres) west of Moscow. This secured the eastern boundary for 37 years. The peace in the east became important when, in 1454, uprising against the rule of the Teutonic knights broke into the open, led by the cities of Gdańsk, Toruń, and Elbląg, which requested Casimir IV to integrate the Prussian–Eastern Pomeranian lands of the Teutonic Knights with Poland (1454). This marked the start of the Thirteen Years' War against the Teutonic Order. The second Treaty of Toruń (1466) recovered for Poland the former Eastern Pomerania (which received the new name of Royal Prussia), along with the bishopric of Warmia (Ermland). The rest of the territory of the Teutonic Knights acknowledged the king of Poland, and Königsberg became the new residence of the Order's grand master. Poland won back the access to the Baltic Sea that it had lost in 1308–09, and Polish rule now stretched from the Baltic Sea to the Black Sea for a period of 32 years. Poland became a trading partner with access to the most profitable trading routes crossing the European continent.

Extension of Polish rule

Extension of the empire. The King's main interest still lay with the so-often-attempted union with Bohemia and Hungary. Casimir IV succeeded in temporarily acquiring the Kingdom of Bohemia in 1471 (finally acquired 1478–79), for his oldest son Władysław (Vladislav). The smaller Bohemian lands, Moravia and Silesia, remained

under the rule of Matthias I Corvinus, the son of János Hunyadi. But after the death of Matthias, even Moravia, Silesia, and Hungary finally recognized Władysław as their king (1491–92). The result of this dynastic union was the creation of one of the largest empires in Europe. By 1491–92 Casimir IV controlled an area of at least 336,000 square miles (870,000 square kilometres), while his son Władysław ruled the Bohemian lands of the crown of Vaclav and the Hungarian lands of the crown of Stephen, an area of at least 174,000 square miles (450,000 square kilometres). This agglomeration of states was, however, as much torn by internal strife as it was exposed to external danger—the greatest external threat coming from the Ottoman Empire and its ally, the tsar of Moscow.

Growth of the power of the nobility. Political authority in Poland was gradually taken over by the aristocracy. Aided by the mechanism of the election of kings, the many wars, and the tax exemptions, they were able to slowly reorganize the kingdom into a monarchy of the estates. The main organ of aristocratic authority was the Privy Council (1385–1493), which became the Senate after 1493. The rights and privileges of the nobility were successively established between 1422 and 1433, securing for them the right to a voice in political matters, their factual exemption from taxation, and finally the political and judicial maxim of habeas corpus, or no imprisonment without trial. After the Union of Horodło (1413), the regional assemblies slowly grew into provincial diets (*sejmiki*), becoming recognized as standing establishments in all parts of the country with the Privilegium of Nieszawa (1454). In addition to the Polish diet (*sejm*), the Prussian diet held a special position with its own particular law. The lower nobility represented its own interests (which were usually directed against the special hegemony of the aristocracy) through elected delegates.

Privileges of the nobility

Development of parliamentarianism. These delegates marked the beginning of parliamentarianism in Poland when they assembled for the first time in a House of Legates at Piotrków (1493). After this, the Polish Senate became the upper house and the House of Legates became the lower house of the Polish Parliament, of Sejm. Lithuania also absorbed the culture and the system of government of Poland through the German municipal laws, introduced by the Polish burghers. According to the statutes of 1386–87, at first only the newly converted Catholic nobles were equal to the nobility of Poland, excluding the Eastern Orthodox Boyars, many of whom seceded to the Tsar of Russia. But the Lithuanian and Ruthenian nobles were gradually assimilated into the Polish sphere of interests, and in 1447–93, the first diets of the great principality were formed.

II. The Polish Commonwealth, 1492–1795

THE "GOLDEN AGE" OF THE POLISH-LITHUANIAN EMPIRE, 1492–1572

After the death of Casimir IV, his second son, John Albert (1492–1501), ascended the throne of Poland, and the two countries were temporarily separated, his third son, Alexander, becoming grand prince of Lithuania. John Albert died early, and Alexander succeeded him as king of Poland (1501–06) through the Union of Mielnik (1501). Poland and Lithuania were thus united again. A statute of 1499 called the combined kingdoms a "republic" for the first time, terming the Union of Mielnik an indivisible body. The social, economic, and constitutional union became even further consolidated under the reign of the fourth of the Jagiellon brothers (Casimir IV's sons), King Sigismund I the Old (1506–48). Ruled by the eldest of the Jagiellons, Władysław (1491–1516) and his son Louis II (1516–26), the combined Kingdom of Bohemia and Hungary remained only loosely connected with the unified Kingdom of Poland-Lithuania.

Union of Mielnik

The Ottoman incursions. The Ottoman Empire directed its attacks mainly against the unified states of the Jagiellons, conquering the commercial trading cities of Kaffa (1475), Kilia, and Akkerman (1484). This meant that Poland was now cut off entirely from the trade with the Orient. In addition, Poland found yet another enemy in the Tatars in the Crimea. These had come under Ottoman rule in 1475, and from 1492 they undertook pillaging and slave raids into Red Ruthenia and Little Poland. The Tatars of the Crimea pushed into Polish territory with the aid of the Ottomans, reaching Cracow in 1498–99. Now Moldavia had also come under the influence of the Ottomans and Tatars. John Albert, the brilliant Polish commander in chief, had always been victorious against the Tatars, but he suffered a heavy defeat at the hands of the Moldavians in the forests of Suceava in 1497.

Russian invasion. Tsar Ivan III the Great took advantage of the situation, and, allying himself with the Tatars of the Crimea, he attacked Lithuania twice (1486–94 and 1500–03), capturing all the principalities on the Oka and Desna rivers as far as the middle of the Dnieper River. Another ally of Russia, the German emperor Maximilian I, laid claim to Bohemia and Hungary and prompted the states of the Teutonic Order to rebel against Poland. The result was that the grand master Frederick of Saxony (1498–1510) refused to swear his oath of allegiance, and his successor, Albert (Albrecht) of Hohenzollern, a cousin of the elector of Brandenburg, went so far as to enter into an alliance with Russia (1517). Poland supplied armed help to Lithuania, but in spite of this Russia took the city of Smoleńsk (1514) in the course of two wars (1507–08 and 1512–22). After this, at the First Congress of Vienna (1515), Maximilian changed sides, renouncing the help of both Russia and the Teutonic Order. Instead, he entered into a mutual hereditary agreement with the Jagiellon kings with respect to the succession in Bohemia and Hungary.

Rebellion of the Teutonic Order

Alliance with the Turks. The Ottoman troops advanced all the way through Little Poland (1524), under the leadership of their Sultan Süleyman I the Magnificent, and at the first Battle of Mohács (1526) they conquered the greater part of Hungary, killing King Louis II, the last Hungarian Jagiellon. The Ottomans next posed a great threat to Vienna (1529), and Poland found it wise to conclude a perpetual peace agreement with them in 1533. This did not, however, prevent the Muslims from threatening Poland's southern boundaries from bases in Hungary—a threat that was to last for about 140 years.

Re-establishment of Polish security. Renewed Moldavian and Russian attacks were repelled by the Polish troops under the brilliant leadership of Jan Tarnowski. Sigismund I the Old was able to incorporate the Duchy of Mazovia into his empire after the death of the last of the Piasts in Mazovia (1524–29). His campaign from 1519 to 1522 also secured Polish sovereignty over Prussia. Albert of Hohenzollern became the first secular duke of Prussia to swear an oath of allegiance to Sigismund (1525), and the land came to be known as Ducal Prussia.

Ducal Prussia

Sigismund II Augustus, 1548–72. The son of Sigismund the Old, Sigismund Augustus, was the last Polish Jagiellon. Trying to insure his country against the threats from the Ottoman and Russian empires, he sought alliances with the emperor Charles V and the Roman king Ferdinand I (1549). This at least brought him Charles's final renunciation of any intervention on his part in settling the question of the constitutional position of both Royal and Ducal Prussia. For 24 years Tsar Ivan IV the Terrible had led a war of conquest in order to obtain the remains of the lands of the Teutonic Order in Livonia (1558–82), and he succeeded in occupying part of the territory. This prompted Sigismund Augustus to enter into an agreement (the Union of Wilno [1561]), with the last grand master of the Order, Wilhelm Ketteler, integrating Livonia directly with Poland while Courland (now a secular duchy) had to recognize the authority of Poland. This antagonized Russia, which invaded Poland and conquered the lands of Połock on the Dvina River in the first Nordic War (1563–70). Under pressure of the threat posed by Russia and also because there was no heir to the throne, the Sejm combined the former crown lands, Lithuanian, Royal Prussia, and Livonia in the Union of Lublin, creating a unified and indivisible republic (1569).

Unification of the kingdom. The Polish population now began to integrate the nobility, the burghers, and a large part of the peasantry in the Lithuanian–Ruthenian

regions in the east. From about 1420, but in particular between 1540 and 1580, the people of densely populated Mazovia began to migrate. Resettling, they colonized the enormous virgin forests in the west of Lithuania— particularly Podlasie. Moving south, they also settled in the Duchy of Prussia. As early as 1582, there were more than 150,000 Jews among the many nationalities of the Polish Commonwealth, more than in all the rest of Europe. The two main parts of the country that comprised the Polish Republic—the crown lands and Lithuania—each retained its own ministers, treasury, and army. Unification was finally accomplished by the consolidation of a common elective kingdom with a functioning Sejm, or Diet, consisting of two houses. The upper house seated about 100 senators, and the lower house approximately 150 delegates. This arrangement created a Royal Republic in 1569, with a constitution that was in many ways similar to the constitutions of England and Hungary, and it remained basically unchanged for the next 226 years.

Establishment of the Sejm

Constitutional change. Between 1493 and 1569 the Polish Estate Kingdom slowly evolved into a parliamentary monarchy, which was governed almost entirely by the nobility. The Sejm of Piotrków, in 1493, was the first of about 200 diets that would be convened in the course of the next 300 years. From the assembly of the Sejm of Mielnik in 1501, the king was regarded as the president of the Senate, and in 1505 the constitutional act Nihil Novi ("No innovations") decreed the making of laws to be the sole right of Parliament. The supreme authority of the aristocracy led to a constantly growing opposition from the ranks of the lower nobility, who were incensed at the aristocracy's virtual escape from the burdens of taxation. The lower nobility at last secured their own appointment to the supervision of the highest courts of appeal and secured the formal recognition of their equality with the aristocracy, together with the prohibition of titles (*e.g.*, duke, prince, margrave, count). The superiority of the aristocracy over the lower nobility continued, however, since the aristocracy still remained in possession of the wealthiest estates of the country, while the lower nobility were very often poor and propertyless.

Trade and wealth. The "Golden Age" of Poland expressed itself in the wealth of the country and a favourable boom in its foreign trade. Raw materials such as grain, lumber, cattle livestock, hemp, flax, and potash showed a lively export trade, especially to the Netherlands and to England, by way of the autonomous free republic of Gdańsk (Danzig), which controlled about 80 percent of Poland's foreign trade on the basis of its monopoly (*jus stapulae*). Second only to Amsterdam, Gdańsk became the most important trading city of central and eastern Europe by 1576. The production of raw materials was in the hands of the propertied Polish nobles, now turned into farming entrepreneurs.

Importance of Gdańsk

The exploitation of the peasantry. The nobles extended their previous manorial domination over the peasant class with statutes passed in 1496, 1518–20, 1532, and 1543 that reduced the peasants to virtually the same status as property; they came under the jurisdiction of the nobles and, as a result, lived under conditions of virtual slavery as cheap labour for the nobles' farmstead economy. The discovery of the Americas paved the way to the second feudal system in Poland, made possible by the imports of gold and silver from the American continent. In contrast, the same era led to the establishment of modern manufacturing plants and industrial enterprises in western Europe, ushering in the period of early capitalism.

Just as in other countries, the wealth of the Golden Age led to a gradual currency inflation in Poland. Rather than being invested in industrial developments, the money was spent on luxury items for the nobles and, above all, on grandiose building constructions.

Cultural developments. The Renaissance and the Age of Humanism set their stamp for ever after on the spiritual and intellectual life in Poland. The art of book printing (in Cracow alone there were eight printing presses by 1580) and general school education spread. As many as one-quarter of the Polish population were able to read and write by 1550–80. Young Poles studied at major foreign universities, particularly at Padua, at Bologna, and at the Sorbonne, as well as at the Protestant universities of Wittenberg, Leipzig, Königsberg, Tübingen, and Basel. The department of astronomy at the University of Cracow achieved world renown through Nicolaus Copernicus (Mikołaj Kopernik; 1473–1543) and his work on the motions of the planets. Polish literary language came into its own, supplanting the use of Latin alone by scholars. The first Polish books were printed between 1514 and 1524, and the first Polish translation of the Bible appeared in 1534. Although still partly German, the burgher class was now integrated completely into the cultural mainstream of Poland. It was also of the greatest significance that the political science philosophy of Greco-Roman antiquity found general acceptance at this time, inspiring such works on the theory of political science as Andrzej Frycz-Modrzewski's *De Republica Emendanda* ("Of the reform of the Republic"; 1551), which was translated into almost every language spoken in Europe.

High level of education

The Reformation. The Reformation came to Poland from Germany, Italy, France, and Bohemia between the years 1523 and 1526. Lutheran and Calvinist denominations were established alongside the originally Hussite Bohemian Brethren and the Anti-Trinitarians, or Arians (from about 1562–65). Under the spiritual leadership of Laelius Socinus (1525–62) and Faustus Socinus (1539–1604) the Arians strove for both a rational secularization of religious life and a radical change in society, including the liberation of the peasantry. In spite of the first suppressive edicts of 1520–26, the Reformation gained strength, particularly after the Sejm suspended the secular execution of ecclesiastical sentences (1552). After this, Poland was the only country in Europe without religious persecutions for about the next 130 years. The Lutheran, Calvinist, and Hussite denominations consolidated in 1570 at the Concord of Sandomierz, but this soon disintegrated because of dogmatic differences. The Compact of Warsaw in 1573 removed the Catholic Church from its dominant position, introducing the constitutional equality of all religions—the first enactment in Europe of complete religious tolerance. The Reformation had affected about one-sixth of all parishes in the Polish and Lithuanian regions by 1562–91. The Eastern Orthodox Church was hardly touched by the Reformation, which received its greatest support from the nobility and the educated burgher classes, while the lower middle class and the peasantry in general remained Catholic. The effects of the Reformation persisted only among the German population of Great Poland and that of Royal Prussia (Gdańsk), in addition to a few larger cities, such as Warsaw and Wilno.

THE "SILVER AGE" OF THE ROYAL REPUBLIC, 1572–1648

The Interregnum and reform of the monarchy. *Constitutional reform.* The Interregnum, or period without a ruler, lasted from 1572 to 1575. During this time the republican constitution was further improved and extended. The election of Henry III of Valois (1573–74)—King Henry III of France from 1574—introduced and established the basic rule that not only the council of the aristocracy but also all nobles and squires should elect the king. This meant a formal victory for the lower nobility and the founding of the Polish rule by the nobility. The fundamental basis of the constitution, the "Henrician Articles" of 1573, required the king to swear an oath reducing his authority to no more than that of a contract partner in a sovereign nation of nobles. The king was also obliged to convoke the Sejm every two years; and in the periods between the sessions he was to be advised by a group of senators. Should the king fail to observe any of these articles, the nation was ipso facto obsolved from its allegiance.

Stephen Báthory (Stefan Batory), 1576–86. The next elected king was Stephen Báthory (Stefan Batory), who had defeated his opponent, the emperor Maximilian II. During his reign the leading statesman Jan Zamoyski

rose to the positions of grand chancellor (1578) and grand hetman (commander in chief; 1580). Zamoyski had been instrumental in gaining the victory for the lower nobility during the Interregnum. Now, Stephen Báthory and Zamoyski defeated Tsar Ivan IV with three brilliant campaigns (1578–81), forcing the Russians to return the occupied territories of Livonia and Połock by 1582.

Successful campaigns against Tsar Ivan IV

Sigismund III Vasa, 1587–1632. Bàthory was succeeded by Sigismund III Vasa, who defeated Maximilian of Austria. Since he was also the legitimate king of Sweden —as the son of John III Vasa of Sweden and of Catherine, daughter of Sigismund I—Sigismund established a brief Polish–Swedish union, but this soon fell apart as the result of Swedish-Lutheran opposition to the Catholic faith of their king.

The Swedish–Polish War of Succession. Revolting against Sigismund III in 1599–1600, Sweden pursued the Swedish–Polish war for 35 years, but Sigismund never relinquished his demands for the Swedish crown. The Polish grand hetmans Zamoyski and Jan Chodkiewicz were at first able to repel the attacks of the Swedes against Livonia (1600–09). They won a famous victory in the Battle of Kirkholm (1605), although outnumbered four to one by the Swedes. In the meantime, a Polish volunteer corps helped the False Dmitry to gain the tsar's throne in 1605–06. Sigismund found it necessary to openly intervene in Russia (1609–13) when a coup d'etat took place there. Under the leadership of Stanisław Żółkiewski, the Polish army won a brilliant victory at Kłuszyn (1610) against a sixfold Russian–Swedish majority. Polish troops occupied Moscow (1610–12), but they were driven out by the new tsar, Michael Romanov (1613–18). The important fort of Smoleńsk remained with Poland (1618–54).

The Thirty Years' War. At the outbreak of the Thirty Years' War (1618), Sigismund remained officially neutral but favoured the Austrian side. The Prince of Transylvania had advanced and occupied Vienna, but a decisive victory by the Polish army in 1619 prepared the way for the Austrian victory at the Bila Hora (Weisser Berg, or White Mountain) in 1620. Although the Poles won a decisive victory against the Turks also at the Battle of Chocim on the Dniester River in 1621, the Turkish attacks did not cease for another decade. Taking advantage of this Polish predicament, Sigismund's cousin, Gustavus II Adolphus of Sweden, occupied Livonia and its important Baltic seaport Riga (1621). At the Peace of Altmark, the Swedes temporarily gained possession of Livonia as well as of some Royal Prussian port cities (1629–35).

Swedish occupation of Livonia

Władysław IV Vasa, 1632–48. A highly talented field marshal, Sigismund's son Władysław IV (1632–48) assured the existence of Poland when he repelled the renewed attacks of the Turks, Russians, and Swedes (1633–34). He also reconquered, in 1635, the Prussian ports on the Baltic Sea that Sweden had occupied. But his plans to establish a Polish navy that would hold its own against the Swedish domination of the Baltic, to regain sovereign rights over Livonia (1639), and to prepare for an extensive war against the Ottoman Empire (1646–47) were all frustrated by the disinclination of the Polish nobles to wage wars. They wanted peace at any price in order to take advantage of the splendid opportunities opening up on the European market in the wake of the Thirty Years' War.

The Cossack revolt. The prominent powers in northern, eastern, and southeastern Europe had attacked Poland constantly and had forced the republic into repeated compromises on the question of sovereign rights over the Ukraine and the Duchy of Prussia. Meanwhile, the nomadic Ruthenian–Eastern Orthodox farmers—the Cossacks—had established themselves in the Ukraine, and their number was steadily growing with the addition of escaping peasants from Poland and Russia. The Polish landed estate owners could hardly keep them under control over the vastness of the barely populated steppe that the Cossacks inhabited. The Cossacks eventually rebelled against the constant attempts on the part of the Polish estate owners to reduce them to the state of bonded peasants, and their leaders demanded a position of equality with the Polish nobility. The Cossacks remained the foremost social danger in the republic, allying themselves with the Tatars in 1648.

Poland's access to the Baltic endangered. In 1618 John Sigismund, elector of Brandenburg (1608–20) also became duke in Prussia, and thus the two crowns were now in personal union, though the sovereignty over Ducal Prussia remained with the Polish crown. His grandson, however, the Great Elector Frederick William (1640–88), by clever manoeuvring between Sweden and Poland, succeeded in removing Ducal Prussia from Polish sovereignty by the Treaty of Welawa (Wehlau) in 1657. From then on, the Brandenburg state of Prussia became the most dangerous of Poland's neighbours, not least because Frederick William intended to establish a territorial connection between Western Pomerania and the Duchy of Prussia, henceforth a leading aim of Brandenburg–Prussian politics until the First Partition of Poland about 115 years later. It threatened to cut off Poland from its connections with the Baltic Sea.

The threat from Brandenburg

Economic prosperity. The booming years between 1577 and 1654 became the high point in the history of Polish economics. At that time, Poland was considered, appropriately, the "granary of Europe." Polish grain was exported to France, England, Spain, and sometimes even to Italy and the Near East via Amsterdam, where the most important grain stock exchange was located. Gdańsk soared in importance to the rank of a trade city of world renown. As a result of this foreign-trade boom, the landowners pushed the farmsteading economy to its very limits, developing estates into tremendous areas. In many cases the servitude of the peasantry become actual slavery. This caused a number of small rebellions by the peasants, many of whom escaped into the Ukraine.

The liberum veto. The economic power of the magnates encouraged the forming of groups of supporters among propertyless voters who would keep their votes at the disposal of one particular aristocrat, thus enabling the magnate to command entire districts of the provincial *sejm* and, if necessary, even to paralyze it. In this way the original, formal "democracy of nobles" based on about 150,000 voters had evolved into an oligarchy of about 300 families. This was to prove a decisive factor in the later disintegration of the Polish republic. The undermining of the power of the *sejm* culminated in a procedure that made it possible for just one delegate to disrupt it through the liberum veto. The liberum veto was based on the assumption of the political equality of every Polish gentleman, with the corollary that every measure introduced into the *sejm* must be adopted unanimously. Consequently, if any single deputy believed that a measure already approved of by the rest of the house might be injurious to his constituency, he had the right to exclaim "Nie pozwalam," ("I disapprove"), the measure in question falling at once to the ground. All efforts toward increasing the power of the central government were held to be against traditional freedom and were opposed. This principle was extended still further until a deputy, by interposing his individual veto, could at any time dissolve the *sejm*, and all measures previously passed had to be resubmitted for the consideration of the following *sejm*. The liberum veto was used for the first time during the *sejm* of 1652, but before the end of the 17th century it was used so recklessly that all business was frequently brought to a standstill. It thus became the chief instrument of foreign ambassadors or native magnates for dissolving inconvenient sessions, as a deputy could always be bribed to exercise his veto.

The Counter-Reformation. Poland experienced the effects of the Counter-Reformation after the Council of Trent (1545–63). Led by prominent scholars such as Cardinal Stanislaus Hosius (Stanisław Hozjusz), who introduced the Society of Jesus, or Jesuits, in Poland in 1564, a total of approximately 50 Jesuit establishments were founded in the republic between 1564 and 1654. The Jesuit colleges in Połock, Smoleńsk, and Černigov and the Jesuit Academy of Wilno (founded in 1578–80)

The Jesuits

were mainly intended for the conversion of Eastern Orthodox Russia to Catholicism. Through their colleges and schools, the Jesuits established a dominant influence on the education of the Polish people that would last for about two centuries. Piotr Skarga was the most prominent theologian and politician among the Jesuits. The Order of the Camaldolese also supported the Counter-Reformation. The Counter-Reformation led to religious disorders; many of the Protestant churches were closed, especially in the large royal cities (particularly between 1574 and 1612); and the number of Protestant parishes diminished, as many of the aristocratic dynasties reverted to Catholicism. By 1650 there were not more than 270 non-Catholic parishes left in the country. Only the established German–Lutheran Church in the city republics of Prussia (Gdańsk, Elbląg) and the Calvinist Church in Lithuania (approximately 140 churches) retained a certain importance. The most significant result of the Counter-Reformation was the Synod of Brześć in 1596, whereby six of the ten Eastern Orthodox eparchies acknowledged the authority of the Holy See in Rome. Declared illegal and schismatic in 1595, the Eastern Orthodox faith was again recognized in 1635. Its spiritual centre became the Academy of Kiev, which had been modelled after the colleges of the Jesuit order. The Protestant academies of Gdańsk and Leszno still retained their European distinction, however, and John Amos Comenius, the philosopher and pedagogue, taught in Leszno.

THE IRON ERA PERIOD OF WARS AND DISINTEGRATION, 1648–97

John II Casimir Vasa, 1648–68. The reign of the last of the Vasa kings in Poland, John II Casimir, is remembered in history as a period of bloody wars and disasters.

The Cossack–Russian invasion. The epoch began with a rebellion of the Ukrainian Cossacks, led by Bohdan Chmielnicki, who heavily defeated the Polish troops in 1648 and 1651, establishing their own military state (1651–54) with the armed assistance of the Tatars of the Crimea and the Ottomans. In the Union of Pereyaslavl, in 1654, the Cossacks acknowledged the authority of the Russian tsar Alexis Mikhaylovich. On the strength of Cossack aid, Alexis attacked Poland, conquering almost the entire eastern part of the country, as far as Lwów, in 1654–55.

The Swedish invasion. Charles X Gustav of Sweden took advantage of the Polish predicament and occupied the Polish and Lithuanian lands of the republic in 1655, receiving some support from the native aristocrats and from Frederick William of Brandenburg. John Casimir fled to Silesia, and the majority of the nobles (among them the future king John III Sobieski) recognized Charles as their king, particularly after he had promised to reconquer the eastern part of Poland from the Russians and Cossacks. The extraordinary brutality of the Swedish army, however, gave rise to a general revolt against Charles, and the Swedes were not able to take the two largest cities of Poland (Gdańsk and Lwów). The brilliant grand hetman Stefan Czarniecki eventually destroyed the Swedish and defeated Frederick William's Brandenburg army between 1655 and 1657. In the treaties of Welawa (Wehlau), Bromberg (Bydgoszcz), and Oliwa (Oliva, 1660), Poland recognized the "superior power of government" of Frederick William over the Duchy of Prussia but reserved the right to reversion with regard to the duchy and did not relinquish this right until 1773.

Flight of John Casimir

Settlements and social consequences of the wars. The Ukraine again entered into a union with Poland in 1658, and Russia was decisively defeated in 1660–62. Nevertheless, Russia obtained Smoleńsk and the eastern half of the Ukraine in the truce of Andruszów (Andrusovo; 1667), after the rebellion of Jerzy Lubomirski against the King of Poland. Having been instrumental in bringing about the greatest catastrophe in the history of Poland, John II Casimir finally abdicated in 1668. The war, the bubonic plague, slave raids, and mass murders had reduced the total population to approximately 4,000,000 people, or 45 percent of the former total population. (This was about twice the number lost in World War II.) The cities suffered worst of all, losing all manufacturing establishments and industrial enterprises along with three-quarters of all their inhabitants. The tremendous costs of the war caused the greatest inflation in the history of Poland. Religious feeling was fanatically inflamed since the war had been conducted in the spirit of a religious struggle, and the Protestant parishes almost totally disappeared. The constitutional maxim of religious tolerance was finally abandoned in 1648–58.

Michael (Korybut) Wiśniowiecki, 1669–73. The short reign of Michael (Korybut) Wiśniowiecki was dominated by internal fighting between the Habsburgs, led by the King, and the French-oriented faction that had gathered around John Sobieski, a military commander of rising fame. The Cossacks seceded again in 1667, allying themselves with the Ottoman Empire (1668), and in the succeeding Turkish attack in 1672–73, the Poles lost Podolia and two-thirds of the Ukraine.

John III Sobieski, 1674–96. A brilliant victory in 1673 established the fame of John Sobieski as the "vanquisher of the Turks," a title that brought him the crown of Poland in 1674. In spite of his many successes in the fight against the Ottomans (1675–76) and despite his liberation of Vienna in 1683—a crucial factor in the development of European history—John III Sobieski was not able to regain the regions in the southeast of the republic held by the Ottomans. He also failed in his attempts to secure the return of the Duchy of Prussia to Poland through treaties with France (1675) and Sweden (1677). Poland's participation in the Holy Alliance against the Ottoman Empire (1684) and the many Turkish campaigns of John III Sobieski (1684–92) eventually aided the development of Austria and Russia as new great European powers. With the Peace of Carlowitz (Karłowice; 1699), Poland again received Podolia and the Ukraine west of the Dnieper River. Given to Russia for two years in 1667 but never returned to Poland, the city of Kiev was recognized as part of Russia in the "perpetual peace" in 1686.

Peace of Carlowitz

THE SAXONIAN ERA AND THE RUSSIAN PROTECTORATE, 1697–1763

Augustus II the Strong, 1697–1733. On the death of John III, no fewer than 18 candidates for the vacant Polish throne presented themselves. The successful competitor was Frederick Augustus I, elector of Saxony, who renounced Lutheranism for the crown. Under his reign as Augustus II the Strong, Poland experienced a new wave of wars and catastrophes. The planned union between the industrially developed Saxony and raw material-producing Poland would have been advantageous for the economy of both countries, but that did not materialize since the King, as ambitious as he was lacking in talent, became involved in a war with Sweden through his alliances with Peter I the Great of Russia and King Frederick IV of Denmark.

The Northern War. These allies attacked Sweden and began the Northern War (1700–21). The Swedish king Charles XII repelled Denmark and Russia in 1700 and marched from Livonia into Poland (1701–02), the latter serving him as the base for his future military operations. Appointing Stanisław Leszczyński (Stanisław II) as puppet King (1704), Charles XII invaded Saxony in 1706 and forced Augustus to abdicate. During the civil war of 1704–10, Swedish troops totally devastated Poland, and the Saxonian and Russian armies fared little better. Cultural losses were enormous, and one-third of all Polish cities lay in ruins. Although by 1702 the population had recovered again, reaching about 8,000,000, war and then the bubonic plague of 1709 took their toll, and other disasters reduced the population to about 6,400,000. Poland was spared any further Swedish occupation when the Swedish troops advanced into the Ukraine, where they suffered a decisive defeat at the Battle of Poltava (1709). Stanisław Leszczyński escaped from Poland in 1710, and Augustus II again assumed the crown with the particular support of Peter the Great of Russia. Breaking all agreements with Poland, the Tsar annexed Swedish-held Livonia (1710), the very land that had given Augus-

Civil war of 1704–10

tus II the pretext to start the war. Then the Tsar turned east and put down the last Cossack rebellion in the Polish Ukraine (1702–04), which the Polish Sejm had provoked by despotically deciding to disband the Cossacks in 1699.

The Prussian situation. The Hohenzollern prince Frederick III had assumed the title of king of Prussia in 1701, with consent of Augustus but without the approval of the republic, which alone had the power to grant this title. In 1715 King Frederick Willaim I tried to acquire the Polish Prussian territory, again with the approval of Augustus. The intervention of Peter the Great, who resided in Gdańsk at the time, frustrated this attempt in 1716, and the Russian troops remained in control of the southern coast of the Baltic Sea. Peter the Great thus assumed the role of protector of the territorial integrity of the Polish republic. From then on, Poland was to all intents and purposes a protectorate of the Russian government, and this relationship lasted until the dissolution of the republic in 1795.

Augustus III, 1733–63. The French-supported candidate, Stanisław Leszczyński, was elected king for the second time after the death of Augustus II in 1733. The Russian and Saxonian armies interfered, however, forcing the election of his opponent, Augustus III, the son of Augustus II. It was in his reign that Prussia definitely and conclusively achieved the position of a European power under its king, Frederick II the Great (1740–86). He annexed Silesia between 1740 and 1745, securing control of the western boundaries of Poland as well as of Polish foreign trade at the points where it was shipped overland to central and western Europe. Inhabited mainly by Poles, the southeastern half of Silesia still retained close political and cultural ties with Poland, but the Prussian annexation ended the church sovereignty of the metropolitanate of Gniezno over Silesia.

The Seven Years' War. During the Seven Years' War (*q.v.*; 1756–63), the Russian army used the territories of Gdańsk, Toruń, and Poznań as operational bases in their war against Prussia. Frederick II ordered the counterfeiting of Polish money, a measure that severely damaged and inflated the monetary system of the republic.

Social, religious, and political change. During the long and peaceful era between 1716 and 1768, the population grew rapidly, due to both the steady immigration of Germans and the mass exodus of an enslaved peasantry from Russia. By 1772 Poland's population numbered approximately 11,420,000. Comprising more than one-half of all the Jews in the world, the Jewish population of Poland lived in all parts of the country. The Counter-Reformation reached its zenith under the spiritual leadership of the Jesuits, who trained 20,000 students of noble birth annually in their 51 greater colleges and other educational institutions. In 1717 and 1733, laws were passed that restricted the spiritual and secular rights of the dissident (non-Catholic) population. Popular education fell to its lowest level. Spreading through all the other parts of Europe, the Age of Reason affected Poland only by virtue of some individual achievements. Parliamentary life came to an almost complete standstill. Under the reign of Augustus II only four of a total of 13 *sejms* came to an orderly conclusion; under Augustus III only one out of 13 did so, as the result of the use of the liberum veto.

Population growth

REFORM AND PARTITION DURING THE REIGN OF STANISLAW II AUGUSTUS PONIATOWSKI, 1764–95

A former diplomat, Stanisław II August Poniatowski gained the crown of Poland with the help of Russian troops in 1764. Aided by his relatives, the Czartoryski, he tried to introduce political reforms, among other things the temporary abolition of the liberum veto in 1764–66, but his plans failed because of the objections of Catherine II the Great, the Russian empress. Taking advantage of the religious quarrels in Poland, the Empress sent Prince Nikolai V. Repnin as Russian minister to Warsaw with instructions that led to further discord in Polish affairs. Of Poland's population of 11,500,000, about 1,000,000 were dissidents, one-half Protestant and one-half Orthodox. For these people, who had been largely deprived of their rights by the laws of 1717 and 1773, Repnin, in the name of the Russian empress, demanded absolute religious and political equality with the Catholic population of Poland. He was well aware that an aristocratic and Catholic assembly like the Sejm would never concede such a demand.

Intervention by Catherine the Great of Russia

Early in 1767 a confederation was formed at Repnin's instigation to send a deputation to Catherine, petitioning her to guarantee the liberties of the republic. Subsequently, but not without a stubborn resistance, the Sejm accepted Catherine's authority, and the so-called fundamental laws were enacted, guaranteeing the liberum veto and all the other ancient abuses as unalterable parts of a Polish constitution ensured by Russia. All the restrictions against the dissidents were repealed at the same time.

Confederation of Bar. These events led to a Catholic uprising known as the Confederation of Bar, aided by the Turks. After four years of fighting (1768–72), the Russian troops were able to gain the upper hand over the confederates. Frederick II of Prussia tried to use this civil war to achieve a partition of Poland, but he met with the resistance of Catherine II. Only the threat of Austrian troops advancing against Russia and the Austrian annexation of some Polish territory in the Carpathian Mountains (1770) forced Catherine to seek help from the Prussian king on his terms—namely, the partition of Poland.

First Partition, 1772. With the First Partition of Poland in 1772, the republic lost about 28 percent of its territory, the Livonian and White Ruthenian regions north of the Dvina River and east of the Druć River falling to Russia. Austria received Little Poland south of the Vistula River and almost all of the Red Ruthenian territory under the name of the Kingdom of Galicia. Prussia obtained the smallest but most valuable property, namely, Royal Prussia, comprising Warmia (Ermland) and a part of Great Poland on both sides of the Noteć River, though without the cities of Gdańsk and Toruń. Through this annexation Prussia had not only acquired the desired land connection between Western Pomerania and East Prussia that it had tried to secure for almost 120 years, but it had also gained control over approximately four-fifths of the total foreign trade of Poland. By levying enormous custom duties on the Polish foreign trade going via this route, Prussia made this its most important source of income. In fact, the partition of Poland was of vital importance to Prussia, for it was now able to shake off the last vestiges of the old Polish sovereign rights. The treaty of 1773 established the full and complete sovereignty of Prussia over its lands.

Prussian control of foreign trade

The period of reform. The shock of the First Partition caused political and economic reforms in Poland—as far as Russia would permit. The taxation system and the army were thoroughly reorganized, and with the Permanent Council, in 1775, Poland installed its first real central government. Many landed estate owners carried out individual peasant emancipations, and almost 200,000 Germans and at least 300,000 Russians emigrated to Poland, a land that was considered free. Slowly, the cities began to flourish again, and the beginnings of an industrial revolution could be observed with the rise in mining activities, the advances of the textile industry, the founding of the first joint-stock companies, and the employment of the first machine equipment. Polish intellectual pursuit and reasoning received fresh impetus when the Age of Reason was finally accepted. The King promoted the political theories of such men as Edmund Burke and George Washington, while the educated Poles advanced the political theories of Montesquieu and Rousseau. But whatever the theory, the concept of modern democracy, including all ranks and classes of society, quickly became established in Poland. The first governmental department for public education in Europe was established in conjunction with the appointment of an education commission in Poland in 1773. Its task was the reorganization of the universities of Cracow and Wilno, as well as the supervision of approximately 80 gymnasiums. The printing of books and the publication of magazines and journals also flourished.

Improvements in education

The Constitution of May 3, 1791. The first codified constitution in Europe since antiquity and the second in



The partitions of Poland, 1772–95.

the world after the United States, the new Polish constitution passed by the "Four Years Sejm" embodied the following ideas: first, the precept of a "people's sovereignty," which included the nobility as well as the metropolitan bourgeoisie; second, the constitutional separation of powers between the executive, legislature, and judiciary; and third, the responsibility of the cabinet to parliament. The liberum veto and all the obstructive machinery of the anomalous old system were abolished. According to its own rules, this constitution had to be revised after 25 years at the latest. But Catherine II considered such a constitution dangerous to the existence and continuation of her own autocratic governmental system; she therefore ordered her troops to invade Poland in 1792, and the reforms were destroyed by force.

The Second and Third Partitions. The Russo–Prussian treaty of 1793 (the Second Partition) placed almost all the Red Ruthenian and White Ruthenian lands under Russian rule, while Prussia received Gdańsk, Toruń, Great Poland, and part of Mazovia. The reaction was inevitable, and in 1794 a general popular insurrection broke out under the leadership of Tadeusz Kościuszko against this latest Russo-Prussian humiliation. Kościuszko achieved a partial liberation of the peasant class and began a general armament of the people, mobilizing more than 150,000 men. At first the Polish forces were almost universally successful; but ultimately Kościuszko was defeated by the Russian troops under Aleksandr Suvorov at the Battle of Maciejowice, and Warsaw capitulated at the end of 1794. Now Russia and Austria—joined later also by Prussia—agreed to the Third Partition of Poland in 1795. Russia annexed the entire territory east of the Niemen and Bug rivers, Austria took almost all of Little Poland, including Cracow, and Prussia received the remaining lands, including Warsaw. Stanisław II abdicated officially in 1795. For all official purposes, at least, this was the end of the Republic of Poland, which was eradicated from the map of Europe for the next 123 years.

Kościuszko

III. Poland under partition, 1795–1914

FOREIGN RULE AND THE DUCHY OF WARSAW, 1795–1815

The Russian sector. With the Third Partition of Poland in 1795, the largest part of the territory fell to Russia. Officially, Catherine II called these partitions a return of Russian territory, even though such a claim dated back 450 years. Proclaiming the unity of all Russians, the largest part of the Uniate Church returned to the fold of the Eastern Orthodox faith. In spite of this, the Polish constitution and culture remained fairly unaffected at first in these "Russian" parts, where at least 1,800,000 Poles, including over 640,000 nobles, still held positions of influence. Because Russia's level of civilization was still extremely low under the reign of the tsars Paul I (1796–1801) and Alexander I (1801–25), the government was unable to administer the area without Polish help. The land had been divided into eight governmental units in 1801; and since the Poles retained basic administrative control, the municipal rights, control of the Polish county courts and of the Polish Sejm remained in force for a number of years (1796–1831). An added factor was that this region had the highest standard of civilization, culture, and education within the realm of the Russian Empire—with the exception of the Baltic Sea provinces—and the fundamental principles of the Russian educational system were based on the essential features of the Polish educational institutions of this region.

The Prussian sector. The Prussian area consisted of the three provinces of West Prussia, South Prussia, and New East Prussia. Together with the province of East Prussia, which had not been under Polish rule since 1657, the country came to be known officially as the Kingdom of Prussia, a part of the Prussian monarchy. Of its population of almost 4,000,000, however, about three-quarters were ethnic Poles.

The Austrian sector. The Kingdom of Galicia, as the Austrian section was known, was repeatedly reorganized administratively. Most of the native intellectuals in both the Prussian and the Austrian sections were replaced by German administrators who governed the lands in their customary autocratic fashion. It is therefore not surprising that in these areas numerous nationalist secret societies were established.

Patriotic and conspiratorial societies

The struggle of the patriots. As early as 1768, Polish emigrant groups were formed in Hamburg, Dresden, Constantinople, and particularly in Paris. The group of Polish Jacobins organized in Warsaw and Wilno (1792–94) had established connections with the French Revolution, especially through Józef Sułkowski, who had politically supervised Napoleon Bonaparte in his capacity as secret police inspector (1796) of the French Directoire. Branching out from their Jacobin origins, the conspiratorial Society of Polish Republicans, formed in Warsaw around 1798–1800, and the Warsaw Society of the Friends of Science, founded in Warsaw in 1800, served partially as a cover organization for Jacobin activities. The centre of the conspiratorial network was in Paris, but it gained a particularly strong hold in Galicia after the French–Austrian War of 1792–1801. Gen. Jan Henryk Dąbrowski, who had gained his fame from the defense of Warsaw in 1794, established Polish legions in northern Italy in 1797, recruiting men from Austrian prisoners of war of Polish nationality. The legions fought without any political success against Austria between 1797 and 1801.

Russian rule. Prince Adam Jerzy Czartoryski, the Russian foreign minister, tried to re-establish the Polish Republic from the sector occupied by Russia. He was aided in this by the tsar Alexander I. The plan was designed to create a buffer zone against Prussia, then allied with France. The idea lost its rationale, however, when Napoleon I of France attacked Prussia in 1806 and Russia came to the aid of Prussia. The French advance against Poznań and Warsaw paved the way for a general Polish insurrection against Prussia, led mainly by Jan Dąbrowski and Józef Wybicki.

The Duchy of Warsaw. Taking some purely Polish areas of the Kingdom of Prussia, Napoleon I created the Duchy of Warsaw, placing it under the rule of the king of Saxony, Frederick Augustus (1807). The constitution dictated by Napoleon I and the introduction of the Code Napoléon called for the personal liberation of the peasants. But the soil remained fundamentally with the landed nobility, and therefore this liberation promoted rather than alleviated the agrarian population surplus in Central Poland, a condition that had existed already for about 300 years. Józef Poniatowski (1763–1813), the nephew of the last king, distinguished himself as field marshal during the war of 1809, when the Duchy of Warsaw acquired the Austrian part of the Second Partition, including Cracow, as well as Zamość, annexed with Galicia in 1772. Thus, the Duchy of Warsaw had by this time grown to almost the same size as the Prussian monarchy. The failure of Napoleon's war against Russia in 1812, however, meant that by 1813 the Duchy of Warsaw had come under Russian administration.

Government under Napoleon

THE KINGDOM OF POLAND AND THE OTHER POLISH LANDS, 1815–31

The Congress of Vienna (1814–15) established the territorial division of Poland (within the boundaries of 1772) until 1918. In accordance with the Prussian demands, the Grand Duchy of Poznań was detached from the authority of the Duchy of Warsaw and incorporated into the Prussian monarchy. Cut off from the Duchy of Warsaw as well, the "Free City of Cracow" was designated to be the "symbolic capital of the divided Poland," though these intentions came to an end when Austria annexed the city in 1846. Galicia remained an Austrian hereditary land.

The Kingdom of Poland. The main part of the Duchy of Warsaw was given the title of Kingdom of Poland in 1815. Approximately four-fifths of its population (over 4,000,000 in 1827) were ethnic Poles, of whom as many as 325,000 were nobles. This area thus came to be considered as the Polish nucleus. After the Congress of Vienna, in 1815, this Congress Kingdom of Poland, or Congress Poland, as it often came to be called, was united with the Russian Empire, though it retained its own governmental system, its own Sejm, and its own army. Tsar Alexander I was now in command of about three-quarters of all the historically Polish territories and, being not only the tsar of Russia but also the king of Poland, he was well advised to try to win the loyalty of the Polish people.

The constitution of the Kingdom of Poland

The constitution of the Kingdom of Poland, passed between 1815 and 1831, was considered the most liberal in all of Europe. The population of the Congress Kingdom of Poland included more voters than France, whose population was six times greater. A severe impediment, however, was the fact that Alexander I reigned with the harshness of an absolute monarch from 1819 on, breaking the constitution several times. This again promoted the creation of secret societies, the most consequential of which was the National Freemasonry (1819–25) under the leadership of Walerian Łukasiński. Similar secret societies were founded in Cracow, Poznań, and especially in the University of Wilno, where the Society of the Philomats functioned as a front for several conspiratorial groups. Even Adam Mickiewicz, the greatest Polish Romantic poet, became involved in the affairs of the Philomats and was sent into exile for some time. The secret societies (such as the National Patriotic Society) continued and flourished, despite repression at the hands of the Russians.

The Rising of 1830. Alexander's successor, autocratic Tsar Nicholas I (1825–55), planned a military campaign against France and Belgium after the revolution in France in July 1830. He was going to send a Polish army to accomplish this, but most of the Polish ensigns banded together to create the November Insurrection on November 29, 1830, under their leader Piotr Wysocki. Almost the entire army and the majority of all Polish politicians joined in the insurrection, and the Sejm proclaimed the ousting of Tsar Nicholas I on January 25, 1831, an act that led to a formal war between Russia and Congress Poland. The war lasted from January to September 1831, the superior Russian forces finally defeating the Poles, who received no response to their appeal for western European protection. About 6,000 leaders of the rising were forced into exile during 1831–32, most of them emigrating to France in what has become known as the Great Emigration.

RUSSIAN REPRISALS AND THE JANUARY INSURRECTION, 1831–64

Repression of Poland. The suppression of the November Insurrection meant heavy reprisals in the Russian as well as in the Prussian Polish areas, even though the latter had no actual part in it. Extensive confiscations of estates, deportations, and enforced recruiting were designed to break the national patriotism of the Poles. The universities of Warsaw and Wilno were suppressed, and Polish students compelled to go to St. Petersburg (now Leningrad) and Kiev in Russian Poland. The remainder of the Uniate Church was abolished and converted to Eastern Orthodox between 1831 and 1839, though continuing to exist in Congress Poland until 1875 and in Galicia until 1945. The Organic Statute of 1832 replaced the constitution, the Sejm, and the army in the Kingdom of Poland; and the land was united with the Russian Empire in a true union, although officially the kingdom retained a few autonomous rights.

The Organic Statute

Polish nationalist groups. The conservative group of emigrants led by Adam J. Czartoryski, the "king in exile," tried to solve the Polish question through legal and diplomatic means. His minimum goal was the restoration

of the legal status of 1815, which had been violated from the very beginning. The radicals had splintered into many factions, but the Polish Democratic Society founded in 1832 soon emerged as the leading organization. Its Manifesto of Poitiers of 1836 was less realistic than the program set up by the conservative side, because its revolutionary agenda demanded an independent Poland within the borders of 1772. But the Manifesto proclaimed the idea of a land reform without compensation to the present owners, and this drew many supporters. In the years between 1833 and 1846, emissaries and leaders of the partisans of the "left" established secret revolutionary cells in all parts of Poland. These cells were persistently traced down by the authorities, and thousands of their members were deported or imprisoned.

The 1846 insurrection. The Polish Democratic Society gained a certain influence in the country with the installation of a committee in Poznań between 1839 and 1846. This committee drew up a plan for a general revolution, but the Prussian police discovered it in 1846. In spite of this, a central government in Cracow, under the revolutionary leaders Jan Tyssowski and Edward Dembowski, tried to get an insurrection started. The uprising spread to Galicia, where Austrian troops soon put it down, at the same time inciting the local population so that many Galician peasants turned against the revolutionaries. The failure of the 1846 revolt severely weakened Polish participation when a general revolutionary wave swept through Europe during the early spring of 1848–49. Originally, Prussia and Austria had nominally supported Polish demands for independence, but now attempts by the Poles to establish a real independence were resisted by armed force in both countries. The course of this revolutionary struggle showed clearly that the fight was no longer only about the ideal of a unified republic based on the boundaries of 1772 but also about the Romantic Idealism of the time, which favoured individualistic, linguistically and culturally separated nationalities. On the other hand, nationalist ambitions now extended beyond the borders of 1772 for the first time. The Polish question also gained more impetus on the international scene. Adam Mickiewicz founded a new Polish legion in Italy and a little later published the moderate socialist newspaper *La Tribune des Peuples* in Paris (1848–49). The landed gentry of Poznań financed the *Neue Rheinische Zeitung* ("The Newspaper of the Rhineland"), the journal through which Karl Marx and Friedrich Engels in 1849 proclaimed the national liberation of Poland to be the most important task of the workers' movement of Europe.

Romantic revolutionary Idealism

Alexander II and the rising of 1863. The Eastern Question and the Crimean War (1853–56) caused violent upheavals within the Russian Empire, which motivated Alexander II (1855–81) to take a more conciliatory stand toward the Kingdom of Poland. A series of reforms were implemented, in their later stages by the autocratic Count Aleksander Wielopolski, who was installed in high office and stood for a national policy of loyal union with Russia. His methods and character lost him the support of the moderate gentry, however, and aroused the outright enmity of the young patriotic radicals, who were influenced by ideas expressed by Napoleon III of France and by the national reunification of Italy (1859–60). (The left wing of these radicals became linked with the Russian liberal movement and the international emigrants in London centred around Karl Marx and Aleksandr Herzen.) Religious ceremonies became occasions of political demonstrations, rioting broke out in Warsaw, and the situation culminated in the widespread insurrection of January 22, 1863, sparked off by Wielopolski's plans to press the revolutionary youth of the cities into the Russian army. The revolutionaries had relied on French and Austrian support, but these hopes proved to be in vain; and to make matters worse, Prussia again cooperated with Russia. As before, the ethnic groups in the Russian sectors—Poles, Lithuanians, and the Catholic White Ruthenes—united in a common war against the Tsar that lasted for 15 months, though the insurrection was finally crushed when the rebel "National Government" was detected by the Russian administrative authorities on April 10–11, 1864. The defeat was followed by executions, confiscations, and deportations, and all hopes for the Polish people to establish an autonomous national state were ended for the next half century. The Russian government did proclaim, however, the liberation of its peasants on March 13, 1861, and it extended this to the Kingdom of Poland on March 2, 1864. By so doing, it took the most important traditional premise for the revolution out of the hands of the revolutionaries. The agrarian policy, however, was expressly calculated to stir up ill feeling between the peasants and the country gentry, whom Russia was determined to punish in every way for their leading part in the insurrection. Yet the agrarian reforms accelerated the growth of towns and industry and the new social structure produced both financiers who exploited the great possibilities of the Russian market and a new working class, among whom the first socialist organizations began their secret activity.

Liberation of the peasants

THE FOUNDING OF MODERN POLAND, 1864–1914

After the failure of the January Insurrection, a policy of harsh repression fell upon the Poles remaining in the Russian and Prussian sectors. The Russian sector was now offically named the Western Region of the Russian Empire, and its governor general, Mikhail Nikolayevich Muravyov (who became known as the "Hangman of Wilno"), made every effort to stamp out Polish culture altogether.

Russification. The Kingdom of Poland was reorganized into the Land of the Vistula under a purely Russian administration. The University of Warsaw became a Russian university in 1869. Russification of all secondary schools was implemented during the years 1869 to 1874 and of all other public schools after 1885: the Russians very successfully accomplished their aim of keeping illiteracy among the common people at as high a level as possible. Consequently, by 1905 there were relatively fewer schools and students in the Land of the Vistula than there had been in 1815. Nearly all private schools and all social and political activities were prohibited.

Prussian and Austrian Poland. Following the Russian example Otto von Bismarck, the German chancellor, instituted a similar policy, and the Prussian sector was incorporated into the newly founded German *Reich*. In 1872 the entire school system was Germanized. The Colonization Law of 1886 tried to integrate German peasants into the native Polish population. Polish autonomy was able to achieve some success only in the Austrian sector, where its most important agency was the Galician Provincial Sejm (Diet). The Poles in Austria became the country's second nationality, and they were allowed to participate in the government at Vienna and retained sovereignty over the Ruthenes in Galicia. Galicia itself became the haven of Polish culture, with the University of Cracow, the University of Lwów, and the Academy of Sciences attracting many Polish intellectuals.

Galicia

Industrial growth and social change. The development of modern industry, notably the textile and metal manufacturing industries, as well as mining, made Congress Poland after the abolition of the customs barrier with Russia the most dynamic, industrialized part of the Russian Empire until about 1890. In Congress Poland alone there were more than 400,000 industrial workers by 1910. It was mainly Polish intellectuals who stimulated industrialization in the Russian and Austrian sections, whereas German intellectuals led the industrialization program of the Prussian part—particularly in Upper Silesia, with its 200,000 Polish miners. Due to the many agrarian crises between 1846 and 1892, most of the 3,200,000 descendants of the former nobility moved to the cities in the Russian sector, where they formed the majority of the liberal professions. In the Polish regions of eastern Germany, Congress Poland, and West Galicia, the intellectual classes were mostly without property and therefore dependent on wages, whereas the Poles in the Western Region of the Russian Empire and in East Galicia were mainly wealthy landowners or belonged to the city bourgeoisie. One of the main features of the period

was the mass emigration of Poles from their native lands into the more industrialized parts of Germany, a movement that was to continue for decades to come. Thus, by 1910 more than 100,000 miners of Polish descent were working in the industrial Ruhr district. The rural population of Congress Poland and Galicia was mostly illiterate, and poverty increased there to such an extent that at least 3,500,000 people emigrated between 1890 and 1914, mainly to the United States.

Mass emigration to Germany

Development of political parties. The first permanent political parties were established in Congress Poland and in exile between 1891 and 1893. The Polish League (LP), founded in 1886–87, was reorganized as the National League (LN) by Roman Dmowski in 1892–93 and later became the nucleus of National Democracy (ND), which was the strongest bourgeois party between 1905 and 1939. The Polish Socialist Party (PPS) was founded in 1892–93 under the patronage of Bolesław Limanowski and subsequently, under the leadership of Józef Piłsudski, became the strongest socialist party (with about 50,000 members) in the entire Russian Empire during the Russian Revolution of 1905. The left-wing socialist party (SDKP; from 1900 SDKPiL, Social Democracy of the Kingdom of Poland and Lithuania), founded in 1893 and led by Rosa Luxemburg and Junian Marchlewski, remained a minor party in comparison. The 1905 revolution, which swept over Russia, changed political life in all the Polish regions. The National Democracy, led by Roman Dmowski and supported by the majority of the landed gentry and the bourgeoisie, restricted its national platform to the demand for Polish autonomy. For this reason it was allowed to operate legally not only in Congress Poland but in Russian Poland, too, after 1905. The same party also controlled the Polish faction in the Russian Duma, or Parliament (1905–17). The PPS split in 1905–06: the patriotic wing under Józef Piłsudski and Tomasz Arciszewski retained its "radical program for independence," in accordance with the teachings of Friedrich Engels, and took part in the abortive revolution of 1905 against the Russian government; the left wing adjusted its program to an internationalist standpoint and so found it easy later on to merge with the SDKPiL, on the extreme left. The introduction of universal suffrage in Austria, including the Austrian sector of Poland (Galicia), for the parliament in Vienna in 1907 promoted the founding of the agrarian Polish Peasant Party (PSL). This party soon spread its influence to southern Congress Poland, but it split in 1913 into a moderate majority and a radical left. Only in Prussia was the repression of Polish cultural and political life continued, and after 1905 this policy was pursued in an even more severe form. This was the main reason why the Polish National Democrats regarded the German *Reich* rightfully as their main opponent from 1905 on.

IV. Modern Poland, since 1914

THE FOUNDING OF THE SECOND REPUBLIC, 1914–21

Events during World War I. At the outbreak of World War I, the governments of Austria and Russia tried to use the Polish question as a political weapon. On August 16, 1914, the Austrian government allowed the formation of a Supreme National Committee (NKN) in Galicia, and the organization of a Polish legion recruited mainly from the membership of the followers of Józef Piłsudski. The ultimate goal of Austrian policy was the incorporation of Congress Poland into Galicia. On August 14, 1914, the Russian government recognized the basic right of the Poles to autonomy, particularly in the lands that the Russian troops hoped to conquer. The Polish National Committee (KNP) was formed in November 1914—its chief members coming from the National Democrats, led by Roman Dmowski. The KNP supported the Russian war effort, and Russian forces conquered almost all of Galicia in 1914.

Piłsudski and Dmowski

Division of Congress Poland. The troops of the Central Powers—Germany and Austria-Hungary—were able subsequently not only to retake Galicia but to advance deep into Congress Poland. The Central Powers divided it into two military occupation zones, with a German governor general in Warsaw and an Austro-Hungarian one in Lublin. The political groups in both of these zones separated into those who believed in adopting a passive stance and the activists. With the majority of the PPS and the left-wing intellectuals on their side, the activists established the Central National Committee (CKN) in December 1915, their goal being to re-establish a Polish republic with the help of the Central Powers.

Creation of a Kingdom of Poland. At the request of the German commander, Gen. Erich Ludendorff, and the German governor general, Hans Hartwig von Beseler, the German and Austrian emperors proclaimed a new Kingdom of Poland on November 5, 1916. The kingdom consisted of both the occupied zones but comprised only about one-sixth of the territory established in 1772. The hopes of both monarchs for more support among the Polish volunteer troops in the Austro-German fight against Russia were in vain, for German aspirations did not match the expectations of the Poles, who had envisioned an eastward expansion of the Kingdom of Poland. Ludendorff, in turn, tried to establish what he himself termed "a national kingdom of Lithuania" in the eastern regions. All these elements now gave the first incentive to an internationalization of the Polish question.

The Russian Revolution. The Russian Revolution in March 1917 brought about the two Poland proclamations of the Soviet of Petrograd and the provisional government of Russia. These recognized Poland's right to independence, while the latter insisted on a military alliance with Russia. After this, three Polish army corps were formed from the ranks of the Russian army between May 1917 and May 1918, and France also recruited a Polish volunteer army. Founded by Dmowski on August 15, 1917, in Lausanne, the Polish National Committee (KNP) worked closely with the governments of the Western Allies. In the Kingdom of Poland, Piłsudski and the Polish Legion, now reorganized as the Polish Auxiliary Corps, with about 20,000 men, clashed with the Central Powers. Piłsudski and the majority of his corps were interned in July 1917. The grant of a limited Polish self-government in the Kingdom of Poland and the installation of a Regency Council on September 12, 1917, came too late to overcome the distrust that now divided the Central Powers and the Poles.

The Polish National Committee

The fight for self-determination. The October Revolution in Russia, as well as the proclaimed principle of the "right for self-determination of all nations" as stated by both Lenin and the U.S. president Woodrow Wilson (though both with very different interpretations), gave the Polish question international significance. President Wilson demanded the creation of an "independent Polish state" with a free and secured access to the sea in his well-known "14-points speech" of January 8, 1918. This program was ratified as an integral part of the future peace settlement by the governments of France, Great Britain, and Italy on June 3, 1918. The Central Powers used the "right for self-determination of all nations" in the peace negotiations of Brest-Litovsk (now Brest, Belorussian Soviet Socialist Republic) in February–March 1918 as a pretext for the separation of the Kingdom of Poland from the Russian Empire, but they apportioned the Chełm province, with its predominantly Polish population, to the Ukraine on February 9, 1918. This infuriated Polish opinion.

Re-establishment of the state. The military collapse of Austria in October 1918 led to the formation of a Polish Liquidation Commission in Cracow on October 28, 1918, which assumed power in Western Galicia. On November 7 a Provisional Government of the Polish Republic was established in Lublin. With the collapse of the German *Reich* in November 1918, Piłsudski was freed from his internment in Germany. Returning to Warsaw, he took control of the whole of Poland on November 11, 1918. The administrations for the regions of Cracow and Lublin also put themselves under his command. November 11, 1918, has thus been generally recognized as the day of the founding of the Second Polish Republic.

Piłsudski's return to Poland

The Second Polish Republic. To begin with, the Second Polish Republic consisted of the Kingdom of Poland

as it existed in 1916, the Duchy of Cieszyn, and West Galicia. Poland's losses during World War I numbered about 800,000 dead as well as the destruction of the majority of industrial plants, communications, and the serious disruption of the Polish monetary system and economy.

Internal developments. Piłsudski secured his position of power in the country by reaching an agreement, in January 1919, with the Polish National Committee (KNP) in Paris, headed by Roman Dmowski. The KNP was the recognized official Polish delegation at the Paris Peace Conference. Ignacy Jan Paderewski, a close associate of Dmowski, formed a government in Warsaw in 1919, but the executive power remained in the hands of Piłsudski, as the Polish "chief of state," from November 22, 1918, until December 14, 1922. The succeeding democratic elections led to the formation of a constitutional Sejm, which operated both as the constituent and legislative assembly (1919–22) after the passing of the "Little Constitution" on February 20, 1919. The strongest parties in the country were the National Democrats and the Peasants (PSL), the latter for the most part cooperating with the National Democrats. The left-wing parties commanded only about one-quarter of the votes. The Communist Workers' Party of Poland was founded on December 16, 1918, and Workers' Councils were set up in November 1918. The Councils were broken up by the government in June 1919, and this established and secured a "bourgeois democracy" in Poland as opposed to the "Soviet democracy" in Russia.

The Treaty of Versailles. At the Paris Peace Conference the Polish delegation demanded the surrender of the formerly Prussian sector, Upper Silesia, and southern East Prussia (Mazuria) from Germany. The largest part of the province of Poznań was already under Polish authority because the December 1918 Polish uprising against the Germans in Poznań had engulfed almost the entire province. The Treaty of Versailles, signed on June 28, 1919, awarded Poland the main part of the Prussian sector, providing the desired direct access to the Baltic Sea, even though the port of Gdańsk (Danzig), to which the Poles had laid claim, became a "Free City." The plebiscite in Mazuria, held on July 11, 1920, resulted in a heavy defeat for Poland, partly because of the Polish–Soviet war, and it brought only a few border regions under Polish authority. The Polish population in Upper Silesia rebelled three times against the German administration. In the plebiscite of March 20, 1921, 682 communes voted for Poland and 792 for Germany, but the Poles secured preponderance in the southeastern plebiscite area, which was industrially the most important. The new frontier of October 20, 1921, divided this area: Poland was awarded 76 percent of the coal production, 22 out of 37 blast furnaces, and nine out of 14 steel works, with about 990,000 people. Populated by Poles, Czechs, and Germans, the former Duchy of Cieszyn was the subject of controversy between Poland and Czechoslovakia. On July 28, 1920, it was divided, but in a manner that left a considerable Polish minority in the Czechoslovakian sector.

Territorial disputes. Poland owed its rapid eastward expansion to its quickly recruited and effective army. Evacuation of the territories occupied in the east by the German troops had to be completed by the beginning of February 1919, but every strip of land left by the Germans was promptly taken over by the Red Army. Thus, between November 1918 and February 1919, that army moved westward from the Dnieper to the Bug River line. Piłsudski decided that it was imperative to throw back eastward the forces aiming at imposing a foreign form of government on restored Poland. The central part of historic Lithuania was occupied by the Poles in April 1919. On July 27 the Supreme Council approved a demarcation line between Poland and Lithuania, leaving Wilno on the Polish side. In eastern Galicia the West Ukrainian People's Republic had been established in November 1918, but Polish troops occupied it in June–July 1919. In addition, the Polish Army extended its sphere of influence over the largest part of White Ruthenia and over the western Ukraine—far beyond the boundary lines of the Second Partition of 1793—without meeting any serious resistance from the Red Army until December 1919, because the Russians were still embroiled in civil war. The Soviets launched a military offensive against Poland between February and May 1920. The militarily and politically ill-advised Polish advance against Kiev in June 1920 led to the Polish–Soviet war that had been brewing for the past one and a half years. The Red Army, under the command of Mikhail Tukhachevsky, advanced, reaching the vicinity of Toruń, Warsaw, and Lwów by July and August 1920. The Polish forces led by Piłsudski defeated the Red Army decisively, however, in a battle on the Vistula between August 16 and 28. This victory crushed the Bolsheviks' last chance of a widespread revolution in central and western Europe. The Polish Army reconquered Central Lithuania with Wilno (Lithuanian Vilnius), western White Ruthenia, western Wolhynia, and East Galicia before the armistice of October 12, 1920. The Soviet government finally recognized this boundary unconditionally in the Peace of Riga on March 18, 1921.

Defeat of the Red Army

POLAND BETWEEN THE WARS, 1921–39

The constitution of March 17, 1921, established a parliament consisting of two houses, the Senate and the Sejm. The political forces in the country had splintered up into numerous parties and groups, and this proved a hindrance to the development of a stable, unified parliamentary government.

Domestic policy. Between 1922 and 1926 Poland was governed by frequently changing cabinets that were based mainly on a coalition between the National Democrats, the Christian Democrats, and the moderate Peasant Party. The primary task of these governments was to attempt to eliminate the inflation that the destruction of World War I and the subsequent conflicts had caused. The government stabilized the situation through the introduction of a new currency, the złoty, on February 1, 1924. But the new currency was in danger again by 1925, mainly as the result of a tariff war instigated by the Germans. The financial crisis that followed led to high unemployment and severe unrest among the working people. This, in turn, gave the Communist Party a steadily rising influence, even though it had been declared illegal in 1923. Piłsudski had not been active on the political scene since 1923, but these new developments compelled him to organize a military coup on May 12–15, 1926, in order to take power. Supporting their former leader with a general strike, the PPS played a decisive role in Piłsudski's victory.

The Piłsudski coup

The Piłsudski government. Piłsudski headed the government from 1926 to 1930. His rule was a moral dictatorship with only a formal preservation of parliamentary authority. His administration aimed at the "moral cleansing" of society in Poland and at eliminating some of the widespread corruption. Piłsudski relied mainly on the army and the conservative nobility of eastern Poland for his support. Between 1926 and 1927 a right-wing opposition emerged, led by Dmowski. Piłsudski confronted them with his Non-Party Block for Cooperation with the Government (BBWR), which was a loose organization held together only by Piłsudski's personality. At any rate, in the elections to the Sejm held on March 4, 1928, the BBWR succeeded in obtaining one quarter of all the parliamentary seats, though a contributing factor for this political success was the increasing economic prosperity of 1926–29.

The Great Depression. The crisis in the world economy began to affect Poland in October 1929. Unemployment hit about 893,000 workers, excluding agricultural workers—or approximately one-third of those employed in 1931. Production sank between 1931 and 1933 to only 46 percent of the 1913 level. Under the pressure of the general misery and need, the PPS, the Polish Peasant Party (PSL), and some others united on November 1, 1929, to form a left-wing opposing coalition (Centrolew, or Centre-Left). Piłsudski employed harsh measures against his opponents, ordering the arrest of about 70 politicians. The Sejm elections of November 1930 gave

the BBWR its desired majority, but the genuine political parties now practically disappeared from the political scene. Piłsudski governed from 1930 to 1935 without holding an official public political office. His authority was based on the support given by the civil service and by his officers' corps and, in particular, the support of the colonels trained in the legions. Piłsudski tried to master the economic crisis through a program of non-intervention and thrift, but he was largely unsuccessful.

The 1935 constitution. A constitutional reform was introduced on April 23, 1935, strengthening the executive branch of government and giving the head of state the power to enact important laws without the approval of Parliament.

Piłsudski's successors. Piłsudski died on May 12, 1935, and the resulting struggle for power, complicated by strikes and demonstrations, soon reduced all parliamentary activity to a complete farce.

Opposition groups. The elections held in September 1935 and November 1938 showed this only too well, and a large part of the population and most of the opposition parties boycotted them. There were violent demonstrations and strikes in 1936 and 1938, and the beginnings of a popular front, bringing left and left-centre parties together, could be noted. The established radical workers' parties, the PPS and the KPP, found a new ally in the Peasant Party, which had now also become radical under its new leader Stanisław Mikołajczyk. Joseph Stalin's reaction to these signs of national unity was to have the Comintern dissolve the Communist Party of Poland in 1938, branding its leaders as provocateurs and capitalist agents. On the right wing, the National Radical Camp (ONR) generated some Fascist groups, but they were dissolved in 1934 and received hardly any support. Founded by the government on March 1, 1937, the Camp of National Unity (OZN) remained an authoritarian association that—like the BBWR—had no real social or political platform.

Dissolution of the Communist Party of Poland

Social problems. The important social questions of Poland remained unsolved in the time between World Wars I and II. Any permanent solution was beyond the reach of the country's economic resources. The surplus in the agrarian population could only have been eliminated by a comprehensive program of industrialization, but there was no capital to do this. Nevertheless, the agrarian reform based on the laws of July 15, 1920, started the distribution of approximately one-tenth of all arable land. Industrialization made some progress with, for example, the energetic promotion of the chemical industry and the building of the "coal-railroads" from Upper Silesia to the Baltic Sea coast, but the funds necessary for investment in these enterprises stretched Poland's national economy to its limits. The entire 1938 production barely reached the 1913 level. Poland remained an agrarian country, and about two-thirds of its total population depended either directly or indirectly on its agricultural produce. Only one-quarter of the country's inhabitants received their income from industry.

Foreign policy. In its foreign policy, Poland attempted to strengthen its position through alliances with France and Romania (1921). Endangered by political and military cooperation between Germany and the U.S.S.R., as well as by their concerted policy aiming at the revision of existing frontiers, Poland found itself in a hazardous position after the treaties of Locarno (1925) and the April 1926 treaty negotiated by Gustav Stresemann and Georgy Vasilyevich Chicherin of the Soviet Union stating that German–Soviet relations should continue to be based on the 1922 Treaty of Rapallo. Acting under Piłsudski's instruction, Józef Beck, Polish foreign minister, concluded a nonaggression pact with the U.S.S.R. in 1932, and a similar one with Germany in 1934—the latter preserving both Poland's alliance with France and its freedom of action as a member of the League of Nations. This safeguarded Poland's position, at least for a short time (1934–38). But the German *Reich* had by now annexed Austria and the largest part of Czechoslovakia, and it established control over the Slovakian and Lithuanian republics. This placed it in such a strategically favourable position toward Poland that Adolf Hitler now dispensed with all caution. His proposals for a global solution were aimed at placing Poland in a situation of complete dependence on Germany so that he would be able to make Poland into a base for his planned attack on the U.S.S.R. Observing the guarantee of March 31, 1939, the British government under Neville Chamberlain gave Poland diplomatic, but no military or financial, assistance. This did nothing to prevent the Germans from attacking Poland or to stop the outbreak of World War II. Hitler, on August 23, 1939, concluded with Stalin a nonaggression pact containing a secret protocol dividing Poland into German and Soviet spheres of interest, circumscribed by the Narew, Vistula, and San rivers. This "Fourth Partition of Poland" proved to be the decisive factor leading to the outbreak of World War II.

Beck's nonaggression pacts

POLAND DURING WORLD WAR II, 1939–45

Defeat by Germany. The fall of Poland came unexpectedly fast during the September campaign of 1939. Warsaw capitulated on September 27, 1939, after the defeat of the Polish Army. The last organized Polish resistance ended on October 5, 1939. Following German demands, the U.S.S.R. ordered two army corps to invade eastern Poland on September 17, 1939, breaking its earlier agreements of 1921 and 1932. On September 28, 1939, Hitler and Stalin again agreed on a new partition of Poland along the Narew, Bug, and San rivers. Hitler's plans for the creation of a Polish puppet state did not materialize because no collaborator could be found. On October 12, 1939, Hitler gave the order to annex outright the former Prussian section, parts of Mazovia, eastern Great Poland with Łódź, and a border strip in Little Poland.

The division of Poland

Soviet annexations. The regions in the Soviet sphere of interest were incorporated into the U.S.S.R. with the added help of supervised plebiscites held between October 22 and November 2, 1939. In contrast with German procedure, the U.S.S.R based its decisions, at least formally, on the "right for self-determination of the nationalities," hoping in this way to get its demands internationally legalized. The annexation carried out by the U.S.S.R. put the British government into an embarrassing situation. On the one hand, it had guaranteed the security of Poland, while, on the other hand, it did not want to risk a conflict with the U.S.S.R. It was for this reason that on October 26, 1939, the British foreign minister Lord Halifax pointed to the fact that the westward extension of the Soviet boundary approximated the earlier Curzon line of December 8, 1919, and of July 11, 1920. He was the first European statesman to do so, but this explanation became the basis for the later establishment of the boundary between Poland and the U.S.S.R. at the end of World War II.

Polish government-in-exile. The forming of a Polish "government-in-exile" in France, based on the 1935 constitution, assured the survival of the Polish Republic. Its leading statesman and supreme commander was Władysław Sikorski. The Polish government-in-exile moved to the United Kingdom after the defeat of France in June 1940. A civil and military resistance movement was also formed in occupied Poland in September 1939.

German repression. From the very beginning, in October 1939, the occupational policy of the Germans was designed to completely eradicate Polish culture. Almost the entire Polish school system was suppressed. Mass executions of Polish intellectuals were carried out, and the infamous concentration camp of Oświęcim (Auschwitz) was set up. The intention was to Germanize the territory, and mass deportations of Poles and mass settlings of Germans in their place were supposed to accomplish this. Germans from the Baltic states and eastern Poland were transferred to the western Polish lands incorporated into the German *Reich*. The Soviet government also participated in mass deportations of the Polish inhabitants from the eastern parts of their country incorporated into the U.S.S.R.

Alliance with the U.S.S.R. The German attack on the U.S.S.R. on June 22, 1941, changed the situation drasti-

cally for Poland since the Germans now began to occupy Soviet-held Poland. In addition, there now existed the possibility of an armed alliance between the Polish government-in-exile and the U.S.S.R. In accordance with the Polish–Soviet agreement of July 30, 1941, a Polish army was set up in the U.S.S.R. under Gen. Władysław Anders. But the Polish units were moved to Persia in 1941 and later fought in the Near East and North Africa. Stalin agreed to the annulment of the "Fourth Polish Partition" agreement between the German and the Soviet governments of 1939, but he declined the restitution of the Polish–Soviet border of 1921. In his debate with the British foreign minister Anthony Eden on December 16, 1941, he proposed the Curzon line as the basis for the future Polish–Soviet boundary.

After the outbreak of hostilities between Germany and Soviet Russia, the German government consolidated East Galicia with the rest of occupied Poland, even though the majority of the population in East Galicia was Ukrainian.

German policy of extermination. The extermination of the Polish intellectuals and the mass deportation of Poles, particularly from the areas around Poznań, Łódź, and Zamość, was vigorously pursued. The Polish citizens of Jewish descent, hundreds of thousands of whom had been interned in the ghettos of the large cities, were deported to extermination camps, such as Auschwitz and Treblinka, where they were killed, along with the Jewish population of the greater part of Europe. Neither the belated ghetto risings of Warsaw, Białystok, and Wilno (April–September 1943) nor the special actions taken by the Polish underground were effective enough to halt this extermination. According to the most careful estimates, about 3,350,000 Polish citizens of Jewish descent were slaughtered. Millions of Polish citizens were sent to forced labour camps in Germany, and hundreds of thousands were executed.

The ghetto risings

Polish resistance. It is not surprising that the Polish underground grew quickly since it considered itself primarily a national organization for self-protection.

From January 1942 on, the Communist and left-wing socialist groups again gained influence. The Polish Workers' Party (PPR) and its military wing, the People's Guard (GL; later the People's Army, or AL) developed from their ranks. One of the secretaries of the PPR was Władysław Gomułka (born in 1905), who was subsequently to become deputy premier of Poland and first secretary of the Polish Communist Party. The PPR became an important and powerful factor when the U.S.S.R. terminated its relationship with the government-in-exile in London after the Katyn affair of April 1943, involving allegations that the Soviet authorities had murdered about 13,400 Polish officers in April 1940. A second government-in-exile was subsequently set up in the U.S.S.R. by the Union of Polish Patriots. The turning point on the Russian front came between June and September 1943, when the Soviet forces began their westward advance. In January 1944 they crossed the frontier of the Riga peace treaty, pushing forward to the Curzon line. By July 31, 1944, the Polish territories all the way to the Narew, Vistula, and San rivers were under Soviet control. On July 21, 1944, the Polish Committee of National Liberation (PKWN) was established in Chełm (later in Lublin) representing the opposing faction to the government-in-exile in London.

The Warsaw uprising

The uprising of Warsaw on August 1, 1944, was staged by the Home Army (AK) to stabilize the administrative power of the London exile government in the capital, which was still occupied by German troops. This insurrection remained without British and Soviet support, however, and the AK had to capitulate. This, in turn, assured the pre-eminence of the Lublin Committee. On January 1, 1945, the Lublin Committee reorganized into a provisional government and exercised authoritarian control in all the territories west of the Curzon line, which had been occupied by the U.S.S.R.

German defeat. The Soviet and Polish armed forces, the latter recruited by the provisional government, began the great winter offensive in January 1945. Taking almost all the lands along the Oder and Neisse rivers by March 10, 1945, they established a decisive precedence over the non-Communist Polish troops fighting in Western Europe as to the making and implementing of political decisions in restored Poland. At the Yalta Conference (February 4–11, 1945), both Franklin D. Roosevelt and Winston Churchill sacrificed the legitimacy of the London government-in-exile to Stalin by consenting to the formation of a mixed interim government in Poland. With the consent of the Soviet government, the provisional Polish government officially assumed sovereign control over all German lands east of the Oder and the western Neisse rivers, in addition to the territory of the Free City of Gdańsk, between February 5 and May 24, 1945.

Emergence of the People's Republic. The provisional government signed a mutual pact with the U.S.S.R. on April 21, 1945. After many difficulties, the new Polish Government of National Unity was finally constituted on June 28, 1945. It included Western-oriented politicians such as Stanisław Mikołajczyk, who became vice premier. But at the same time it assured itself of the predominance of Soviet-inclined parties, especially the PPR, under Władysław Gomułka and Bolesław Bierut. Harry S. Truman and Winston Churchill appreciated this bias in the Polish government and decided on a policy of strong resistance to Stalin.

Potsdam and the Oder–Neisse line

With this policy in mind, they stipulated at the Potsdam Conference (July 17–August 2, 1945) that the Polish western boundary along the Oder and Neisse rivers should be recognized as a temporary boundary. Despite this, the Allied governments removed all ethnic Germans from the territory east of the Oder and Neisse rivers, thus securing Polish sovereignty. The Polish–Soviet treaty of August 16, 1945, confirmed the Polish eastern boundary along the Curzon line.

Effects of the war. About 6,000,000 Polish citizens, or 22 percent of the total population, died during World War II. Of these, only 600,000 were actual war casualties. The greatest loss, approximately 4,500,000, was due to mass extermination, deportation, exile, and emigration. The damage to industries, communications, housing, buildings, and to culture in general from the war has been impossible to estimate. The Germans were responsible for at least four-fifths of these losses; the rest was the responsibility of the U.S.S.R. World War II was thus the greatest catastrophe in Polish history in terms of total loss of life and property and as a result of the thorough and premeditated attempt at genocide.

POLAND AFTER WORLD WAR II

Formation of the People's Republic. Under the Polish Government of National Unity that was in power between June 28, 1945, and February 6, 1947, Poland was again a formal democratic republic, recognizing the constitution of March 17, 1921. But the nationalization of the large- and medium-size industries in 1946 and a few changes in the constitution, endorsed by a plebiscite on June 30, 1946, were a prelude to the reorganization of the Polish government into a People's Democracy. Strong resistance was raised against this, and only one-quarter of all the soldiers who had fought in the armies under British command actually returned to Poland. In Poland itself an effective opposition was organized against the often despotic measures taken by the government. Influential groups of the grand old parties, such as the National Party, PPS, SL, and SP, supported the resistance movement, which was founded on September 2, 1945. The resistance movement had as its aim a "democracy in a western European interpretation" and "Christian ideals." The Polish security corps, the police, and the Polish Army quelled the partisan war waged by the resistance movements with extreme harshness, particularly during the period between March 1945 and April 1947. This resistance later made it all the more difficult to establish legal bourgeois parties. The National Party and other right-wing associations were declared illegal from the very beginning, but the political groups that collaborated with the government formed a "bloc of democratic parties," led mainly by the PPR, under Bierut and Gomułka, and the left wing of the PPS, under Józef Cyrankiewicz and Adam Rapacki. In the elections for the legislative

Resistance to sovietization

National Assembly on January 19, 1947, this democratic bloc received 80.1 percent of all votes cast, according to official figures.

Socialist Poland. The gradual conversion of Poland to a purely socialist state took place under the presidency of Bierut (1947–52) and under the premiership of Józef Cyrankiewicz. Gomułka became deputy premier. The entire national economy was nationalized and brought under public ownership. The exception was the agricultural sector, which was only partially collectivized. On December 15, 1948, the PPR and the PPS merged to form the Polish United Worker's Party (PZPR). With the growing political hostility between the U.S.S.R. and the Western powers between 1947 and 1949, Stalin ordered the purging of those politicians who were considered "rightist-nationalist" deviationists. Being part of this group, Gomułka and his supporters were taken into custody (September 1948 to November 1949). This left Bierut as both president and first secretary of the PZPR (1948–56). The planned socialist economy was finally introduced with the establishment of the State Planning Commission on February 4, 1949, and the adoption of the six-year plan on July 21, 1950.

The purges ordered by Stalin

The new constitution. On July 22, 1952, a new Soviet-type constitution was passed and was slightly revised in 1972. This made Poland a People's Republic. Under this constitution the supreme organ of state authority was the Council of State; and the supreme organ of state administration was the Council of Ministers appointed by and responsible to the Sejm. In the elections of October 26, 1952, the PZPR received about 64 percent of all the seats in Parliament. The constitution remained purely formal, however, considering the factual dictatorship of the PZPR.

The post-Stalin period. After Stalin's death, Cyrankiewicz returned to the prime minister's office (1954–57). Political repression was relaxed, and about 100,000 political prisoners were freed in 1955 and 1956. But the real de-Stalinization was not implemented before the 20th congress of the Soviet Communist Party in 1956, after the sudden death of Bierut on March 12, 1956. The leadership of the PZPR considered the severe workers' unrests in Poznań in June 1956 a warning signal, and they therefore found it expedient to endorse a deep consideration of possible reforms. Reinstated on August 5, 1956, Gomułka again took over the leadership of the PZPR since he was the only man who could restore party unity and count on wide popular support.

Re-emergence of Gomułka

The leadership of Gomułka. The party leadership of Poland was still autocratic under Gomułka (1956–70), but the country experienced a much more liberal government than the other states in the eastern bloc. The administration allowed the return of private ownership in small businesses, great efforts were made for the construction of new homes, the collectivization of agriculture was revoked, and the freedom of the press was considerably extended. From 1956 to 1971 the PZPR maintained a steady 51 percent of all seats in Parliament due to predetermined selection of candidates, but a certain political codetermination of other parties or groups was now tolerated. There were power struggles within the PZPR from which the "partisans" under the leadership of Gen. Mieczysław Moczar, minister of the interior, emerged as the strongest faction in 1964. Moczar and his supporters suffered a setback in connection with the student unrests and with their policy of anti-Zionism in 1968, but they were not totally removed from their position until 1971.

Cultural developments. The scientific achievements and the general level of education reached a very high level under the Gomułka regime. The Catholic Church was able to expand its work relatively freely after an agreement reached in 1956. The Polish episcopacy was instrumental in preventing open rebellions and thus indirectly preventing Soviet military interventions, especially in October 1956 and in December 1970.

Agriculture and industry. The loss of human life during World War II, the land distributions in the western districts of Poland, and the agrarian reforms of 1944 and 1945 much improved the long-standing agrarian crisis in Poland. It was finally completely eliminated under Gomułka, who followed the principle of private ownership for the farmer. Consequently, an average of about 85 percent of the arable agricultural land was in private hands during the years 1956 to 1972. Growing industrialization, which had begun even before 1956, started a mass exodus of people from the rural areas into the metropolitan centres. The system of a centrally controlled economy proved too stifling and inhibiting as the industries began to expand. For this reason, the fifth convention of the PZPR in 1968 alleviated the restrictions. The single-handedness of Gomułka's decisions, his disregard for popular opinion, and finally the bad harvests in 1969 and 1970 led to severe unrest among the working people in the Baltic seaport cities in December of 1970. He was forced to resign on December 20, 1970. His successor, Edward Gierek, secured popular support through his policy of extending the consumer industries and through promoting public discussions and rehabilitating many members of the creative intellectual class, although he was unable to eliminate the danger of creeping unemployment.

In March 1972 Henryk Jablonski was named to the presidency, and Pyotr Jaroszewicz to the premiership.

Foreign policy. During the Cold War and due to the U.S. policy of disengagement in eastern Europe between 1949 and 1956, Poland's foreign policy remained closely tied to that of the U.S.S.R. In 1949 Poland joined the Council For Mutual Economic Assistance (Comecon); and it became a member of the Warsaw Pact in 1955. The central problem of Polish foreign policy has been the legal, international recognition of its western boundary (the Oder–Neisse line). The German Democratic Republic recognized this boundary in 1950, but the Federal Republic of Germany disputed it at least until 1963. The governments of the U.S., U.K., and France recognized this western border line tacitly after the gradual cessation of the Cold War from 1956 onward. The Polish–Soviet agreement of 1965 renewed the Soviet border guarantee that had existed since 1945. The Polish foreign minister Adam Rapacki tried to slowly break through the iron curtain with his plan for disarmament, put forth for the first time on October 2, 1957. After the nuclear test ban treaty, concluded in 1968, and the start (November 1969) of the strategic arms limitation talks between the U.S. and the U.S.S.R., the Rapacki plan again became the basis for serious negotiations toward a planned European security system. In connection with the policy of general relaxation, Gomułka, the West German chancellor Willy Brandt, and Cyrankiewicz signed in Warsaw on December 7, 1970, a treaty declaring that the existing German–Polish border line, determined at Potsdam on August 2, 1945, by the governments of the United States, Great Britain, and the Soviet Union, "forms the western state frontier of Poland"; the Federal German government "affirms the inviolability [of this frontier] now and in the future."

The Rapacki plan

A visit to Warsaw by U.S. president Richard Nixon (May 31, 1972) resulted in the establishment of closer trade contacts with the U.S.

BIBLIOGRAPHY. The standard work for the political history of Poland until 1936 is *The Cambridge History of Poland*, 2 vol. (1941–50), with chapters written by the acknowledged Polish and British scholars of the period between World Wars I and II. Contributions by leading historians of the People's Republic of Poland are included in ALEKSANDER GIEYSZTOR *et al.*, *History of Poland* (1968), which covers events until 1939. A group of scholars in exile have published the representative collection: HENRYK PASZKIEWICZ *et al.*, *Polska i jej dorobek dziejowy* (1956), recounting Polish history until 1918. Of further importance are the chapters on Poland in *The New Cambridge Modern History*, vol. 1 (1961). The results of recent historical research in Poland are available in essays and synoptic form in English, French, or German in the *Acta Poloniae Historica* (annual). Of considerable value also are the Polish contributions to the international congresses for historical research. For an evaluation of Polish history from the Catholic point of view, see OSKAR HALECKI, *La Pologne de 963 à 1914* (1933; Eng. trans., *A History of Poland*, rev. ed., 1960). The recognized standard work on Polish culture and its history of ideas is ALEXANDER BRUCK-

NER, *Dzieje kultury polskiej*, vol. 1–3, 3rd ed. (1957–58), and vol. 4, 2nd ed. (1946). Still absolutely necessary for the social sciences and the development of economics in Poland is JAN RUTKOWSKI, *Historia gospodarcza Polski*, 2 vol. (1947–53). The collective work *Miasta Polskie w tysiącleciu*, 2 vol. (1965–67), contains a wealth of material on the subject. For a concise history of the development of Polish government, the *Historia ustroju Polski w zarysie* by STANISLAW KUTRZEBA, completed by ADAM VETULANI, 8th ed. (1949), is still basic. It should be complemented by the more detailed work of the two scholars in constitutional history, JULIUSZ BARDACH and B. LESNODORSKI, *Historia państwa i prawa Polski*, 2nd ed., 2 vol. (1964–66), which also features an excellent bibliography. Polish church history is synthesized in *Kościół w Polsce*, 2 vol. (1968–69), written by a group of researchers of the Catholic University of Lublin under the editorship of JERZY KLOCZOWSKI. JAKUB SAWICKI has published numerous studies on the religious minorities in Poland, such as *Studia nad położeniem prawnym mniejszości religijnych w państwie polskim* (1937), on the legal position of religious minorities in Poland. The Polish government has recently republished the history of former eastern Poland and its minorities in single volumes. Beyond these are OSKAR HALECKI, *Borderlands of Western Civilization* (1952); and ADAM ZOŁTOWSKI, *Border of Europe* (1950). MARIAN KUKIEL, *Zarys historii wojskowości w Polsce*, 5th ed. (1949), is still fundamental for the country's military history. The era of the founding of the state and the time of the early Piasts has been most thoroughly researched after 1945, mainly as the result of new excavations. Important in this respect is WITOLD HENSEL, *Polska przed tysiącem lat*, 3rd ed. (1967; Eng. trans., *The Beginnings of the Polish State*, 1960). The Middle Ages and early modern times are all discussed in the above-mentioned works. The following monographs in western languages are currently available: HERBERT H. KAPLAN, *The First Partition of Poland* (1962); JEAN FABRE, *Stanislas-Auguste Poniatowski et l'Europe des lumières* (1952); WALERJAN KALINKA, *Der Vierjährige Polnische Reichstag 1788–1791*, 2 vol. (1896–98); R.H. LORD, *The Second Partition of Poland* (1915); M. HANDELSMAN, *Napoléon et la Pologne, 1806–1807* (1909), now somewhat dated; MARIAN KUKIEL, *Czartoryski and European Unity, 1770–1861* (1955); R.F. LESLIE, *Polish Politics and the Revolution of November 1830* (1956) and *Reform and Insurrection in Russian Poland, 1856–1865* (1963). Outstanding on the subject of the founding of the Second Republic is TITUS KOMARNICKI, *Rebirth of the Polish Republic* (1957), containing an extensive bibliography and documentation. A recent modern history of Poland is HANS ROOS, *Geschichte der polnischen Nation* (1961; Eng. trans., *A History of Modern Poland*, 1966).

(Ha.Ro.)

Polar Biomes

The polar regions comprise the Arctic, in the north, and the Antarctic, in the south, and although these two are apt to be confused, they are indeed "poles apart" as regards their animal and plant associations, or biomes. The few similarities that exist result chiefly from the coldness, widespread aridity, and level of latitudes. (Because of the latitude, day and night practically divide the year: the sun is continuously above the horizon in summer and below it in winter.) Even the criterion of coldness is elicited mainly in contrast with more hospitable regions, for in the Arctic there is probably no lowland area that does not enjoy for at least one month in each year a mean temperature appreciably above the freezing point, although in the Antarctic only the off-lying islands and the coast of Graham Land on the Antarctic Peninsula are favoured with this warming.

Boundaries of the polar regions

Because the arbitrary latitudinal Arctic and Antarctic circles have no meaning as natural boundaries for plants and animals, biologists regard the polar regions in general as lying poleward of the limits of tree growth. One generally recognized natural boundary of the Antarctic is the Antarctic Convergence—a delimitation in the sea, where cold surface-water layers spreading northward from the continent meet with and sink below warmer mixed sub-Antarctic waters. This boundary embraces to its south some of the sub-Antarctic islands, such as Heard Island and South Georgia.

In the present context, biomes are taken to be more or less climax formations of plants and animals considered together and comprising the biota—apart from recent introductions and other disturbances caused by man.

THE ENVIRONMENTAL SETTING

Fossil evidence indicates that in earlier geological ages many parts of the Arctic and Antarctic had temperate or warmer climates favouring quite luxuriant tree growth. Over millions of years the change to relative barrenness has occurred probably both through variations in world climate (the ice caps are no more than 4,500,000 years old) and through the drifting of landmasses poleward from lower latitudes. In any case the polar regions today —especially the Antarctic—are characterized by extremely harsh climates. Precipitation is generally so low —often only a few centimetres (a centimetre is about 0.4 inch) a year—in the highest latitudes that, coupled with the low temperatures and permanently frozen subsoil, there prevails a desert-like barrenness. Nevertheless, the availability of snow meltwater, often throughout the short growing season, and the long-lasting summer daylight favour some limited plant growth.

Desert-like barrenness

The Arctic. The highest latitudes in the north are occupied by the Arctic Ocean, the farthest north land being in Greenland and the Canadian Arctic Archipelago. To the south stretch islands or major landmasses, which in many sectors of the northern polar cap extend into the sub-Arctic and then temperate zones, thus affording more or less continuous dispersal routes for plants and animals from the south. The Arctic Ocean is deep, but the surface is largely covered by floating ice. The land area is commonly rugged, with widespread mountain ranges and glaciers descending to the sea—especially in Greenland, occupied by the world's second-largest ice cap (see ARCTIC OCEAN).

The Antarctic. The highest latitudes in the south are occupied by the continent of Antarctica, which is largely covered under the world's greatest ice cap. Only on projecting mountains (nunataks) and limited ice-free tracts near the shores, comprising about 4 percent of the total area, can land plants grow. The continent is far more isolated than is any Arctic land, being fully 800 kilometres (500 miles) from the nearest continental landmass, though scattered and usually small sub-Antarctic islands exist and very probably act as "stepping stones" for some limited plant and animal dispersal. Nevertheless, for practical purposes the Antarctic is isolated by formidable sea barriers that have made it extremely difficult for colonization by plants and animals since the last glacial period (see ANTARCTICA).

THE BIOTIC COMPONENT

Arctic and Antarctic plants and animals are extremely well adapted to their harsh and inhospitable environments. The form of adaptation is often similar in both polar regions, but rather few species are common to the two. While their terrestrial and freshwater ecosystems (ecological communities considered as a unit together with the nonliving factors of their environment) are widely different, the far more prolific marine ecosystems are comparable; indeed in both the Arctic and the Antarctic, it is chiefly in the seas that life abounds.

Richness of life in polar seas

Vegetation. Polar food cycles, from plant producers through various animal consumers, are relatively simple, with few organisms linked together. Unfavourable climatic and soil conditions combine to limit severely plant growth in the Antarctic. But in the Arctic, vegetation is often fairly luxuriant in the southern tundra ranges, and its component plants are very diverse—as also are the dependent animal species. Of Arctic flowering plants and ferns, about 1,000 different species are known from regions north of the tree line. Some of these are confined to the Arctic, but the vast majority range far southward—especially on mountains, in characteristic Arctic-Alpine distribution. The more primitive "lower" plants are less known but probably include almost as many species of mosses as of vascular plants and some 300 species of liverworts. Lichens are probably at least as numerous as the mosses, and the fungi are surely far more so, fully 850 species of fungi having been recorded from Greenland alone. Arctic algae are very numerous, —probably more than 2,000 different species, a considerable number of which are limited to the Arctic Ocean. The

bacteria of the Arctic have been insufficiently studied, but their numbers must be considerable, for a wide range of types occurs in both salt- and freshwaters, as well as in bottom deposits and soils.

In marked contrast to the Arctic, the Antarctic proper possesses only two native flowering plants: a grass, *Deschampsia antarctica,* and a relative of the pinks, *Colobanthus quitensis.* Though both of these are limited to favourable exposures in the Peninsula sector, they also occur on the sub-Antarctic islands and in southernmost South America. Apart from some algae and other relatively simple organisms, terrestrial life is restricted to the 4 percent of land area that is ice-free. There occur more than 70 species of mosses but relatively few liverworts. Lichens are the most successful and widespread group of plants in Antarctica because of their ability to colonize bare rock surfaces and their physiological adaptation to low temperatures and arid conditions; about 400 species have been recorded from the continent, although this number may be reduced by critical taxonomic study in the future. They extend to slightly over 86° S, and include *Verrucaria serpuloides,* the only known marine lichen, which passes its entire existence underwater. Of algae, approximately 360 nonmarine species have been recorded from Antarctic ice, snow, soils, and temporary summer ponds. The snow algae, which give rise to reddish, yellow, or green patches of snow, are found only in the coastal areas where permanent snowfields undergo extensive annual thawing.

Compared with the Arctic, fungi and bacteria are few in the Antarctic. Bacteria that were isolated from lakes, and yeasts of the genus *Candida* isolated from rock detritus and snow, proved remarkably well adapted to low temperatures. In the seas more than 200 species of algae have been recorded, the majority being diatoms; bottom-dwelling algae in shallow waters sometimes form extensive miniature "forests".

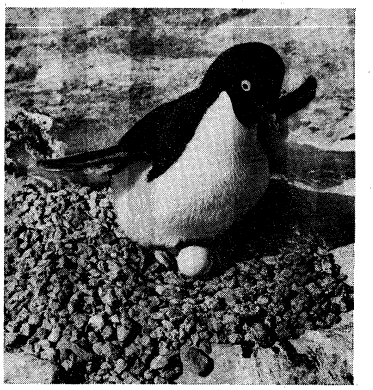
Animal life. Arctic animals tend to be far less numerous than plants, both in species and individuals. In North America the most notable among Arctic mammals on land are the Arctic hare, wolf, fox, collared lemming, polar bear (chiefly along coasts), short-tailed weasel, caribou, and musk-ox. Except for the musk-ox, these animals are all represented in Arctic Eurasia by the same or related species. All these mammals reach the far north, most of them practically to the northernmost limits of land, but some others, including the wolverine, shrews, ground squirrels, and some additional small rodents and weasels, range well into the Arctic without reaching so very far north. This is also the case with the brown bear, red fox, and ermine.

Range of animal life in the Arctic

The Arctic seas are the habitat of various seals, whales, walrus, and narwhals, while white whales (belugas) frequent narrow inlets and the mouths of rivers. (Man in ever-increasing numbers threatens to change the fragile ecosystems of the Arctic as never before; see CONSERVATION OF NATURAL RESOURCES.)

Although a few marine birds range over the Arctic Ocean north of the northernmost coasts, most Arctic birds are limited to the vicinity of land and southward. Almost all of them migrate south to spend the winter, but Arctic ptarmigan regularly and snowy owls sometimes remain in the far north. On land, numerous ducks, geese, and waders, particularly, come to nest and pass the summer, as do some swans, while pipits and ravens are widespread, as are various falcons and other birds of prey. Often nesting in vast numbers collectively on bird-cliffs or on the ground are various kinds of guillemots (murres) and terns, as well as puffins, little auk, and fulmar. Loons, or divers, call eerily from the numerous shallow lakes, and various kinds of gulls and skuas, or jaegers, screech from above. The Arctic, in fact, seems to abound in bird life in summer. At this time of year also the air inland (except in the far north) teems with insects, especially mosquitoes. Still farther south the insects probably contribute the greatest number of species to the total fauna; they diminish drastically in the far reaches of the north, except for the wingless springtails (Collembola), which occur even on snow and ice.







Animal life in Antarctica.
(Top left) Emperor penguin, the only Antarctic bird that breeds in winter. (Top right) Adelie penguin, and (bottom) Weddell seal, two of the few types of animals native to Antarctica.
By courtesy of (top right and bottom) the U.S. Navy; photograph, (top left) Popperfoto

Freshwater fish also extend well into the Arctic, the northernmost apparently being the Arctic char (*Salvelinus alpinus*) found in lakes around 82° N. Other prominent Arctic fish include cods (Gadidae), eelpouts (Zoarcidae), sculpins (Cottidae), and lumpsuckers and related forms (Cyclopteridae); a few sharks also occur. The Arctic marine-invertebrate life is often particularly rich and includes many sizable Crustacea; this fauna is dependent for food on the seasonally abundant phytoplankton and in turn supports most of the vertebrate animals.

Like the land vegetation, the animal life of Antarctica is extremely poor. Truly terrestrial vertebrates are lacking, though various sea-feeding penguins and fulmar and other petrels come ashore to breed—as do also the great skua, Dominican gull, Antarctic tern, and Antarctic shag. There are four kinds of truly Antarctic seals, all belonging to different genera—the Weddell seal, the leopard seal, the Ross seal, and the crabeater seal, of which the last-named is the most abundant seal in the world, numbering an estimated 5,000,000 to 8,000,000 individuals. Several of the largest species of whale are still found in the open seas around Antarctica, though their populations have been drastically depleted by excessive hunting. By extreme contrast the largest terrestrial animals of these rigorous regions are two species of small midges, one of which is winged and the other wingless; these, in addition to a limited array of mites, springtails, micro-

Scarcity of animal life in the Antarctic

scopic rotifers, tardigrades, protozoans, nematodes, and rare enchytraeid worms, constitute the native land fauna. In the adjacent seas there is an abundance of the crustacean krill (*Euphausia superba*), constituting the greater part of the food of the blue, fin, and humpback whales, and a vast array of other zooplankton, often in great abundance, even at depths where the temperature is constantly below the freezing point of freshwater.

COMMUNITY STRUCTURE

Unlike the situation in most parts of the world where plants form the main embellishment on land, the Arctic and, to a far greater extent, the Antarctic tend to have the framework of their terrestrial ecosystems dominated by the physical structures that locally make up the earth's surface. In the Antarctic, nesting penguins, because of their accumulated mineral-rich guano (excrement), may create barrens on which no plant can grow; but, on the other hand, large seaweeds in the sublittoral (below tide) zones of the sea create suitable habitats for a wide range of marine animals. Thus on land it is chiefly in the low-Arctic and the sub-Antarctic islands that plants really affect the landscape.

In the polar regions the struggle of the totality of organisms tends to be more with the extremely exacting environments than with one another, though some competition for space and food does exist. This occurs among both animals and plants, but, in general, what one immediately sees of living materials in these regions consists almost entirely of plants, the greatest development of which is the Arctic tundra of low-growing species interspersed with lichen-covered rock fields and numberless ponds (see TUNDRA).

Besides marshy and other seaside vegetational types of limited extent, there may be extensive areas of sedges and cotton-grasses (*Eriophorum*), and of true grasses, such as *Arctophila* and *Arctagrostis*, around lakes and ponds in the Arctic. Luxuriant "patchwork quilt" communities of mixed mosses and lichens are often found around and above bird-cliffs, nourished by mineral-rich bird excrement. Where the snow drifts so deeply in winter that the growing season is drastically reduced by its late melting, zoned series of vegetation types develop. The seas abound in tiny free-drifting organisms, or plankton, of which the plant component, or phytoplankton, occur as deep as light penetration allows; these last support a wealth of crustacean and other zooplankton and, ultimately, fish and mammals (*e.g.*, seals and whales). Even floating ice bears a considerable population of diatoms on both its upper and lower surfaces, making the ice appear a dirty brown in summer.

Effect of snow on vegetation

The Antarctic, by contrast, is impoverished. There is no true tundra on the continent, but only occasional thin tracts of sparse mossy vegetation and algal and lichen patches in especially favoured places. These tracts are found along coasts where the climate is mildest; they contribute to the peat formations on the west side of the Antarctic Peninsula. Other terrestrial habitats include rocky coastlines, the slopes of mountains projecting above the continental ice cap, glacial moraines, summer meltwater channels, and warmed areas around volcanically active fumaroles (holes emitting hot gases and vapour). There are also some lakes and numerous freshwater ponds and puddles. In all such areas on land and in freshwaters, however, the biomass, or totality of living things, is such that vegetation appears to the uninitiated to be virtually lacking except in especially favoured areas that are merely tinted by it. By contrast, in shallow marine waters the biomass of living matter may be considerable, being often greater than in the Arctic. This richness of sea life results apparently in part from the upwelling of nutrient-rich waters in the Antarctic Ocean.

The sub-Antarctic islands south of the Antarctic Convergence show polar affinities and are commonly included as part of the Antarctic in the broad sense, though on the islands of Kerguelen, Heard, and South Georgia, there are commonly coastal tussocky grasslands of *Poa* or *Festuca* species and inland fields of herbaceous plants, with mainly lichens and mosses on higher land. Especially notable is the Kerguelen cabbage (*Pringlea antiscorbutica*), the sole representative of its genus. Some ferns and a club moss occur in protected habitats. Animal life on these islands is more prolific than on the Antarctic continent; insects and birds are numerous.

Antarctic islands' vegetation

COMMUNITY DEVELOPMENT AND PRODUCTIVITY

Although in and around the nesting rookeries of seabirds (except for the penguins as described earlier) the plants may to a considerable extent follow the animals, in general the opposite obtains. Vegetation develops how and where it can, and the animals follow—often becoming regular and important associates, doubtless helping to spread the plants by disseminating seeds and other plant parts. Thus, characteristic biomes develop, often in spite of cyclic or other changes in which the polar regions abound. As for successional (seral) changes in the Arctic, marshy and boggy formations, arid rock and other land formations, and flower slopes and herb fields, all theoretically develop into relatively stable communities, within which the organisms habitually group themselves into smaller units called microbiomes.

Except in the more favoured marine habitats, as already indicated, biological productivity in the polar regions is rather meagre at best, tending to be higher in the Arctic than in the Antarctic. The polar organisms, however, are usually well adapted to withstand rigorous conditions, an impressive instance being some Antarctic lichens, which, after artificial frost treatment by slow stages down to −198° C (−324.4° F) and then gradual warming, recovered normal activity within one day. Many lowly polar plants are capable of photosynthesizing at temperatures even several degrees below freezing point and can carry on vital activities even when the surrounding air temperature is far below 0° C (32° F); differentials as high as 25–30° C (45–55° F) between plant and air may be observed in both the Arctic and the Antarctic. In such circumstances, annual net production may reach the order of some hundreds of grams of dry weight per square metre in moss communities at favourable sites even in Antarctica. In the more favoured Arctic, although plant productivity is low daily production during the short growing season is comparable with that of many herbaceous communities of temperate regions.

Annual biological production

In the matters of productivity and food webs of polar and allied biomes, much work was under way in the early 1970s through the International Biological Programme, acting in several countries, which should yield further interesting results in due course. Preliminary reports indicate that in Char Lake on Cornwallis Island, the phytoplankton, consisting mainly of flagellates and diatoms, has an annual production of 5–7 grams per square metre (0.15–0.21 ounce per square yard) where the average depth is 10.4 metres. The average daily wet weight of phytoplankton and the mean chlorophyll content were surprisingly constant from April to October, though the surface of the lake was only ice-free from mid-August to mid-September. Only two species of zooplankton were present, but there were numerous Arctic char, which fed mainly on immature stages of midges.

BIBLIOGRAPHY. *Antarctic Research Series*, from vol. 1 (1964, issued irregularly), mostly monographic treatments of various topics; MARTIN W. HOLDGATE (ed.), *Antarctic Ecology*, 2 vol. (1970), a particularly valuable symposium of numerous specialist papers; NICHOLAS POLUNIN, *Botany of the Canadian Eastern Arctic, Part III: Vegetation and Ecology* (1948), and *Circumpolar Arctic Flora* (1959); see also his articles "Arctic Flora," "Arctic Vegetation," and "Tundra" in *Encyclopedia of the Biological Sciences*, 2nd ed. (1970).

(N.Po.)

Polemoniales

This order of flowering plants, consisting of eight families and over 4,000 species, includes some of the most handsome and favoured of garden and wild flowers. European and American gardens sparkle in spring and summer with the colours of *Phlox, Polemonium* (Jacob's ladder or Greek valerian), *Ipomoea* (morning glory), *Myosotis* (forget-me-not), and *Phacelia*. Woodlands, glades,

mountains, and deserts, particularly in North America, are often brilliantly coloured with their petals. Many other species, however, are often considered weeds, and a few justly deserve that appellation. Most members of the order are annual, biennial, or perennial herbs, many of which are delicate and ephemeral desert inhabitants. Several others, however, are woody trees, shrubs, or vines of great duration, but because of their limited size, number, and distribution are not important timber producers. Two families of the order are unusual in consisting of nonphotosynthetic parasites.

GENERAL FEATURES

Sweet potato the principal food plant

Economic importance. On the whole, the order is of minor economic importance. The one species of major importance as a food plant is the sweet potato, *Ipomoea batatas* (family Convolvulaceae). The nutritious tuberous roots of the plant have been an important food source since prehistoric time. Utilization is now worldwide, but long before European man began to eat the plant its use was well established in South America and in the Pacific Islands. The fact that it was well-known as a crop plant in two isolated geographical areas before botanists began to study it presented an interesting dilemma for students of the origin of cultivated plants. Was it originally American or Polynesian? Even though the sweet potato is not known as a wild plant, its apparent close wild relatives are found in South America. This, and other compelling reasons, led most botanists to the conclusion that the cultivated plant originated in South America and was transported at an early time to such diverse regions as New Zealand, Hawaii, and Easter Island. The mode of transport and the migration route of the sweet potato to these remote regions are matters not yet resolved.

Other members of the family Convolvulaceae (morning glories) have aroused recent interest because their seeds contain significant quantities of the hallucinogenic chemical substances lysergic acid amide and isolysergic acid amide, which are similar to the drug LSD. This psychotomimetic attribute was known to the Aztecs, and certain Mexican Indians still use the seeds for divinatory purposes. Species of the genera *Ipomoea* and *Rivea* possess significant quantities of the drugs, the different species varying in their potency. The natural quantitative variation in the drug content of the seeds leads to public-health problems among illicit drug users because it is difficult to determine the drug content of a seed before it is eaten. Serious psychic effects and even death have resulted from uncontrolled use of the seeds. Since many *Ipomoea* species are common weeds, it is unlikely that legislation would be effective in controlling the use of morning glory seeds for psychotomimetic purposes.

Garden ornamentals

Use of members of the order for aesthetic purposes as garden ornamentals is another economic application of these plants. Compared with the cash value of sweet potatoes, the commercial value of these plants as ornamentals is not very great, but the annual spring perusal of seed catalogs is a much richer experience because of the inclusion of members of the families Convolvulaceae, Polemoniaceae, Hydrophyllaceae, and Boraginaceae. Species of the genus *Dichondra* (Convolvulaceae) have unique value as turf plants in tropical and subtropical regions. They are tough, low-growing, small-leaved, small-flowered plants that grow appressed to the ground, qualities that enable them to compete successfully with grasses. They are superior to grasses in one respect—they need little or no mowing. A vigorous stand of *Dichondra* provides one of the most beautiful lawns possible.

Size range and diversity of structure. As treated here, the eight families of the order Polemoniales can be grouped into two categories. The three small families, Cuscutaceae, Lennoaceae, and Hoplestigmataceae, are sufficiently distinct individually that each will be discussed separately. The five larger families, Polemoniaceae, Convolvulaceae, Hydrophyllaceae, Boraginaceae, and Loasaceae, however, show a similar range of variation so that one family may serve as a model to portray the whole group of five. The family Polemoniaceae is ideally suited for this, since it is large enough to illustrate most of the variation encountered in the other families, yet small enough to have undergone critical study.

Growth forms. The form of members of the family Polemoniaceae and the entire order includes a great range of morphological (structural) diversity. Large shrubs and small trees grow in the Andes in South America. In North America, the nearest approach to this growth habit is seen in some Mexican plants, but most others are at best low shrubs or woody-based perennial herbs. The latter constitute a significant part of the family in western North America. Large herbaceous perennials with extensive underground stems are another important growth form in the family. There are other plants that may be weakly perennial, but that tend for the most part to be biennials. Such plants often attain large size. By far the largest number of species in the family are annuals. Some tend to grow for many months when conditions are favourable, but many others are desert annuals, whose growth is rapid and usually complete in just a few weeks. Some annual members of the family are extreme dwarfs that flower while seemingly mere seedlings. Some of the latter rarely produce true foliage leaves but exist on the produce of their cotyledons ("seed leaves"), photosynthetic stems, and floral bracts.

Stems. Stem growth varies from extensive in the trees and shrubs to very little in desert annuals. The patterns of stem growth include underground rhizomes (rootlike stems), thickened woody crowns, typically branched aerial stems, and vines.

Leaves. Leaf form is similarly varied, ranging from pinnately (lateral leaflets arranged along a central axis, comblike) or palmately (all leaflets radiate from a common point) compound to simple needles and from large to nearly lacking. In a few cases, the tips of compound leaves grow out into tendrils, which provide support in the climbing vines. The general outline of the leaf includes those that are broadly elliptical or lance-shaped, those that are elongate but narrow, and small needles or scales. Leaf margins may be lobed, toothed, or smooth. Veins may be extensive, branched, and prominent or so poorly developed as to be nearly lacking in some leaves. Leaves are often clothed with a hairy covering that may often consist of cobwebby, prickly, or glandular plant hairs. In most members of the family the leaf arrangement is alternate along the stem but in two genera it tends to be opposite.

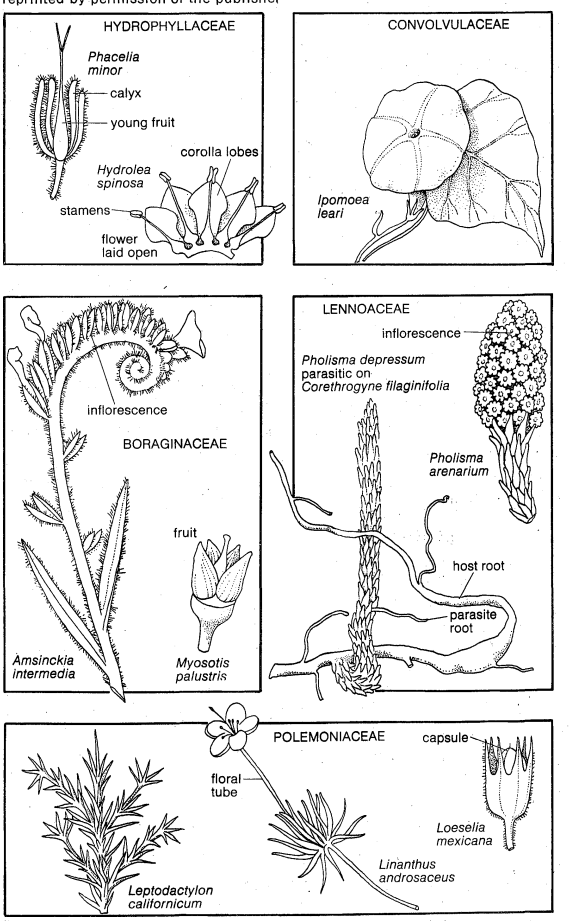
Plant hairs

Flowers and fruits. Flowers of the order Polemoniales are usually disposed in clusters in which the first flower to bloom is the apical one; thus, they are cymose. The cluster may be loose or, in a few cases, a compact head. Some species have clusters consisting of hundreds of flowers whose period of bloom may last many days. Others have only a few flowers per inflorescence (flower cluster), and a few species have a solitary flower at each branch tip. In these, the flowers are either very large, or, if smaller, the plant is so compact that it may literally be covered with flowers—each flower solitary at the end of a crowded branch. The flowers of the order Polemoniales are highly varied in structure, form, size, and colour; however, they share a basic plan, consisting of hypogyny (sepals, petals, and stamens attaching at the base of the ovary), five united sepals, and five united petals, to which the filaments of the five stamens are fused; they generally have three carpels (ovary wall structures) united into a three-locular (chambered) compound pistil (female structure). Nectar, produced by special glands at the base of the pistil, often collects in the corolla base. The sepals vary in shape from a large saucer-shaped structure to a constricted tube. The tissue along the midrib of each sepal tends to be green and dense, but the tissue between the lobes is thin and membranous. The membrane may be flat, corrugated, or prominently ridged. In some genera, the calyx (all sepals considered together) enlarges bladderlike in fruit, whereas in other cases the membranes are ruptured as the enclosed fruit enlarges. The most striking variation in flowers is seen in the petals or corolla. Usually the corolla is radially symmetrical, but in some genera it tends to be bilabiate (two-lipped). In other aspects of form it may be described as rotate (round, wheellike),

Flower shapes

campanulate (bell-shaped), tubular, funnelform, or salverform (trumpet-shaped). The individual corolla lobes may be rounded, truncate, pointed, or erose (with eroded margins, as if gnawed). In size the corolla may be minute or up to several centimetres long or broad. Colours vary from white to yellow through orange, pink, red, mauve, or blue. The flowers may be uniform or multicoloured, often with conspicuous spots or streaks arranged in definite patterns. Stamens (male reproductive organs) likewise vary in numerous parameters. They may be large or small, set deeply within the flower, variously disposed upon the petals, set at the orifice of the corolla, or strongly exserted (extended beyond the opening of the corolla tube). Flowers of the Polemoniaceae family are typically protandrous; *i.e.*, the stamens mature before the stigma (tip of the female organ, where pollen is deposited) is receptive to pollen. The pistil and resulting fruit may bear one seed, a few seeds, or many seeds. Generally, the mature fruit is a dry capsule that is often dehiscent (capable of splitting along definite lines) into three parts. The stigma, as is the case with the stamens, may be set deeply within the flower or long exserted. The seeds vary in size, weight, colour, markings, in the presence or absence of a mucilaginous coat, and may have a broad wing, especially in the tropical woody species. Pollen grains, elaborately sculptured, show variations that may exceed the variations in other features of the plants.

Drawing by M. Moran based on *(Phacelia minor, Amsinckia intermedia, Leptodactylon californicum)*. Reprinted by permission of the publisher, from Lyman Benson, *Plant Classification* (1957), D.C. Heath and Company, Lexington, Mass.; *(Hydrolea spinosa, Pholisma arenarium* (J. Hutchinson, *The Families of Flowering Plants*, vol. 1, the Macmillan Company from *Taxonomy of Vascular Plants* by G.M.H. Lawrence, Copyright 1951 by the Macmillan Company; *(Myosotis palustris)* after Eichler, Berg and Schmidt, Gurke, Gilg, Warburg, Hegi, Lawrence in A. Engler, *Syllabus der Pflanzenfamilien*, Gebruder Borntraeger, Stuttgart; *(Pholisma depressum* parasitic on *Corethrogyne filaginifolia)* originally published by the University of California Press, reprinted by permission of the Regents of the University of California from J. Kuijt, *The Biology of Parasitic Flowering Plants* (1969); *(Loeselia mexicana, Linanthus androsaceus)* K. and V. Grant, *Flower Pollination in the Phlox Family*, p. 128, p. 110 (1965), Columbia University Press, reprinted by permission of the publisher



Representative plants and structures of five of the eight families of the order Polemoniales.

Other variations. The previous examples, taken mostly from the family Polemoniaceae, reflect some of the basic features of the order as a whole. There are, however, some outstanding specializations found in other families that deserve additional mention. In the matter of floral biology, the family Boraginaceae shows heterostyly. This is the condition of the flower in which members of a species population differ in the length of the stigma relative to the length of the stamens. Most often encountered is the distylous condition, wherein the style is very short and the anthers stand near the throat of the corolla (a condition known as thrum form), or the styles are elongate, with the stigma standing in the throat of the corolla and the anthers disposed in the base of the flower (pin form). This is an arrangement that enhances crossing because pollen from a thrum plant normally leads to fertilization in a pin flower, and pollen from pin leads to fertilization in thrum. In regard to seed dispersal, many unusual modifications related to dispersal by fur-bearing animals are found in the Boraginaceae. In most members of the family the two-carpellary ovary divides into four nutlets at maturity—the fruit wall of each nutlet tightly embracing a single seed. The outer surfaces of the nutlets bear hooks, barbs, bristles, and similar structures that adhere to fur. The family Loasaceae, which is usually not placed in this order, differs in having an inferior ovary (*i.e.*, the sepals, petals, and stamens appear to attach at the upper end of the ovary, their basal portions enclosing it). Along with certain members of the families Boraginaceae and Hydrophyllaceae, they usually have harsh, often stinging, trichomes (plant hairs).

NATURAL HISTORY

Pollination. Variation in floral features within the order is matched by variation in pollination mechanisms. The range of pollinators known in the order includes bees, hummingbirds, flies, beetles, butterflies, moths, and bats. Many species are self-pollinated. Floral structure is closely correlated with mode of pollination, and the order exhibits nearly all the floral modifications associated with such a diversity of pollinators (see POLLINATION).

Dispersal. Seed-dispersal mechanisms in the Polemoniales order are varied but not nearly so diversified as the pollination mechanisms. In nearly all species the fruit is a dehiscent capsule, the rupturing of which allows a slow sifting of seeds from it, or else an explosive split scatters the seeds forcibly. In a few cases the fruit does not split and it (and its seed or seeds) is dispersed as a unit. Once liberated from the capsule, the seeds may simply fall to the ground immediately about the parent plant, but those with explosive capsules may eject the seeds a metre or more. In both of these cases the seeds are subject to being carried about by animals, wind, water, or other agents. The tropical members of the order with winged seeds may be dispersed great distances by strong winds. A peculiar adaptation of many annual species is the development of a mucilaginous layer on the outside of the seed coat. When wet, this layer may increase in thickness to the diameter of the seed itself and is very sticky. These sticky seeds may be eaten, especially by birds, and may, by chance, stick to their feet or beaks and thus be carried great distances. Some North American species are represented by disjunct (widely separated) populations in South America, and it is reasoned that they may have been established there by transport on migratory birds. This must be thought of as the extreme form of dispersal, and most movement is probably the result of fitful small jumps repeated over thousands of generations.

Agents of seed dispersal

Physiology. The physiological diversity in the family is evident from the ecologically diverse habitats occupied by some populations. The perennials of Arctic or Alpine regions are adjusted to short, cold growing seasons and long dormant seasons. In contrast, some tropical species may grow the year around, with only slight variations in growth brought on by fluctuations in available moisture. Seasonal adjustment is very evident in the perennial plants of eastern North America, where a succession of species develop and flower, starting in the early spring, through summer, and, finally, to a species whose last flowers are killed by the first autumnal frost. The desert annuals have perhaps the most rigorously programmed

Life cycle of desert annuals

physiological systems. Their seeds germinate only when certain minimal moisture and temperature requirements have been satisfied. Then they grow rapidly to flowering and fruiting before the onset of the next drought period. Most populations vary in their requirements for germination, so that not all seeds germinate every year. Many seeds, in fact, may remain dormant for several years before germination occurs. This variation ensures that some seeds will always be available in the event of a crop failure caused by the extreme rigours of the fluctuating desert environment.

Parasitism. Two families in the order, Cuscutaceae and Lennoaceae, consist of totally parasitic plants. With the exception of certain species of *Cuscuta* (dodder or love vine), they are rarely observed. The dodders are thought to be closely related to, and derived from, the twining members of the family Convolvulaceae. In *Convolvulus* the advantage of twining seems to be purely mechanical; whereas in the nearly leafless and achlorophyllous (*i.e.*, without chlorophyll—nonphotosynthetic) *Cuscuta*, twining brings the stem of the parasite into intimate contact with the host plant, where specialized organs (haustoria) penetrate the nutritive as well as the water-conducting tissue of the host, withdrawing sustaining substances. Dodder plants flower and produce an abundance of seeds with ample reserve food in the endosperm (starchy nutrient tissues). Typically, the seeds germinate in soil, producing first a club-shaped rootlike structure and later a wiry stem that may establish a new parasitic relationship with a suitable host. Soon the root dies, leaving only the haustorial connections established between parasite and host. Since dodder plants are lacking in the usual photosynthetic machinery, they die unless a satisfactory parasitic union is established. Seedlings, however, may live for several weeks and grow many centimetres before their successful establishment or their starvation occurs.

The family Lennoaceae shows no obvious resemblance to *Cuscuta* in its general morphology or mode of parasitism. Usually only the flowering portion of the plant emerges from the sandy soil where it grows, but an extensive underground root system is present. This system apparently ramifies randomly until host root systems are encountered. When a root of the parasite approaches a host root, a haustorial type of root system develops that penetrates into, and absorbs from, an appropriate host root. The mechanisms of lateral rather than vertical root growth and the sensing of an appropriate host plant are unresolved questions regarding the physiology of these plants.

EVOLUTION

There is very little fossil evidence that would be helpful in understanding the relationships of the order. The few fossils known are isolated organs scattered geographically and stratigraphically. Consequently, phylogenetic relationships must be deduced by comparing modern species.

It is conceded by phylogenists that the order Polemoniales is an advanced plant group. Its high level of specialization is evident in the flowers, fruits, vegetative morphology, and physiological adaptations. Because of its advanced state, however, it is difficult to deduce a group ancestral to the order and to demark clear lines of relationship within the order. These are problems common among many flowering plant orders. Taxonomic groups referred to as Polemoniales, or that have included all or most of the families appearing here, range in size from three families to 26 families. The eight families here comprising the order bear a relationship to the order Gentianales (*q.v.*), especially in several floral and anatomical characters. It is difficult, however, to conceive of one of these living orders as having been derived from the other. The two orders could better be thought of as having been derived from a common ancestor. Some of the more primitive members of the groups within—or closely related to—the orders Rosales and Dilleniales, such as the family Pittosporaceae, have been suggested as possible ancestors from which both the orders Gentianales and Polemoniales arose.

Relationships to other plant orders

The lines of relationship within the order are also obscure, since most of the families show parallel specializations. It seems, therefore, unrealistic to attempt a sequential arrangement of the families Convolvulaceae, Polemoniaceae, Boraginaceae, Hydrophyllaceae, and Loasaceae, when surely none exists. On the other hand, the family Cuscutaceae seems clearly to be a parasitic derivative of the family Convolvulaceae, whereas the families Hoplestigmataceae and Lennoaceae show a strong similarity to either or both of the families Boraginaceae and Hydrophyllaceae.

CLASSIFICATION

Distinguishing taxonomic features. The families of the Polemoniales order are delimited primarily on the basis of flower and fruit characters, their vegetative, chemical, and ecological considerations being of secondary importance. Most families have gynoecia (female apparatus—stigma, style, and ovary) of two carpels, but the family Lennoaceae is exceptional in having six to 14 carpels, the family Polemoniaceae regularly has three, and the family Loasaceae may have as many as five. In the latter family, the ovary is inferior, but in the rest of the order the ovary is superior. The petal number (or lobes of the corolla) is mostly five in the order, but the family Loasaceae often has from four to ten, and the family Hoplestigmataceae has from 11 to 14. Occasional individuals or even some unusual species may have either four or six petals in other families. In the families Hydrophyllaceae and Boraginaceae the petals often bear appendages or crests. Nearly all members of the order have five stamens per flower, but the family Loasaceae often has many, and the Hoplestigmataceae family has from 20 to 30 stamens. The tendency in the Boraginaceae for the fruit to ripen as four-dividing nutlets is unique to that family within this order, as is the gynobasic style (*i.e.*, the style arises from the base of the four-parted ovary). Vegetatively, the Hydrophyllaceae, Boraginaceae, and Loasaceae families often possess harsh epidermal trichomes (plant hairs) that, in some species, have stinging capability. Woodiness is lacking only in the parasites, the families Lennoaceae and Cuscutaceae. Both species of the family Hoplestigmataceae are trees. In the remaining five families, trees and shrubs or woody vines occur, along with herbaceous species.

Annotated classification. The classification presented here includes only a small portion of the families usually treated together as an order in most former classification systems, along with three families of doubtful affinity. On practical and theoretical grounds, the first five families presented below are difficult to separate from the remainder of the approximately 26 families of those systems. The other three families (Lennoaceae, Hoplestigmataceae, Loasaceae) have rarely been included with them. They do bear some relationship to the families Boraginaceae and Hydrophyllaceae, however, so they may rest about as comfortably in the Polemoniales as they would in other orders.

Scope of classification system

ORDER POLEMONIALES

Flowering plants with mostly sympetalous (fused) corollas. Flowers generally 5-merous (in parts of 5—5 sepals, 5 petals, etc.), actinomorphic (radially symmetrical), stamens 5 and alternate with corolla lobes, or numerous. Fruit a capsule or of 1 to 4 nutlets. Plants typically trees, shrubs, vines, or herbs, except for a twining parasite and a few root parasites. Primarily in North and South America, but some also in Eurasia and in Africa as well. There are eight families, 208 genera, and 4,446 species.

Family Polemoniaceae (phlox family)

Herbs, shrubs, small trees, or vines; leaves mostly alternate, rarely opposite, simple or compound. Flowers bisexual, regular or rarely bilabiate (2-lipped), 5-merous, inflorescence a cyme or head, flowers rarely solitary; sepals 5, united; petals 5, united, often rotate (wheel-shaped, often flat and circular in outline), campanulate (bell-shaped), salverform (trumpet-shaped), or tubular, contorted in bud. Stamens 5, adnate (fused) to corolla, sometimes with unequal insertion (*i.e.*, they arise at different points along the corolla tube), exserted (extending past the corolla opening), or included in corolla tube. Ovary superior, carpels 3 or rarely 1 or 2 by reduction, surrounded by a nectariferous (nectar producing) disk, style 1, stigmas 3, rarely 1 or 2. Fruit a dehiscent capsule; seeds 1 to

numerous, large or small, sometimes winged, mostly small, often mucilaginous when wet; endosperm present; embryo straight or slightly curved. Eighteen genera, over 300 species, mostly in North and South America, but a few species in Eurasia. Some prominent examples include the genera *Polemonium, Phlox, Gilia,* and *Linanthus.*

Family Convolvulaceae (morning glory family)

Herbs, shrubs, trees, or vines; chiefly twining or trailing, with milky juice. Leaves alternate, simple, entire, or lobed. Flowers bisexual, regular, rarely irregular, 5-merous, inflorescence a cyme (a flower cluster that matures from the apex downward) or flowers solitary; sepals 5, united; petals 5, united, often campanulate or cylindrical, contorted in bud; stamens 5, adnate to corolla, exserted or included in corolla tube. Ovary superior, carpels 1 to 4, mostly 2, surrounded by a nectariferous disk, style 1, stigmas 1 or 2; fruit a dehiscent capsule. Seeds few, large, often hairy; endosperm present; embryo curved, cotyledons large and crumpled. Fifty genera, over 1,400 species, widely distributed in the tropics and temperate regions. *Convolvulus, Ipomoea, Dichondra.*

Family Cuscutaceae (dodder family)

Parasitic herbs lacking chlorophyll. Stems slender, twining, threadlike, often in great masses, with haustorial connections to host, but rootless after seedling stage. Leaves reduced minute scales. Flowers bisexual, regular, 5-merous (rarely 4-merous), white or waxy, inflorescence a compact cyme; sepals 5, united; petals 5, rarely 4, united, corolla campanulate or cylindrical, often appendaged, lobes imbricate (overlapped) in bud. Stamens 5, inserted at corolla throat, alternate with corolla lobes. Ovary superior, carpels 2, styles 2 or united basally; fruit a capsule, circumscissile (opening by a regular transverse circular line of division) or irregularly dehiscent. Seeds 1 to 4, embryo coiled within the endosperm, cotyledons minute or lacking. One genus (*Cuscuta*) of over 150 species, widely distributed, occasionally a noxious pest of agricultural crops or indigenous flora.

Family Hydrophyllaceae (waterleaf family)

Herbs or shrubs. Leaves usually alternate, but often opposite or in a rosette, simple or deeply lobed to compound, coarsely hairy, sometimes with stinging hairs. Flowers bisexual, regular, 5-merous; inflorescence, usually of coiled cymes. Sepals 5, united, often with appendages between the lobes; petals 5, united, rotate, campanulate, or tubular. Stamens 5, adnate to base of corolla, often associated with corolla appendages, exserted or included. Ovary superior, carpels 2, style 1, stigmas 2. Fruit a dehiscent capsule; seeds numerous, highly ornamented; endosperm present; embryo small and straight. Twenty genera, nearly 300 species, mostly in North America, with a few on other continents but none in Europe or Australia. Important examples of genera include *Hydrophyllum, Phacelia, Nemophila, Nama,* and *Eriodictyon.*

Family Boraginaceae (borage family)

Herbs, shrubs, or large trees. Leaves usually simple, entire, opposite, alternate, or whorled, often coarsely hairy or scabrous (scaly). Flowers bisexual, regular, 5-merous; inflorescence often a coiled cyme. Sepals 5, united, slightly irregular, petals 5, united, crested or appendaged, rotate, campanulate, salverform, or tubular. Stamens 5, adnate to corolla, usually included in corolla, but may be long-exserted in a few, sometimes heterostylous (differing lengths of stamens and styles in different flowers—the "thrum" and "pin" forms mentioned in the article). Ovary superior, carpels 2, often deeply lobed and style gynobasic, style and stigma usually 1. Fruit usually of 1 to 4 smooth or highly ornamented nutlets; endosperm present or absent; embryo straight or curved. One hundred genera and 2,000 species, with worldwide distribution but prominent in North America and the Mediterranean region. *Cordia, Ehretia, Heliotropium, Lithospermum, Echium, Cryptantha,* and *Cynoglossum* are some important genera.

Family Lennoaceae (sand food family)

Fleshy erect parasitic herbs, lacking chlorophyll, attached to roots of other plants. Leaves simple bractlike scales. Flowers bisexual, regular, 5- to 8-merous; inflorescence a compact spike or head. Sepals 5 to 8 linear, united at base; petals 5 to 8, united into a flaring tubular corolla. Stamens 5 to 8, inserted at corolla throat, anthers barely exserted. Carpels 5 to 15, united, placental column (the area of ovule attachment) massive, style 1, stigma 1, but with several small lobes. Fruit capsule-like but eventually separating into several nutlets; seeds small; endosperm present; embryo undifferentiated. Three genera (*Lennoa, Ammobroma, Pholisma*), 4 species, distributed in western North America.

Family Hoplestigmataceae

Trees. Leaves large, simple, entire alternate. Flowers bisexual, regular, number of floral organs indeterminate, inflorescence cymose. Sepals united, splitting irregularly at anthesis (time at which ovules are fertilized); petals 11–14, united basally, irregularly imbricate. Stamens numerous, irregularly arranged, not adnate to corolla. Ovary superior, carpels 2, united; placentae parietal (ovule attachment on the inner wall of ovary); style 1, short; stigmas 2, elongate. Fruit a drupe; seed with endosperm; embryo large, straight. One genus (*Hoplestigma*), 2 species, in western tropical Africa.

Family Loasaceae (loasa family)

Herbs or shrubs. Leaves opposite or alternate, simple or divided, often coarsely hairy, sometimes stinging. Flowers bisexual, regular, 4- to 5-merous; inflorescence a cyme or head, or flowers solitary; floral tube adnate to ovary; sepal lobes 4 to 5, usually persistent on fruit; petals 4 to 5, often prominently clawed. Stamens 5 to many, often in 5 bundles opposite the petals. Ovary inferior, carpels united, style 1, stigmas 1 to 5, placentas parietal; fruit a capsule; seeds 1 to many; endosperm prominent; embryo straight. Fifteen genera and 250 species, in North and South America, except for 1 genus (*Kissenia*) in Africa. *Mentzelia, Loasa,* and *Cajophora* are important examples of genera.

Critical appraisal. The families grouped here as the order Polemoniales have been treated in a variety of contexts in different systems. Most systems, however, include the families Convolvulaceae, Cuscutaceae, Polemoniaceae, Hydrophyllaceae, and Boraginaceae in the same order, whether it is designated Polemoniales, Tubiflorae, Solanales, or by some other name. In addition, several families such as Solanaceae, Nolanaceae, Verbenaceae, Scrophulariaceae, and other families with tubular corollas and superior ovaries are often included in an ordinal group along with more typical Polemoniales. The family Lennoaceae has sometimes been placed in the order Ericales, but its relationship to the Hydrophyllaceae and Boraginaceae families now seems firmly established. The arrangement presented here deviates most from other proposed systems in the inclusion of the families Hoplestigmataceae and Loasaceae. That the relationship of the first is uncertain is clearly exemplified by its inclusion in the orders Lamiales, Violales, Bixales, and Ebenales by various authorities. Similarly, the family Loasaceae has been placed in the orders Cistales, Violales, and Loasales by different workers. It seems particularly out of place in the Polemoniales because of its inferior ovary, separate petals, and multiple stamens; however, its resemblance to the families Hydrophyllaceae and Boraginaceae in characters of indument (covering of plant hairs), seed, embryo, and endosperm is considered adequate to permit its placement in the Polemoniales order as defined here.

Uncertain status of some families

These variations in treatment of the families are the result of having inadequate data and also disparate opinions about the value of certain characters in phylogenetic taxonomy. Until more data are available, particularly additions of critical material to the fossil record, concepts of the primitive condition of angiosperms (flowering plants) will have to remain theoretical. In view of the variations in thought on this subject, there is certain to be a continuing disparity in the evaluation of the evolutionary significance of plant characters and their utilization in the construction of classifications purporting to be phylogenetic (*i.e.*, reflecting actual evolutionary relationships).

BIBLIOGRAPHY. LEROY ABRAMS, *Illustrated Flora of the Pacific States,* vol. 3, *Geraniums to Figworts* (1951), an illustrated guide to many western North American Polemoniales; K. FAEGRI and L. VAN DER PIJL, *The Principles of Pollination Ecology* (1966), a critical treatment of pollination biology with several detailed case histories; VERNE GRANT, *Natural History of the Phlox Family,* vol. 1, *Systematic Botany* (1959), an exhaustive presentation of the systematic botany of the family Polemoniaceae; and with KAREN GRANT, *Flower Pollination in the Phlox Family* (1965), a summary of pollination biology in the Polemoniaceae including detailed documentary descriptions for many species of the family, especially from western North America; JOB KUIJT, *The Biology of Parasitic Flowering Plants* (1969), a worldwide treatment of all groups of parasitic angiosperms including original observations on the family Lennoaceae; ROBERT SCHERY, *Plants for Man* (1952), an economic botany textbook that includes many obscure and little-known aspects of plant utilization; L. VAN DER PIJL, *Principles of Dispersal in Higher Plants* (1969), a critical presentation of the biology of dispersal in plants emphasizing ecological and evolutionary aspects.

(D.M.Sm.)

Police

The term police is now used to denote a body of people organized to maintain civil order and to investigate breaches of the law. In the 18th and 19th centuries the term was used more generally to denote the internal civil administration of a state or a town, including (to quote Adam Smith) "the domestic regulation of such matters as taxes, bounties, monopolies and the exclusive privileges of corporations." In France and countries influenced by the Napoleonic style of administration, the duties of the police have been more widely defined to include the enforcement of regulations concerning public health, the erection of buildings, and so forth.

All societies need some means whereby order is maintained. In the smallest societies lacking written laws, informal sanctions discourage deviation. In the simplest forms of state organization, the ruler has agents to enforce his decisions. The police function is universal in society, though it is only in the larger and more complex states of modern times that full-time officials are appointed with special police responsibilities regulated by law. In modern industrial societies most policemen are enrolled in state, provincial, or municipal forces concerned with criminal investigation, traffic regulation, and preventive patrol. A central government generally has special sections of personnel responsible for the collection of political information, counterespionage, the protection of important persons, and so forth; and not all of this work can be accounted policing in its modern connotation. The ordinary police services are often supplemented by special police forces, separately organized, for the supervision of railways, harbours, docks, airports, and government establishments. Recent years have also seen a rapid growth in private security organizations, though countries differ in the extent to which they accord legal recognition to the creation of private police forces for the supervision of commercial establishments, university campuses, and the like.

POLICE AND SOCIETY

The diversity of police activities. There is a remarkable historical, geographical, and organizational diversity in the activities of persons who are, or have been, counted as policemen. Within any one country the work of policemen today is very different from what it was 200 years ago. There are also major differences between countries—policing New York City is bound to have little in common with policing the Solomon Islands. The police forces of some countries, having been created by the central government for its own purposes, have grown from the centre outward. Elsewhere, communities have maintained local police arrangements, and the central government has stepped in to consolidate and coordinate them so that growth has been from the outside inward. The diversity is so great that the onlooker may wonder if the different kinds of activity and organization have sufficient in common to be classified together.

Principles of police accountability and adaptation

The principle of police accountability is helpful in explaining the diversity noted above. The problem is summarized in a Latin question: *Quis custodiet ipsos custodes?* ("Who guards the guardians?"). From one point of view, it is desirable that a police force should be as efficient as possible. It should be rationally and economically organized so that specialist services can be developed without any overlap or conflict between the responsibilities of different units. Such considerations favour the establishment of national police forces, which can take advantage of the economies of scale in training, promotion schemes, specialist organization, and so on. But it often has been felt that the existence of a national force places too much power in the hands of those who direct it; that there is a danger that the government will use its control of the police to keep itself in office; and that the police will not be accountable to the public. Therefore, some countries favour a local basis for police organization.

A second principle that helps explain the diversity of police activities in different societies is that of police adaptation. If the society is agricultural, as in rural India, the police will be concerned with the state of the crops, irrigation of fields, the conditions of roads and paths, private feuds and quarrels, the registration of births and deaths, feasts, fairs, and all manner of private and public events. If local landowners are influential, as they were in 19th-century England, much police effort will be expended in the enforcement of laws against poaching, and much less in the investigation of ways in which the behaviour of employers infringes the rights of employees. If offenses are committed in moving vehicles, then the police must be motorized. If criminals use firearms, then the police are likely to be armed. Police activity must be adapted to the society that is being policed. The spread of industrialization in the 20th century has meant that otherwise different societies have adopted similar modes of production and have experienced similar problems. The control of automobile traffic has aspects common to Los Angeles, Tokyo, and Leningrad. Police organizations in different countries therefore tend to respond in increasingly similar ways. A later section will outline constant features in the activity of policemen in industrial societies.

Social control in small communities. Most people obey most laws willingly, whether or not there is a policeman looking on. They comply with the laws because they believe them right and perceive that in the long run it is in everybody's interest to observe them. But this voluntary compliance is more general in certain kinds of community and with respect to certain kinds of law. In small communities in which everyone knows everyone else, people who live up to the shared ideals are rewarded by the esteem of their fellows. If they fall short of other people's expectations, life is made more difficult because others are no longer so ready to help them in times of trouble. This pattern of informal rewards and punishments is, in all kinds of society, the most potent force in law enforcement, but it is stronger in the small community because there people meet one another in different roles. The village grocer may represent his district on the county council; and he may meet his neighbour, the schoolteacher, in his capacity of councillor, or of grocer, or of parent of a school child, or of neighbour, or in one of many other roles. Should there be two grocers, people may feel obliged to patronize the one who attends the same church as they do. Thus, in the village there is a relatively dense network of social ties, which imposes more restraint upon social deviance of any kind.

The forces that order the life of the small community make the policeman's task much easier. Instead of imposing requirements, he has only to lubricate the mechanisms inherent in social relations, and police action is needed only when the informal controls have proven insufficient. The ideal of the well-integrated local community has, therefore, had a romantic appeal for some social philosophers, and many policemen sometimes think nostalgically of the life of the village constable. Though it is an unreal abstraction—like the idea of a frictionless pendulum—the notion of a self-policing community can serve as a model when studying what is required of the police. And the notion need not be limited to local communities. Professions such as law and medicine also form communities of practitioners who are able to police themselves with respect to the practice of professional skill, but, naturally, the members cannot police one another's conduct in respect to nonprofessional matters such as motoring offenses.

Social control in large societies. The bigger a society becomes, the weaker its self-policing elements are likely to be. In the first place, when people are involved with those whom they are unlikely ever to meet again, there may seem to be less reward for honest dealing; and, if one party feels he has been cheated, there is less chance of his being able to bring informal pressures to bear upon the offender. Most contacts are impersonal, and since the parties do not meet in other capacities, the informal controls cannot play so great a part. In the second place, with technological progress there comes a need for new laws. With the growth of automobile traffic, for example, many new laws governing vehicle construction, licensing, and

manner of driving become necessary. Because many such laws do not have the same moral significance as laws penalizing personal violence or dishonesty, the reporting of offenses cannot be left to members of the public. Formal controls are needed in the shape of trained inspectors, officials, and traffic police. When the legislature declares that certain actions are in the future to count as offenses, it increases the amount of crime. There is a danger here that people who are convicted of the one offense—say, a motoring offense—may feel aggrieved and in the future be less willing to cooperate with the police or to obey the law when they are not being observed. The self-policing element is therefore weakened. In the third place, as societies grow it becomes more difficult to relate public interests and private interests. If an employer catches a group of his employees committing an offense within the workplace, he may not wish to call in the police. He may believe that, if he deals with it in his own way, production and profit may suffer less; the public interest, however, is not served thereby. Similarly, if valuable property is stolen, an insurance company may be willing to pay for its return without obtaining the information that would make it possible for the criminals to be prosecuted. Supervision of the way private bodies respond to crime is necessary if the total level of crime is to be reduced.

The hazards of centralized policing. For these and other reasons, it would appear that communities cannot wholly be left to police themselves, and that a central government should impose its standards upon local communities. If, however, the central government itself is corrupt or self-seeking, the outcome will be worse than if communities were left to enforce law and order according to their own lights. Moreover, good intentions on the part of the central government are in themselves obviously not sufficient. Much will depend upon whether it can induce its agents to use their powers in desired ways.

French police in the ancien régime

A striking example of action by a central government to enforce its own standards is provided by the French political police. Though some writers have looked critically at the use of police by French kings to effect a personal rule, the fact is that community controls could not have achieved the same end, and there would seem to be little doubt that in the 18th century the citizen of Paris was far better protected from crime than, for instance, the citizen of London. Louis XIV had 40 inspectors of police and many paid agents, who supplied the king with masses of detailed information about the conduct of private persons. Louis then issued *lettres de cachet*, ordering the imprisonment of persons of whose guilt he was satisfied even though there was insufficient evidence to bring them to trial; returning to their hometowns noble ladies who had been misbehaving; and, generally, exercising a kind of summary justice. The attendant influx of criminal information and salacious anecdotes concerning all orders in society helped to create the image of an omniscient political police. Indeed, at one time it was estimated that one-quarter of all the housemaids and lackeys in Paris were receiving money from the police in return for information about their employers and acquaintances. The police chief Sartine once boasted to Louis XV, "Sire, whenever three people speak to one another in the street, one of them will be mine." At the same time the ordinary police took an active part in many aspects of the city's life. The lieutenants general of police built markets, paved and constructed streets, founded the Bourse (stock exchange), moved a cemetery, provided street lighting, established a system of state pawnshops, created a hospital for children, founded a veterinary college, built schools for children of the poor, sought work for the unemployed, and organized fire and river rescue services. The police took stock of the city's food supplies and forbade wine sellers to have leaden counters and milkmen to use copper vessels. They censored posters and regulated the employment of foster mothers. They persecuted Protestants, controlled the sale of lottery tickets, and inspected drains. They had the streets cleaned. They supervised cabdrivers, made regulations about dogs, checked food prices, surveyed dangerous buildings, read people's letters in the course of post, and even stopped fishmongers from selling oysters between April 30 and September 10. The police inspected prisons, read theological works for signs of heresy, and had a financial agency for the encouragement of overseas trade.

French police after the Revolution

When the Revolution came, the French police made not one effective move against the revolutionaries; they identified themselves with the revolutionary government rather than with the old law. In 1796 a new Ministry of the General Police of the Republic was created and soon came under the control of Joseph Fouché, who used his position to establish a political police making extensive use of paid informers. As a consequence, a separate police body, the Prefecture of Police of Paris, had to be created to take responsibility for more conventional duties. One agent within the Prefecture, François-Eugène Vidocq, became head of the Police Judiciaire (the department for criminal investigations) and was celebrated for his disguises. Generations after his death his name was still execrated by ordinary Frenchmen as the epitome of police espionage. Though the legend bears little relation to Vidocq's actual history, it testifies to the fears of citizens that the state may use the police to serve the ends of those in power rather than the community as a whole.

Present-day schemes of police administration. Although the belief still persists that decentralization favours police accountability (and thus serves to steer clear of the abuses exemplified by the French police of an earlier day), some centralization is required for economic and operational efficiency. As the period of training needed for effective police work has grown longer and work hours have been reduced, police manpower has become more expensive; and it has become important to see that men are not employed on tasks for which less highly trained employees could be used. Consequently, policemen in most industrial countries are now assisted by a host of auxiliaries: typists, photographers, mechanics, fingerprinting technicians, forensic scientists, traffic wardens, and so on. It has become advantageous to have national schemes for fixing the pay scales of both policemen and auxiliaries; for recruiting; for seeing that similar standards are employed all over the nation; for establishing a common system of criminal records; for developing national facilities for higher training and the selection of top administrators; for forming specialist squads to investigate particular kinds of crime; for research; and, generally, for seeing that the police service keeps abreast of the changing pattern of criminal activity. Such tendencies make for increased centralization, and they operate in a similar fashion in all industrialized countries. As a result, regardless of the way in which police services have started out—whether, like the French, they have been highly centralized or, like the British, have taken community policing as their model—all have tended to develop common features; and within apparently different systems there are comparable devices for balancing central and local interests.

Police reorganization in West Germany and Japan

The examples of police reorganization in West Germany and Japan illustrate the way in which demands for economic rationality serve to offset the belief that decentralization favours police accountability. In 1949 the Council of the Allied High Commission in West Germany ruled that the police were to be administered at the level of municipalities and rural districts. There were to be no provincial police or national police, lest they assume a paramilitary character or otherwise threaten either democratic government or the security of the occupation forces. The police were to be responsible for the maintenance of public order and safety, the prevention of crime, and the bringing of offenders to justice. But in 1950 provision was made for provincial (*länder*) police forces. Subsequently the Federal Criminal Police Bureau (Bundeskriminalamt) was set up in Wiesbaden to coordinate the detective work of local police forces and prosecute criminals whose activities were not confined to one province. A national Frontier Police Force (Bundesgrenzschutz) was established and became the biggest police force in the republic. Organized in companies, it is controlled by the minister of the interior. An Emergency

Police (Bereitschaftspolizei) was created to deal with disasters, such as floods and landslides, and with riots and other emergencies. The arms available to this force include machine guns. Though the provincial governments are responsible for personnel costs, the federal government provides the equipment and appoints the inspector general. Since the inspector general, on behalf of the federal government, is responsible for inspecting the provincial police, and, since the Emergency Police also has to enlist and train recruits, the central government has considerable influence over the entire police service. The training function is particularly important. Recruits first attend basic training courses at state police schools and then serve in the Emergency Police for approximately two years before returning to a local force. At the same time, the provinces watch very carefully to see that their prerogatives in respect to police control are not impugned. Apart from the period of the Third Reich, local control has strong roots in German police tradition.

In Japan, in 1947, U.S. Gen. Douglas MacArthur introduced a new organization to decentralize the police and subject them to democratic control. A national rural police was established, but all towns and cities of over 5,000 inhabitants were to have their own police forces. In this way 1,605 communities acquired their own police. Because this multiplicity led to overlapping jurisdiction and confusion, however, all police forces in 1954 were amalgamated into a single national service administered in 45 prefectural districts. At the same time, the chairman of the National Public Safety Commission was given the status of a minister of state so that the government might assume responsibility for public peace and order.

MAJOR MODELS OF POLICE ADMINISTRATIVE ORGANIZATION

Systems with established police forces. *France.* In France, there are two national police forces: there is the Police Nationale, with units in each *département* commanded by the prefect of the *département*, who, in turn, is responsible in police matters to the director general of the Police Nationale; and there is the Gendarmerie Nationale. The Police Nationale was formed in July 1966 by the amalgamation of the Prefecture of Police of Paris and the Sûreté Nationale, which had been responsible for policing the provincial towns and for certain specialized work. The director general's headquarters includes seven divisions covering general intelligence, judicial police, public safety, security, regulations, training, and personnel. The National Police has a strength of some 91,000, the Gendarmerie 61,000. In the larger towns there are *Gardiens de la Paix* who serve as municipal police, responsible for preventive patrol and the control of traffic.

The Paris police has five sections: municipal police; the Police Judiciaire (for criminal investigation); an intelligence unit; a technical unit; a unit for economic, financial, and commercial offenses; and a section responsible for the safety of the president and of other important persons. Within Paris there are 20,000 Gardiens de la Paix (guardians of the peace) who serve as municipal police, responsible for preventive patrol and the control of traffic. In the provincial *départements*, while the commissaire de police is responsible to the prefect, he is expected to cooperate with the mayor and the municipal administration. For preventive work and the maintenance of order (traffic, patrols, meetings, and so on), the commissaire has at his disposal a Corps Urbain composed of uniformed Gardiens de la Paix and higher ranking officers; for criminal investigation and administrative work he has plain-clothed Officiers de Police and Inspecteurs who form the Sûreté Urbain. Major decisions on working methods and administration are, however, taken at the Paris headquarters.

The Police Nationale also has several specialized divisions. One is the provincial Police Judiciaire, which is organized in 17 regional branches. Others include intelligence and counterintelligence services; mobile reserves, or riot police; the Republican Security Companies (CRS), whose 13,500 men, organized in companies of about 200 Gardiens each, are housed in barracks. They became notorious as a result of brutality in dealing with student demonstrators in Paris in May 1968.

Gendarmerie Nationale

The Gendarmerie Nationale is composed of soldiers formed in strictly disciplined units and quartered in barracks. The force includes the Mobile Gendarmerie and the Departmental Gendarmerie, a section of which covers each military area (several *départements*) under the command of a general. The central headquarters for the Gendarmerie, under the direct control of the minister for the armed forces, is charged with planning, staff organization, and administration. The mayors of certain communes have communal or rural policemen at their disposal; but, as a rule, police work in country districts and in towns with less than 10,000 inhabitants is undertaken by gendarmes. They also police certain main roads. Some are appointed as officers of the judicial police and are thereby empowered to inquire into and deal with all types of crime, including murder, kidnapping, wounding, and the like.

These divisions within the French police service, and the separate ties that units have with local administrators, suggest that, though centralized, the French police is far from monolithic. There are many opportunities for one police body to check upon the activities of another and to prevent any one organization from being unduly corrupted by its powers.

Belgium. Belgium, another country with an outwardly centralized police system, displays similar features. There, a Judicial Police, responsible to the minister of justice, is staffed by civil servants and concerned with the investigation of crimes and with bringing offenders before the courts and takes no part in administrative police work. Separate police forces exist for the railways, state security, the military, and the protection of atomic research establishments. The Gendarmerie is part of the kingdom's armed forces; it is responsible to the minister of the interior for its duties as administrative police, the execution of laws and arrests, and the maintenance of public order; to the minister of justice for its duties as judicial police, including the national security; and to the minister of national defense for duties concerning military deserters and mobilization arrangements. Unlike the Communal Police, who are appointed by the various towns and communities, members of the Gendarmerie have power to act in all territories of the kingdom. Though financed by local communes, these policemen are subject to controls exercised by the minister of the interior.

Italy. Italy, too, has several national police forces. The Public Security Guards (Guardia di Pubblica Sicurezza), which is part of the armed forces, is charged with the maintenance of internal public security and order, with the protection of life and property, and with preventing and checking crime and gathering evidence in criminal matters. Although the force is of a semimilitary character, it performs all normal police functions and is responsible to the minister of the interior. The Corps of Carabinieri is a military-like body housed in barracks; it is responsible to the minister of defense in certain matters, but to the minister of the interior in regard to police duties. As an executive organization for the detection of crime, the corps carries out duties assigned to it by law, under the direction of judicial authorities to whom it reports on investigated crimes. At every level, commanders of the corps must maintain liaison with the army, political, judicial, and police authorities. A Guardia di Finanza is charged principally with protecting the national frontiers from activities connected with smuggling, counterfeiting, illegal entry, and tax evasions. In addition to the national forces, municipalities maintain their own police, the Vigili Urbani, which undertake certain kinds of municipal administration but are increasingly concerned with traffic control. Frequent attempts have been made to amalgamate the Vigili with the Public Security police, but the municipalities have resisted such proposals.

Carabinieri

Spain. Spain also has several police forces with interlocking responsibilities. The two principal forces, organized on military lines, are equipped with arms and are

closely associated with the armed forces. The Civil Guard (Guardia Civil) is responsible for maintaining security and order and repressing crime in the rural areas. The Armed and Traffic Police (Policía Armada y de Tráfico) has similar responsibilities in the towns. The Civil Guard operates with detachments assigned to local posts, being supplemented by mobile reserves that keep order on the highways. The Armed Police patrol the towns. Superior "agents" belonging to a third State Police Corps (Cuerpo General de Policía) operate from stations of the Armed Police. They are civil servants who work in plain clothes on criminal investigation duties. In the towns there are also unarmed municipal police, paid out of the municipal budget, who are responsible to the mayor. If they arrest a criminal they hand him over to the Armed Police.

Japan. Japan's police force was a centralized, national organization under the minister of home affairs from 1871 to 1947, when, as noted above, the U.S. occupation forces initiated a plan to decentralize and "democratize" the police. When the experiment failed, all police forces were amalgamated in 1954 and placed under the direction of the National Public Safety Commission (NPSC) and the National Police Agency (NPA). Today, branches of the NPSC and the NPA operate in all the nation's 46 prefectures.

The NPA is responsible for directing general police activities, though each prefecture enjoys considerable autonomy. On the other hand, the NPSC exists primarily to guarantee the democratic administration and the political neutrality of the police. It has power to intervene in the operation, administration, financing, and personnel appointment of police forces in the various prefectures. The chairman of the NPSC is a minister of state directly responsible to the prime minister. (The latter may, with the Diet's approval, control all or sections of the police during a state of national emergency.) Likewise, the director general of the NPA is appointed by the NPSC with the prime minister's approval.

Thus, Japan's police, which in 1970 consisted of about 175,000 officers, is, on the one hand, a decentralized, autonomous force and, on the other, a national, government-directed force. In addition to the general police charged with law enforcement, there are various agencies that carry out specialized work. These include the riot police (numbering 9,700 in 1970).

Scandinavian countries. The advantages of a unified police service can most easily be combined with accountability to the public in small homogeneous nations with a relatively high level of civic education. In the Scandinavian countries—for instance, Denmark, Finland, Norway, and Sweden—all have national police forces and no separate forces for railways, docks, or similar establishments. In Finland the police is an organ of the central government and is financed by the state. In Norway only about 10 percent of police costs are met by the municipalities; each of 54 police areas is under the command of a Politimester, who must hold a degree in law and who is responsible to the minister of justice and police. Before appointment as a constable, a person must have served as a police aspirant in one of the major forces, have successfully completed a ten-month course at the police school, and have undergone four months local training. In Denmark the central government appoints a commissioner of police, but the 72 local chief constables are independent commanders of the personnel placed at their disposal by the state.

Administrative reform in Sweden

The pressures for rationalization are well exemplified by the reform of police administration in Sweden. Because of its technological and social development, the country is experiencing an acute shortage of labour coupled with a high level of urbanization, leaving large provincial regions with very low population density. Since 1965 the police have been nationally administered, the number of local police districts has been reduced to 119, and the number of police stations to 510. This has been made possible by greatly increased use of motor vehicles and the expansion of the radio and telephone networks. But, because nearly 50 districts can muster fewer than 50 police officers, further rationalization is likely. Responsibilities for prosecuting and seizing the property of debtors have been removed from the police. The internal organization of the police now divides into two principal divisions responsible for patrol work and criminal investigation. Local police committees have been established on an advisory basis. The intention has been to create a forum in which the police chief can discuss law enforcement with representatives of local government. Local police budgets are considered by these committees. The police chiefs are responsible to the National Police Board, which reports to the Ministry of Justice.

Decentralized systems. *United Kingdom.* The United Kingdom of Great Britain and Northern Ireland has no national police force, although, in several important respects, the central government exercises strong controls over all police forces; and there has been a continuing tendency to eliminate small forces by amalgamating them with larger units. Different arrangements still obtain in various parts of the kingdom. Since 1829 the home secretary has been answerable to Parliament for the Metropolitan Police (21,000 officers in 1970), who patrol the London area. Other police forces in England and Wales numbered 226 in 1860, but by 1969 they had been reduced to 46 (78,000 officers). The home secretary can, if necessary, require a chief constable of one of these forces to retire in the interests of efficiency, but he cannot give them orders, and they are answerable in operational matters only to the laws of England. In Scotland in 1969 there were 22 forces (11,000 officers), and their position was in many respects similar to that of English provincial forces, with one major exception. In England anyone may prosecute in the courts; and, although the police often prosecute in many cities, the municipality employs its own lawyers to take in court all except minor traffic cases. In Scotland responsibility for deciding whether or not to prosecute lies with the lord advocate, one of the government's law officers, and with his local subordinates (procurators fiscal), to whom the police are accountable in regard to criminal investigation. In Northern Ireland there is a state police, the Royal Ulster Constabulary (3,000 officers), and its senior officer is responsible to the minister of home affairs. Within the United Kingdom there are a number of special police forces for docks, railways, and defense establishments, the members of which are sworn constables with police powers of arrest. Chief among these is the power to arrest a person on suspicion that he has committed a felony. Civilians are entitled to arrest persons whom they know to have committed a felony. The various factory police units and private security services have only civilian powers of arrest.

Combined local and central controls

One aspect of British police administration most likely to be of general interest is the way in which the centralization of certain services is combined with forms of local independence. A Royal Commission on the Police reported in 1962:

> There is a substantial case for creating a national police service. Its organization would be more logical, and a number of us think that it would prove to be a more effective instrument for fighting crime and handling road traffic. . . . We do not regard the creation of a national police service as constitutionally objectionable or politically dangerous.

One of the commission's members added:

> The danger in a democracy does not lie in a central police that is too strong, but in local forces that are too weak.

The commission recommended the continuance of separate local police forces. In defense of its decision it can probably be argued that the local system does provide a measure of democratic control, and that such control, once lost, probably could not be regained without a terrible struggle. Localism obliges police chiefs to seek the support of local councils and, therefore, makes them more responsive to public opinion than they might otherwise be. Provided that the inspector of constabulary certifies that a force is efficient, the central government meets half its cost. The chief constable is, therefore, encouraged to collaborate not only with local government but also with the central government, which can establish uniform standards and conditions of service and with-

hold its share of the finances if these are not met. Chief constables are appointed by committees representative of the relevant local government councils, but their nomination requires the approval of the home secretary. This will be withheld unless the nominee has a period of recent service in some other police force. This, and similar provisions, indicate the growth and spread of national standards in police work.

New Scotland Yard

The British police, through the Home Office, operate certain common services that lead to better cooperation between forces. These include recruit training, advanced training at a police college, research and planning, forensic-science laboratories, wireless depots, central and regional criminal records offices, and regional crime squads. Local forces can solicit assistance from New Scotland Yard in the investigation of murder or from the City of London and the Metropolitan Police in the investigation of serious company frauds. New Scotland Yard is the headquarters of the Metropolitan Police; and, thus, unlike the Federal Bureau of Investigation (FBI) in the United States, it is not a national police agency.

British police experience is of general interest with respect to the question of whether policemen should be permitted to join trade unions. At the end of World War I, there was considerable unrest because rates of police pay had been allowed to fall to a level at which it was difficult for a man to feed his family. Some London and Liverpool policemen went on strike. Pay scales were improved, and the 1919 Police Act was passed declaring, in effect, that no police officer had the right to strike, and that those who did would be dismissed and subject to imprisonment. Every constable, sergeant, and inspector became automatically a member of the Police Federation, which is a representative body with many of the characteristics of a trade union. Acting either locally through branch boards or centrally through the police council or its own conferences, the federation can make representations to a chief officer or the home secretary on matters affecting the welfare and efficiency of members of police forces generally, but not on matters of discipline and promotion affecting individuals. Higher ranks are represented by the Superintendents' Association and the Chief Officers' Association.

The regular police are complemented by special constables (volunteers who perform short periods of police duty in spare time from their normal occupations), a First Police Reserve composed of former or pensioned policemen, and traffic wardens (civilians who work under the direction of the police in the enforcement of parking laws and the direction of traffic). It is important to note that the uniformed, detective, and other branches of the police are not separate career lines. Men are transferred from one to the other, and an officer who is ambitious for promotion will seek to broaden his experience.

British police concept

At the heart of the British conception of uniformed police work is the idea of the constable patrolling his beat. He is regarded as the main instrument of crime prevention in a police service that, since 1829, has believed that "the primary object of an efficient Police is the prevention of crime." Most police officers are employed in the uniformed branch. Up to 10 percent of any force will be in the Criminal Investigation Department, or detective branch, whose members work usually in plain clothes and undertake the investigation and classification of crime reports, the collection of information relating to crime or suspected crime, and the preparation of crime statistics. The constable on patrol carries a short and unobtrusive truncheon. British policemen do not carry firearms, though on special occasions when there is reason to fear that the police may be fired at, trained officers are issued weapons for very limited periods. Firearms control in Britain is strict, and it is difficult for the ordinary citizen to purchase arms.

The United States. The United States provides a picture of what may be the most decentralized police system in the world, accompanied by an extraordinary degree of duplication and conflicting jurisdiction. Since there is no federal government control to ensure uniform standards, every community is entitled to run its own police department but cannot exclude federal or state officials from investigating offenses over which they have jurisdiction. There are five major types of police agency: (1) the federal system, consisting of police officers attached to the Department of Justice (the Federal Bureau of Investigation, together with 700 United States marshals), the Bureau of Internal Revenue (investigating tax evasion), the Secret Service Division, the Narcotics Bureau, the Post Office, and many others; (2) police forces and criminal investigation agencies established by each of the 50 states of the union; (3) sheriffs and deputy sheriffs in over 3,000 counties, plus a few county police forces that either duplicate their police jurisdictions or displace them; (4) the police forces of about 1,000 cities and over 20,000 townships and New England towns; and (5) the police of about 15,000 villages, boroughs, and incorporated towns. To this list must be added some special categories, such as the police of the District of Columbia; various forces attached to authorities governing bridges, tunnels, parks, and parkways; the university, or "campus," police forces; and some units for policing special districts formed for fire protection, soil conservation, and other diverse purposes. The usual estimate is that there are about 40,000 separate, public police agencies in the United States and that the majority of them consist of units of one, two, or three men who are not employed full time. Many are not paid wages but are compensated by the payment of fees for duties undertaken.

These police forces are controlled in different ways. The federal bodies are responsible to the federal authorities, and the state forces to the governors of their states. At local levels there are variations. States are divided into counties, and every county has its sheriff, usually elected for a term of two to four years and with responsibilities for the prosecution of crime and the supervision of the custody of prisoners. He may have deputies performing duties as uniformed policemen. In some states there are county police forces under the control of county boards, whereas, elsewhere, appointments are made by state governors on the recommendation of the county authorities. In some places county police forces are completely independent of the sheriff's office.

Police commissioners

In a major American city, like New York City, the administration, government, disposition, and discipline of the police department are in the hands of a police commissioner, who is appointed by the mayor for a fixed term of years and is likely to be someone with considerable career experience as a policeman. There are deputy commissioners with responsibility for legal matters, policy, public relations, administration, licensing, and juvenile aid. The police department budget is part of the city budget, for which annual appropriations are made out of general tax returns. Whether or not other cities utilize this sort of structure for governing the police depends, in part, upon whether executive responsibility is held by a mayor, a city manager, or a council. Even if a mayor holds executive responsibility, the chief of police may, as in Atlanta, Georgia, be responsible to the police committee of the board of aldermen. In Atlanta, when a disciplinary charge is made against an officer, the chief suspends him; the charge is heard at the next meeting of the police committee, which decides both the verdict and the sentence. Elsewhere, the administrative structure has been supplemented by special citizen review boards charged with investigating citizen accusations against the police. These boards have achieved only limited success and are disliked by the police, who point out that lawyers and doctors do not have similar bodies and that, if a colleague misbehaves, he should be disciplined by a body of fellow professionals, who are the best placed to assess the factors involved. This argument is rejected by those who maintain that American police have failed to discipline those of their number who behave unlawfully. In New York City a proposal for the creation of a civilian review board was rejected in the 1966 citizen referendum.

Many American cities have failed to pay salaries that would attract to police employment the sort of men capable eventually of taking charge of large organizations. In such cities there have been few policemen with manageri-

al talent. Morale has often suffered because serving officers have supported political groups in the hope of promotion if the group is successful at an election. The Detroit department had thirteen police commissioners in the period 1941–71, and there are many other cities in which the turnover has been equally rapid. Some cities appoint as commissioner of police someone who is not a serving policeman but a part-time administrator, who can supervise the managerial side and relate police operations to the political structure. He can then be replaced after an election without upsetting the pattern of relations within the police force. Recent studies have shown that enforcement patterns vary widely when the police are dependent upon the local government. For example, business interests may not want parking regulations enforced strictly, and in the desire to obtain local support, the police may be willing to adapt their activities to community expectations.

American police officers have no national police union or federation to serve as a watchdog over such matters as wages, hours, working conditions, and grievances. Many belong to the Fraternal Order of Police and similar social organizations. In about 45 of the major cities, bodies resembling unions have been formed that appeal to public opinion and use persuasion and arbitration to advance their members' interests. There has been no formal police strike since the one in Boston in 1919, although, as in some Canadian and Scandinavian cities, there have been occasions when large numbers of men have simultaneously reported "sick" as an informal act of protest.

The Tenth Amendment of the U.S. Constitution reserves to the states and to the people those powers not delegated to the federal government. Consequently, state, county, city, and township governments can enact laws and employ policemen to enforce them, provided, of course, that there is no conflict with the Fourteenth Amendment guaranteeing citizens protection by due process of law and equal protection of the law. For this reason, citizens' conduct is regulated by the enactments of different bodies. If a civil rights worker is murdered in a Southern state and the local police are unwilling to investigate the matter properly, the FBI may be sent to investigate whether anyone attempted unlawfully to deprive the dead man of his civil rights. This is a less serious charge than murder, but it is a federal offense and gives federal agents rights of jurisdiction. Similarly, if a European visiting the University of California were one weekend to go to Nevada and bring back with him a girl whom he took to his office and murdered, he could find the case being investigated by the university police, the city police, the local county sheriff, the FBI, the immigration authorities, and, if the girl were a drug user, the Narcotics Bureau. In such circumstances, the different investigating bodies often compete with one another instead of cooperating.

The existing American police structure to some extent reflects public opposition to any concentration of police power. It has been argued that the nation would suffer should all offenses become federal offenses and should power be transferred to Washington and local government be enfeebled. Local problems require local remedies, and the weaknesses of big government are all too apparent. On the other hand, it is argued that the integration and consolidation of police forces could result in an improved service that would reduce costs, increase efficiency, and eliminate the evils of overlapping jurisdictions. The first steps in this direction have been taken in respect to training. Since 1960 no police officer in the state of New York has been able to assume his duties until he has received a certificate for successfully completing basic training. California has introduced grants-in-aid for training that are so attractive that most municipalities are taking advantage of them. Other states have adopted similar schemes. More recently the federal government has made available to the police, through the Law Enforcement Assistance Administration, substantial funds for improving police resources and their management.

Certain other countries with a federal political structure have federal police forces complementing state forces, but the principles involved are no different from those exemplified by the United States. In Australia, for example, each of the six states has its own laws but does not legislate in matters pertaining to federal organizations and cannot pass laws at variance with those of the commonwealth. Each state has its police force. The federal attorney general controls the Commonwealth Police Force responsible for the enforcement of federal legislation. Other police forces under federal control are responsible for overseas territories, the northern territory of the subcontinent, and the capital territory.

The U.S.S.R. In many countries, recent changes in the structure of police administration have been in the direction of greater centralization; so it is of interest to consider a country where the change has been in the reverse direction. After the death of Joseph Stalin the police force of the Ministry of State Security of the Soviet Union was absorbed into the force of the Ministry of Internal Affairs (MVD) under Beria, who was arrested a year later. This force was then divided. The organization dealing with state security was transferred to a Committee of State Security under the Council of Ministers of the U.S.S.R.; known by the initials KGB, it is understood to be primarily a counterintelligence agency. The body that carries out duties performed elsewhere by the normal uniformed police is the militia. Prior to 1956 it was centralized under the Ministry of the Interior, but in December of that year the administration was decentralized. Units are now subject to control by the higher organizations of the militia, by the soviet of the local administrative area, by the local representative of the procurator general of the U.S.S.R., and by the courts. In 1960 the MVD was abolished, and in some areas the militia was itself replaced by Public Order Detachments made up of workers, builders, farmers, teachers, and students. Two and a half million citizens were enrolled into these detachments as voluntary militia. Two years later these new detachments were being criticized in the Soviet press for using their powers in too high-handed a way. To the outsider it would seem as if the Soviet Union found that a highly centralized police service was not accountable to the public and that better responsiveness required drastic decentralization, even to the extent of occasionally relying upon citizen volunteers.

POLICE WORK AND LAW ENFORCEMENT

Routine police activities. Earlier reference was made to the way in which police activity is adapted to the kind of society that is to be policed. There are features common to the work of policemen in differing societies that are founded upon similar technology. Yet within the same society, at times within the same police force, there are variations in the policies of police chiefs. One police administrator, because of his personal beliefs or because of his perception of public opinion, may allocate more men for the investigation of vice than of fraud, may allocate more or fewer men for traffic control or crime prevention. A police chief may instruct his men not to bother about prostitution, provided it is confined to a given locality. Thus, the police in different neighbourhoods may follow different patterns of law enforcement. One study of six American cities identified three such patterns, or styles: a *watchman* style found in departments that use the law more as a means of keeping the peace than for giving effect to the intentions of the legislators; a *legalistic* style, found in a department that acts as if there were a single standard of community conduct, so that all laws are to be enforced, and each policeman may be expected to produce a quota of charges; a *service* style, which is an adaptation to the wants of homogeneous middle class suburban communities, so that, in dealing with minor offenses by local people, the police prefer to activate informal controls rather than invoke the law. These three styles are to some extent associated with city size, population composition, community power structure, and so forth; but they also depend upon the personal decisions of the police chief.

Watchman, legalistic, and service styles of law enforcement

Within the framework of enforcement policy, police work is divided into various branches. The largest num-

ber of men is usually allocated to uniformed patrol, either on foot or with motor transport. Studies of the activities of policemen on patrol indicate that only a small portion of their time goes to the making of arrests or to initiating formal actions under the criminal law. A study of the calls received by a police station in suburban Boston revealed that over half did not involve matters of a police nature, many being simply requests for information. A study in New York City found that less than one recorded incident in 40 led to arrest. Nevertheless, the patrol officer is performing a valuable crime-prevention function by dealing with potentially troublesome situations that, if unattended, could lead to something more serious. Domestic disturbances often culminate in assault, sometimes in homicide or a broken home that can foster juvenile delinquency. The man who is drunk and incapable of looking after himself may be robbed. The prostitute may engage or cooperate in the robbery of her patrons. But, though policemen insist that the heart of police work is preventive patrolling, this activity is not so easily adapted to a motorized age. Policemen in cars are removed from most ordinary contacts with members of the public, and in a big city it becomes increasingly difficult for the policeman to keep a finger on the pulse of his neighbourhood.

"Unit beat policing"

An interesting attempt to come to terms with the problems of policing industrial cities is a scheme of "unit beat policing," first introduced in England in 1966. To each patrol area there is allocated a team that provides for each of three daily shifts—two area constables, one small, conspicuously marked police car and driver, a detective, and a collator. The applications of unit beat policing may vary in accordance with local circumstances and it still has an experimental quality, but a general characterization is possible. The area constable, who is expected to reside within or near his beat, normally works in uniform and is allowed to work out his own plans for covering his area, though he may be expected to spend some of his time patrolling shopping precincts or other areas where the public expects to see a policeman. He is required to maintain personal contact with members of the public, like the old-style village "bobby"; and he should be the "eyes and ears" of the other members of the team. He is the local contact man in any investigation of complaints about the police coming from people in the neighbourhood. The area constable maintains radio contact with these other officers, with his sergeant, and with headquarters. The driver is regarded as another patrolman who has the use of an automobile to take him from one part of the area to another and can therefore attend incidents more quickly than the men on foot. He is not expected to spend his time driving about, and, in one force at least, the optimum mileage for a six-hour tour of duty is regarded as 25 miles (40 kilometres). The collator is a sergeant at headquarters who, with assistants, records all the information sent in by the patrol teams so that he has a daily record of all information relevant to crime, an index of criminals or suspects living or working in the area, an index of such persons' motor vehicles, with additional records for persons subject to driving disqualifications or convicted of drug offenses; there are further records of premises requiring special watch, of the *modus operandi*, or methods, employed in local crimes and of methods used by local criminals, together with a monthly analysis of the pattern of crime in the area. The collator prepares a daily bulletin on local happenings for the benefit of the area team. Because the success of unit beat policing depends in part upon the availability of up-to-date information of this kind, the effectiveness of the area constable is judged by the quality and quantity of information he sends to the collator. The scheme is also dependent upon the use of new methods to establish good communication between members of the patrol team; special conferences attended by the local detectives may be held at regular intervals to bring together as many as possible of the police officers concerned with the patrol area.

The work of police detectives and their new techniques of investigation (such as scene-of-crime fingerprint identification) is relatively well known. It is useful to mention, however, that police forces in many countries have, in recent years, been devoting more and more resources to new departures; *e.g.*, crime-prevention campaigns. Businessmen and householders are advised about weak points in their security and remedies suggested.

Police speakers visit meetings and conferences to educate the public in what needs to be done to combat crime. Police visit schools and talk to students about what the law requires; they may stage mock crimes and trials to give their teaching more impact. In many cities policemen on special duty keep in touch with juveniles considered risks with respect to delinquency, and many officers help with youth clubs in their spare time. If the police are to be active in preventive work, they need men additional to those employed on other duties. This raises the question of what is a reasonable police strength for a given population. Comparisons are difficult, but in most industrial countries there is a police officer for every 600 of population (United States 1:591; United Kingdom 1:571; Japan 1:653). In nonindustrial societies the ratio is higher. Countries like Belgium, France, and Spain have a low ratio if the gendarmerie is included in the calculation; otherwise they are within the normal distribution.

Police and courts. One of the most important ways in which the police are held accountable for the manner in which they perform their duties is through the courts. In France (and in countries with similar juridical procedures) the police making inquiries are under the direction of the investigating magistrate, whose task it is to decide whether there is a case for trial and, if so, what it is. The French think it improper that policemen should inquire into a citizen's activities without his knowledge that they are doing so and without these investigations being under judicial control. The English and American procedure is very different. Anyone may prosecute. The judge listens to the prosecution's argument and evidence that the accused has broken a particular law, and he listens to the defense. Then he decides between them. If he finds the case proved, he then inquires into the accused's previous record and determines the sentence. Should it transpire that the defendant has broken a law but not the one he is accused of having broken, the case must be dismissed.

Police discretion in arrest and prosecution

If the accusatorial system is to function justly, the police must bring all cases of lawbreaking before the courts. It is for the court to decide whether they merit punishment. But, in practice, the police exercise considerable discretion as to whom they will prosecute; and, although their use of this discretion is not at present very well controlled by any agency, the principle of police discretion can be defended on several grounds. In several American cities, for instance, police chiefs who have instructed their men to bear down on offenses and bring offenders to court have had to revert to previous, less stringent practices, because the rate of reported offenses had increased so much that the courts had been unable to cope with the pressure of cases. Apart from such considerations, there are three chief arguments in favour of police discretion. First, it has not yet been possible to draft and keep up-to-date a criminal code that unambiguously encompasses all conduct intended to be made criminal and none other; there will always be technical offenses or offenses that public opinion no longer regards as culpable. Second, those charged with enforcing the law do not have sufficient resources to enforce all the laws all the time, so that enforcement must be selective. Third, to bring a technical offender to court may, in practice, damage a man's reputation.

The second of these arguments was well illustrated in 1966–68 in London, where the law governing gaming houses held that gambling games must be so conducted that the chances would be equally favourable to all players. Casino operators attempted to find ways around this law by such means as using roulette wheels with a zero to increase slightly the odds in the favour of the house. Offenders were convicted but appealed. Believing the law to be uncertain and believing that the expense involved in providing detectives to watch such clubs was not justified

unless there were complaints of cheating or reason to believe that a particular club had become a haunt of criminals, the police commissioner issued instructions in which he said:

> For the time being all applications for authority for an inside observation in licensed or registered clubs for the purpose of detecting gaming are to be submitted for my approval.

A private citizen went to the courts to complain that the law was not being enforced. The court of appeal recognized that police chiefs have a discretion with which the law will not interfere, covering such matters as the disposition of their resources and the decision on whether to prosecute in particular cases. But this instance did not fall within that range. It revealed what one judge called a deplorable state of affairs. Had the commissioner not already withdrawn his instruction, the court very likely would have ruled in favour of the complainant.

There are circumstances in which the police overenforce the law, as when they prefer a "holding charge" to detain someone against whom they are preparing more serious charges. More usually, they underenforce the law by failing to prosecute when they believe that the ends of justice would not be served by so doing. For example, they are usually less ready to prosecute for drunkenness on Christmas Eve. Also, policemen in many countries have decided that if a wife complains on a Saturday night that her husband came home drunk and assaulted her, it is unwise to arrest the husband, lest the wife, who has to live with him when the trouble is over, later asks for the charge to be withdrawn. If the case does go to trial and the man is fined or imprisoned, as much of the burden falls upon the wife and children as upon the offender. In such circumstances, legal sanctions are often clumsy weapons.

Rights of the accused

The courts control police activities in other ways. In England a set of publicized "Judges' Rules" outlines safeguards for accused persons while under investigation. If, in court, it can be proved, for example, that the police failed at the proper points to warn a man that he was under suspicion of having committed a particular offense and that any statement he might make could be given in evidence, then the prosecution might fail. In the United States, the Supreme Court has held that an accused need not answer police questions and that he has a right to consult a lawyer before questioning; if he cannot afford to hire a lawyer, one must be provided. In other decisions, policemen have been prevented from submitting at a trial any evidence that was obtained by unreasonable search and seizure, and, until abolished by Congress, another rule held that, if a person arrested by a federal officer was not taken promptly before a judicial officer and charged with a criminal offense, no confession obtained during a period of delay could be used against him. American police officers have maintained that such rulings hamper them in preparing cases against criminals, and it has been pointed out that one danger in such a situation may be that the police will seek ways around the law.

British and American criminal procedures favour public trials so that all evidence can be challenged. Sometimes offenders are punished more severely by the publicity attached to a trial than by the sentence of the court. According to one point of view, this is, in general, beneficial, since concern for reputation is a powerful sanction on deviation, and it follows that, if offenders receive private trials, they will be less deterred from illegal behaviour. According to the other point of view, publicity may for some people be double punishment and a denial of equal justice. It is worth noting, therefore, the procedure known in Norway as the *forelegg* and in Sweden as the *polisförhör*. This is employed only in connection with minor offenses, such as might be dealt with by financial penalties. On the basis of the police evidence, the public prosecutor notifies the accused that, if the case goes to trial, he will ask the court to impose a certain fine. If the accused pleads guilty and pays the fine, the case need not go to court at all. A similar procedure is adopted elsewhere for on-the-spot fines for motoring offenses. It reduces the burden upon the courts; but, on the other hand, this procedure introduces a new and difficult element into the control of the discretion to prosecute by policemen or by public officials.

Crowds and riots. The first problem in dealing with large masses of people is that of mobilizing police strength. A distinction is therefore to be drawn between the planned assembly for which the police have been able to make preparations and the unanticipated gathering, especially the riot, when the police find themselves unprepared and under strength. The London police have developed great skill with crowds but have no experience of urban rioting on the scale with which American police were confronted in the 1960s. Both crowds and riots, however, require patrol officers to adopt a very different approach from that to which they are accustomed. Normally, policemen are free to handle members of the public in an individual manner. But the control of civil disturbances requires large numbers of disciplined personnel working as a team under unified command. They must not arrest offenders unless it is essential, for, if they do, they will have to guard the arrestee or take him away and cannot then be available for controlling the crowd.

The "new police" were introduced in London in 1829 because the army had been ineffective in controlling mobs. From the beginning, they were careful not to overreact when handling crowds. In Britain, if the police fail to control a riot, recompense for the damage has to be paid by the police authority for the district. Were the police to use, say, water cannon, an injured onlooker could sue the police for assault. The London police use no riot squads, no protective clothing, no arms, no riot gas, and no barriers for crowd control. Constables are drawn from ordinary duties when there is a big parade or demonstration. The police line has always to be in close contact with the crowd, preventing the stone throwing and violence that occurs when there is a gap between them and the citizenry. The advantages of the passive police stance encouraged by their legal accountability were exemplified by four London demonstrations in 1967–68 against United States policy in Vietnam. At the fourth of these, in an atmosphere made tense by press apprehensions of revolution, some 50,000 people marched through the city. In February 1971 a demonstration of some 140,000 individual workers passed off without one arrest. Police efforts have been directed to preserving the citizen's right to demonstrate peacefully while isolating any trouble-making minority and preventing their spoiling the occasion.

Disturbances that erupted in more than 50 American cities in 1967 posed serious problems for the police. Most of these began during the evening watch, when a police department normally has about 13 percent of its uniformed force on duty. At this time, a city like Cincinnati, with a population of about 500,000 and an area of 77 square miles (199 square kilometres), would have about 95 uniformed policemen on duty. The National Advisory Commission on Civil Disorders found that many of the disturbances began when allegedly abusive or discriminatory police actions set sparks among the grievances of Negro ghetto dwellers. The police responded to the disorders cautiously, trying to disperse crowds, observing developments or cordoning off the area. In more than half the cases studied by the commission, rumours significantly aggravated the tensions. Some ghetto dwellers seized the opportunity to raid stores believed to charge high prices. In 13 cases the police were unable to control the resulting situation and the National Guard, composed of private citizens who undertake military duty for at least two weeks during the year, was summoned. The commission concluded that they were inadequately trained for riot duties and that their performance in 1967 raised doubts about their capabilities for such assignments. The rumours running through Negro communities were matched by rumours among the whites of widespread sniper fire. In Newark, N.J., National Guardsmen and police fired 13,326 rounds of ammunition in three days. By contrast, regular troops restored order in Detroit by calm discipline and little use of firearms.

The experience of the 1967 disorders suggests that giv-

ing the police more destructive weapons is not sufficient to maintain order when a substantial section of the population is disaffected. Repressive tactics cannot deal effectively with roving bands of rioters engaged in window breaking, looting, and fire bombing. Nor do policemen think it fair that they should have to shoulder the consequences of society's failure to solve long-standing problems.

Police and private security. Recent years have seen a rapid growth in private security companies, because the police in most countries are not able to provide as much protection as is often desired for private industrial premises or the transport of valuable goods or money. Coupled with this development, and in part responsible for it, has been the increase in the opportunity for theft within industrial concerns. Some firms now assume that at least 2 percent of the goods they handle will be stolen at some point by their own personnel or their customers. The security companies themselves provide opportunities for crime, and the police are concerned that these companies sometimes in good faith employ men whom the police know to have been involved in crime. In Sweden private security companies may apply for official registration, in which case, personnel appointed to sensitive positions have first to be given police approval. It is likely that the question of delimiting the responsibilities of police and private security bodies will attract increasing attention, not least in view of the activities of organizations that inquire into the credit worthiness of private citizens. In Britain, security men carry weapons that are not available to the police and wear uniforms of a kind that sometimes leads people to mistake them for police officers. Some managers discharge employees who have been stealing from the company instead of reporting them for prosecution. But this may be unfair to other employers who may subsequently offer them employment without knowing their background. It is also contrary to the public interest that crime should go unpunished. How private bodies can be made to attend to the public interest is of growing concern to police officers.

Public relations. Since the end of World War II, police forces in many industrial countries have feared that their relations with the public have been deteriorating and that this was having an adverse effect upon their work. Many policemen see the public as having become more ready to question authority of any kind. They believe that the increase in motor traffic has brought citizens who would otherwise be well disposed toward the police into relationships that easily generate friction and resentment. Accordingly, the police have launched public relations campaigns, appointed press officers, and devised new procedures for investigating public complaints. The changes in public attitudes, however, are far more complex than the foregoing would suggest; many individuals will express favourable sentiments toward the police in one situation and hostile ones in another. There may be grounds for believing that the key to this problem lies not in any generalized public attitude or in formal controls upon police conduct, but in the informal relations with the public that stem from the way in which the policeman's role in society is defined.

The policeman in society

Because of the special powers vested in him, a policeman is bound to be set apart, to some extent, from other members of society. In many countries, policemen, in theory at least, are never off duty. They are expected to take police action whenever necessary, even if they are out with their families seeking recreation. Police regulations usually imply that a policeman when out of uniform must do nothing to bring the force into disrepute and may stipulate that officers may have no connection with people licensed to sell alcohol, provide entertainment, run gaming houses, and so forth. Policemen dislike unnecessary restrictions upon their social lives and argue that they can do their jobs better if they can be on easy terms with members of the public, so that the latter see them simply as men with duties to perform like anyone else. It is probable that the more a policeman is hindered from participating in the community, the less he will understand public sentiment, the less he will exercise his discretion, and the weaker will be the informal controls over his conduct. If this is true, then members of the public may be more likely to be irritated by his behaviour and to treat him differently in social contacts—thus tending to isolate him even more. Hostility and isolation feed upon one another. This proposition can be illustrated from colonial territories in Africa, where policemen were at times required to enforce laws not meeting the moral approbation of the population, were isolated from the people, and exploited their powers. In some of the northern cities in the United States, police departments have been demoralized by political control, poor leadership, and low rates of pay; the men have come to feel themselves pariahs. The dangers of social isolation are now more widely recognized.

Police work as an occupation. As has been seen (for example, in the reference to the police in Italy), some police bodies are constituted as part of the armed services. In some European countries, it has been normal to recruit as policemen men who have served as regular soldiers or sailors. It has been even more common to appoint retired army, navy, or air force officers to command police forces. Since World War II the trend has been in the contrary direction.

Police work is seen as a specialist career, and it is argued that the senior positions should go to men who have learned the work by practice. This poses certain problems of recruitment. It is argued that the police should be able to offer appointments as cadets to young school leavers who are keen to become policemen. If they have to wait until they are 21 years or so before they are eligible for appointment, they may be lost to the police. On the other hand, it is said that a youth who has come straight into the police from school has had too little experience of life and is likely to be too narrow in his outlook. Cadet-training schemes therefore seek to give such entrants a wider experience by sending them to social-service agencies and other bodies.

A similar problem is that of officer entry. Many young entrants to the police have the physique and outlook to do the work but have not been among the able scholars and are unlikely to make good administrators. It is not clear whether police should recruit university graduates with the promise that they will be appointed to supervisory positions. The French recruit graduates to the Police Judiciaire by examination, and in Norway only men with legal qualifications are eligible for appointment as police chiefs. In the United States many universities offer courses in police science, and those who obtain qualifications in this field have a certain but limited career advantage. In Britain, graduates are welcomed, but have to compete with other entrants for promotion. The most able among the younger officers are selected for a 12-month course at the national Police College at Bramshill; they divide their time so that half is devoted to professional and half to general studies, by contrast with the rather shorter courses at the Federal Bureau of Investigation Academy at Quantico, Virginia, where all the teaching concentrates upon police subjects.

Police work as a profession

There is a strong movement in the United States to define police work as a profession. There are respects in which policemen exercise a professional skill and provide a service comparable to that furnished by doctors and lawyers. On the other hand, a disciplined body of men incorporated in a strongly hierarchical organization can never have clients of their own or the same kind of obligation toward clients that doctors and lawyers have. The model of a profession does not appear really to fit the police, and, when police chiefs advocate professionalism, they are really using this expression as a slogan in the struggle to improve police standards and win more recognition from the public. It is easier to see the policemen's job as the practice of a craft, for the policeman acquires his skills in the course of an experience that is limited to him and those fellows of his who are bound to the same occupation. Conscious of having a special part to play, members of a craft see the occupation as one that aspirants can enter only through a selective apprenticeship. Unlike professionals, policemen are subject to close con-

trols by superiors, they have no serious professional society, and they do not yet contribute new knowledge about their craft in systematic written form.

INTERNATIONAL ORGANIZATION

The International Criminal Police Organization (Interpol) has a history going back to 1914, though its present constitution dates from 1956. Each member nation has established a bureau that maintains relations with the General Secretariat in Paris. The bureaus transmit criminal information that may be of interest to other countries; they undertake, within their own countries, inquiries, searches, and arrests requested by other countries; and they take steps to implement resolutions voted by the annual assembly. Interpol can act only within the framework of national laws; criminals can be returned only if an extradition treaty is in force and the offender is a national of the country requesting return. The International Association of Chiefs of Police, with headquarters in Washington, draws its members largely from the United States and is the leading voice in America for professional police standards. It is active in training, research, and public relations. There is a parallel body bringing together the representatives of the lower ranks, known as the International Conference of Police Associations. The International Police Association was founded in 1950 as a social organization. Its membership of over 40,000 is strongest in Europe. The Association grants scholarships for study travel and arranges annual conferences.

BIBLIOGRAPHY. JAMES CRAMER, *The World's Police* (1964), provides a remarkably comprehensive survey of modes of police administration. Regrettably there is no reference book listing current changes in this field, and the reader will need to refer to national reports prepared in the countries concerned. THE INTERNATIONAL POLICE ASSOCIATION has published a useful *International Bibliography of Selected Police Literature*, 2nd ed. (1968). There are histories of the police forces of many countries: T.A. CRITCHLEY, *A History of the Police in England and Wales* (1967); M. LECLERE, *Histoire de la Police*, 3rd ed. (1964); JAMES RICHARDSON, *A History of the New York Police from Colonial Times to 1901* (1970); PAUL RIEGE, *Kleine Polizei-Geschichte*, 3rd ed. (1966); PHILIP JOHN STEAD, *The Police of Paris* (1957). For popular accounts of police work and the problems of accountability two recent British books deserve mention: BEN WHITAKER, *The Police* (1964); and PETER LAURIE, *Scotland Yard* (1970). MICHAEL BANTON, *The Policeman in the Community* (1964), is a sociological comparison of the police role in Scotland and the United States. JAMES Q. WILSON, *Varieties of Police Behavior* (1968), compares the enforcement policies of eight American police departments, while JOHN A. GARDINER, *Traffic and the Police* (1969), reports on a more restricted comparative study. JEROME H. SKOLNICK, *Justice Without Trial* (1966), discusses the detective's role in California and is sensitive to the legal as well as the sociological issues. ARTHUR NIEDERHOFFER, *Behind the Shield* (1967), gives a sociological viewpoint from a former New York policeman. The current problems of police operations and accountability are also featured in the U.S. President's Commission on Law Enforcement and Administration reports, *The Challenge of Crime in a Free Society* (1967), *Task Force Report: The Police* (1967); and the *Report of the National Advisory Commission on Civil Disorders* (1968). Comparable studies for countries other than Britain and the United States are scarce, but reference may be made to DAVID H. BAYLEY, *The Police and Political Development in India* (1969); and D. CHAPPELL and P.R. WILSON, *The Police and the Public in Australia and New Zealand* (1969).

(M.P.B.)

Police Technology

Police technology covers the methods and techniques of, and the equipment available to, the law-enforcement agencies of the world. Through the 20th century, improvements and innovations have been made in police transportation, communications, weaponry, laboratory facilities, and other areas; but these have scarcely more than kept pace with the pressures generated by denser urban populations, greater sophistication on the part of criminals, social disturbances, and new types of crime made possible by new economic or technological developments—*e.g.*, aircraft hijacking.

Police technology increasingly involves borrowings from the physical and social sciences, as well as from the disciplines of jurisprudence and education. Professional publications, educational programs, workshops, and international conferences of police reveal the growing importance attached in recent years to the technical side of police work. Tomorrow's policeman will unquestionably be impressively equipped and educated; yet it was only in the 1960s that police technology became recognized as a distinct discipline in its own right. The present article is concerned with police technology in surveillance and crime fighting; traffic-control technology is described in the article TRAFFIC CONTROL. The technology of prison systems is described in PRISONS AND PENOLOGY.

HISTORY

Many examples of an incipient police technology can be found in the ancient and medieval world.

Long before the birth of Christ, Egyptians used detailed word descriptions of individuals, a concept known today as "portrait parle." The Babylonians pressed fingerprints into clay to identify the author of cuneiform writings and to protect against forgery. Fingerprints were also used by the Chinese at an early date for purposes of identification. The concept of comparison and identification of handwriting was established in Roman times, as is evidenced by the Justinian Code of the Eastern Roman Empire, enacted in AD 539. Branding and mutilation were used from time immemorial to identify convicted criminals. Trial by ordeal and by combat were medieval methods of interrogation of suspects, with the survivors judged innocent. A more humane medieval method, and a step toward modern concepts, was compurgation, in which the disputants took oaths on sacred relics.

Precursors of the modern lie detector

An Asian interrogation method involved scientific principles based on the lessening of salivation under nervous tension. The mouths of several suspects were filled with dry rice; the suspect exhibiting the greatest difficulty in spitting out the rice was judged guilty. Another precursor of the modern lie detector was employed in India. Suspects were sent into a dark room where a sacred ass was stabled and directed to pull the animal's tail. They were warned that if the ass brayed it was a sign of guilt. The ass's tail had been dusted with black powder; those with a clear conscience pulled the tail, whereas the guilty person did not; an examination of the suspects' hands revealed the person with the guilty conscience.

Modern police methods were beginning to develop in Europe by the 18th century and, in the 19th, were reinforced by a great number of technological advances. When the British Parliament enacted the reforms of Robert Peel in 1829, the importance of keeping good police records was recognized. In principle, Peel stated that police records were indispensable for the proper distribution of police strength. His reforms laid the foundation for the development of the modern police communications system and provided for a coherent police organization.

Until the mid-19th century, however, police signal systems consisted of whistles, raps on the pavement with a nightstick or baton, and flashing of lights. In 1849 the central police station at Scotland Yard and the district police stations of London were connected by a telegraph line. Other cities followed, and telegraph boxes placed at points on patrolmen's beats permitted swift calls to headquarters when assistance was needed.

The telephone was first installed in a police station in Washington, D.C., in 1878. Two years later Chicago took the important step of installing telephone boxes on patrolmen's beats.

In the area of identification, the late 19th century brought important advances. In 1882, Alphonse Bertillon of France introduced a means of identifying suspects by using measurements of the body, head, and limbs. This system was largely supplanted by fingerprinting, which dates from about the same period. The Henry system of classifying fingerprints (by patterns and shapes) was adopted by Scotland Yard in 1901, and this system, or variants of it, soon became standard all over the world.

The idea of tracing and identifying an individual typewriter by means of the peculiarities of its type characters probably first appeared in Arthur Conan Doyle's story "A Case of Identity," written in 1891, three years before the first mention of the subject in technical literature.

First use of the automobile

The automobile, first used in police work in Akron, Ohio, in 1899, became the basis of police transportation in the 20th century. Closely linked with it was the adoption of radio communication. Other communications innovations were the teletypewriter and facsimile. The teletypewriter evolved from printing telegraphs and developed rapidly during World War I; in the 1920s it became the workhorse of police communications in most countries. Facsimile transmission (telephotography) dates from the early 1900s, and, as early as 1908, a criminal was captured as a result of his photo being transmitted by facsimile from Paris to London. After World War I this technique also became widespread.

Perhaps the single most important invention in the development of police communications was the transistor, which in the 1950s provided small, lightweight, reliable radio transmitter-receivers.

In the 1920s the technique of identification of firearms on the basis of striated markings on bullet jackets was introduced. As these and other new techniques became familiar, the first police crime laboratories were instituted in Europe and the U.S. One of the earliest was that established at Lyons, France.

There have been few innovations in weaponry. The nightstick (club) and firearm, invariably an adaptation of a military weapon, remained virtually unchallenged in the police arsenal until the 1960s, when chemicals, tranquillizing darts, light- and sound-energy techniques, and other devices began appearing.

MODERN SURVEILLANCE SYSTEMS

Police have a number of surveillance techniques available, but chief reliance is almost universally placed on patrolling.

Patrol systems. Patrols vary in methodology, assignments, structure, and mobility. Their flexibility has been significantly enhanced by improved methods of communication, especially the two-way radio and by the availability of numerous types of vehicles; the automobile and motorcycle are the most prominent, but the bicycle remains valued in many countries. The mobility provided by modern communications and transportation techniques is particularly useful in dealing with civil disorders.

The areas covered by metropolitan police patrols vary enormously in population density, character of commerce and industry, terrain, and roads. The helicopter, whose value in traffic surveillance was established in the 1960s, promises in the 1970s to be used for general patrol work in metropolitan and rural areas. In some countries the short-takeoff-and-landing aircraft has been adopted for patrol and rescue work.

Police dogs. The training of dogs for police work was originally developed in Ghent, Belgium, about 1900 and was soon copied elsewhere. Though certain breeds with especially keen senses are used for special purposes, such as detecting caches of illegal drugs or for tracking fugitives and missing persons, the most widely trained dog for regular patrol work is the German shepherd, or Alsatian. Other breeds sometimes used include boxers, Doberman pinschers, Airedale terriers, rottweilers, schnauzers, and bloodhounds. Selection and training has become highly specialized. Selection involves use of standard criteria relating to breeding and physical measurements, and training is rigorous.

Electrical protection systems. Electrical protection (alarm) systems substantially reduce burglaries and contribute to capture of many criminals. The development of alarm equipment and the evolution of operating procedures in the U.S., Canada, and western Europe has been significantly influenced by the published standards of Underwriters' Laboratories, Inc. (UL). Central-station systems and local-alarm systems are the two basic categories of UL-regulated burglar-alarm systems.

Central stations

Central stations are commercially operated control centres serving metropolitan and surrounding areas. Burglar-alarm protective circuits and devices automatically transmit signals to these centres. Armed guards are in attendance at all times and are dispatched to conduct immediate investigations of unauthorized entry of protected properties. Central stations generally have direct communication channels to police control centres. Some such centralized operations in large European cities control several completely automatic satellite stations in nearby cities; the local police are dispatched by the central station.

Local burglar-alarm systems are used primarily in small cities and towns where central-station services are not available. Sensitive warning mechanisms are connected to loud alarms attached to outside walls of buildings and, often, to a signal system in local police stations. (A description of proprietary detection and warning systems is given in SECURITY AND PROTECTION SYSTEMS.)

Technical surveillance. The application of technology to surveillance operations has greatly expanded the capabilities of law-enforcement agencies and has led to creation of a special field of police technology known as technical surveillance.

Audial surveillance. Audial surveillance, or electronic eavesdropping, became practical for obtaining evidence and investigative leads as a result of intensive development of the magnetic tape recorder and solid-state electronic components during World War II and subsequent years. The field has two technically and legally distinct areas: the surreptitious interception of oral communications, commonly known as "bugging," and the surreptitious interception of wire communications, commonly known as "wiretapping." Typical devices used include miniature microphones, miniature radio transmitters, and a variety of radio-receiving and voice-recording equipment.

Microphones, of course, are always required for the interception of oral communications and may be connected by cables directly to recorders located at remote locations or may be connected to transmitters. Excellent quality microphones of very small size are available throughout the world. Microphones located in the handsets of telephones may be used to intercept nearby conversations through modifications and the use of companion equipment connected to telephone lines.

Radio transmitters of cigarette-package size, equipped with integral batteries and microphones, are commonly known as "wireless mikes." Most radio transmitters used for audio surveillance operate within the very high frequency (VHF) range.

The most common wire-communication-interception activity consists of monitoring telephonic conversations. Equipment used for telephone tapping ranges from the telephone repairman's handset to automatic voice recorders connected to target telephone lines.

Optical surveillance. This type of surveillance makes use of such devices as binoculars and telescopes; cameras with long-focal-length, or telephoto, lenses; closed-circuit television; and videotape recorders. Closed-circuit television is a versatile surveillance method that has become widely used both by police and by private security systems. From a central location, a supervisor can monitor events at one or several remote locations. Cameras may be fixed or oscillated to sweep a large area.

Devices for low light levels

Recently available low-light-level devices have greatly extended passive optical-surveillance capabilities. Practical low-light-level television systems are manufactured for optical surveillance operations under light intensities as low as 10^{-8} lumens per square metre, or virtual darkness. Night-vision telescopes using two and three light-amplification stages are also in use.

Tagging and tracking. Object tagging and signal tracking comprise a third category of police technical-surveillance operations. Devices emitting a unique signal are attached to vehicles or implanted in target objects. Compatible receiving or activating equipment is used to track the tagged objects. A typical device used in tracking is a miniature radio transmitter using distinctive tone modu-

lation, commonly known as a "bumper beeper," which is surreptitiously attached to a vehicle. The target is discreetly tracked with one or more surveillance vehicles equipped with appropriate receiving equipment.

CRIME REPORTING AND RECORDS

The efficiency and effectiveness of a police department depend in large degree on complete, readily available, and up-to-date record keeping.

Information must be reported, collected, classified, analyzed, processed, and stored. Numerous files must be kept, of which the most important include fingerprints, *modus operandi* of criminals, and various investigation and reference files. There should be as much cross-filing as possible.

Fingerprint classification. Fingerprint records are among the most significant files that are maintained. Inked fingerprints from all ten fingers of an individual can be classified; *i.e.*, put into narrow categories for file purposes and ultimate retrieval (see illustration). In the United States, the Henry extension system is used. This system is a modification of the Henry system, which was developed prior to 1900 in England. These two systems vary greatly in presentation but are similar in content. In parts of South America, the Vucetich system is used. In the Henry extension system, each of the ten individual fingerprints of a person is classified as to arches, loops, whorls, or composite style prints. The type of design, plus subclassification based on locating certain fixed points and counting the ridges lying between the points, is used to make the classification.

By courtesy of the New York State Identification and Intelligence System



Classification of a fingerprint for computer search and storage from a facsimile copy of a suspect's fingerprints.

Fingerprints obtained from individuals are classified and compared with the fingerprints of other individuals having the same classification. With this system, personnel in a police centre need not compare a set of fingerprints with all the millions of cards on file but only with the few in the same classification. (For the technique of fingerprint photography and identification, see below *Laboratory procedures.*)

Modus operandi. If a criminal has been successful in the use of certain methods and techniques, he tends to repeat the same operational procedure. Thus, evidence of a definitely characteristic behaviour and procedure at the crime scene can serve to identify the offender if the pattern is recognized by the police. Although classification of criminals according to operational method and manner had, by midcentury, lost a good deal of its effectiveness in the investigation of burglaries, large law-enforcement agencies still maintained *modus operandi* files to recognize a pattern of behaviour, to associate a group of crimes with a single perpetrator, and to enable them to approximately predict the next target of the criminal. The *modus operandi* file is most effective in crimes involving personal contact, such as felonies against the person, confidence games, and forgery.

Investigation and reference files. Informational files are maintained by law-enforcement agencies. The primary purpose of these records is to identify the source of evidence discovered in connection with a crime. For example, a file of automotive paint samples will assist in identifying the make and model of a vehicle involved in a hit-and-run case. Automotive registrations and records of firearms will assist in establishing ownership. In addition to their own files, the police have access to the records of many governmental agencies that record titles, contracts, vital statistics, significant events, and official acts in general. Among other files maintained by the police the following prove exceptionally useful.

Missing persons. Each year many thousands of persons are reported missing to the police. In addition, the police must investigate cases of unidentified persons, living and dead. Descriptive files are maintained of missing persons under the age of approximately 19, unidentified dead, and unidentified persons who are physically and mentally afflicted.

Laundry and dry cleaners' marks. A valuable clue to the identity of an unknown person can be found in the mark or marked tag placed in his clothing by the laundry, dry cleaner, or tailor. The mark usually consists of a set of numbers or letters identifying the retail store for the wholesale processor. Samples of these marks are filed systematically by larger police departments for the purpose of tracing a lost, dead, or wanted person.

Lost or stolen property. Many cases of burglary and larceny are solved simply by tracing stolen property to the possession of the criminal. For this purpose, police maintain a central cross-indexed file of property lost or stolen and of articles pledged at pawnbrokers or sold to secondhand dealers. The common objects of larceny—automobiles, watches, jewelry, furs, and cameras—are identified by means of an accurate physical description, including serial marks and personal inscriptions.

Computerized records. The record-keeping tasks of police work were long handled by individuals using a diversity of manual systems. Quick nationwide or international utilization of records and files was impossible. Now, however, electronic data processing has revolutionized information-handling techniques. Manual filing systems have been superseded by computers that operate from central geographic points, storing data and making it instantaneously and accurately available to many police constituents. Furthermore, the computer can be linked with other computers in the same country or in different countries.

Electronic data processing

In addition to serving as a depository and dispenser of information, the computer can provide statistical analysis for research and management, develop intelligence data bases and analysis systems, assist in the analysis of physical evidence, facilitate swift document retrieval, and help in such areas as identification of fingerprints. In the analysis of fingerprints, the computer is not yet economically effective because there is no simple means by which single prints can be directly identified, but promising research is underway to develop automatic scanning and classification systems.

COMMUNICATIONS

A total police communication system can be extremely complex, but its complexity varies with the needs associated with particular police organizations and their operational environmental situations. Each police function or situation calls for its own subsystem of communication.

In a large city, the sources of information that must be fed into the central system are remarkably varied. Police may be on foot or motorcycles, in patrol cars, in patrol boats, or in aircraft. In addition, the system must encompass links to fire stations, ambulance services, traffic-control systems, rescue teams, alarm systems, emergency utility services, crime-information systems, and crime

laboratories. A variety of fixed and mobile systems are in use, including radio, television (including closed-circuit), telephone, teletypewriters, and facsimile transmission. Research and development are active in such areas as more effective use of available radio frequencies, improved mobile teletypewriters, improved facsimile systems for fingerprints and documents, and direct vehicle-to-computer communications systems.

Police radio-communications systems, which profited considerably from the advances in electronics during and after World War II, also profited from the development of computers, which made possible the quick retrieval of information on stolen property, wanted persons, and other police intelligence for patrolmen as well as headquarters. The augmented flow of information has placed an added burden on existing communications systems, which have consequently grown more complex. Among numerous technological innovations, the development of the modern police communications centre merits particular mention.

Functions of a police communications centre

A communications centre performs several functions: (1) it receives and processes police calls from the public; (2) it audits and dispatches police resources in response to police calls; (3) it receives and responds to information inquiries as well as calls for assistance from police patrols; (4) it maintains moment-to-moment status information on police resources for command and control purposes.

The geographic distribution of police communication centres has been broadened to include locations in many countries of the world. These centres facilitate police work needed to deal with the international mobility of modern criminals. The extensive radio network maintained through Interpol serves to promote international cooperation among world police organizations.

INTERROGATION METHODS AND EQUIPMENT

Policemen and other investigators continue to depend on interrogation as a principal means of determining facts and resolving issues. Reliance on interrogation, however, involves certain problems: ascertaining when a suspect or witness is telling the truth, evaluating memory, allowing for the physical and mental condition of a witness or suspect, and understanding the problems created by an individual's perspective.

Interrogation methods and equipment have evolved in response to these problem areas. The psychological, psychophysical, and physical sciences have played vital roles in police interrogation techniques.

The polygraph. The most dramatic gains in interrogation technology have come through the polygraph, or so-called lie detector. The polygraph monitors and records selected body changes that are affected by a person's emotional condition. The recorded changes are then studied, analyzed, and correlated in respect to specific questions or other stimuli.

The first modern polygraph was constructed in 1921 by John A. Larson, a medical student at the University of California, working with a member of the local police department. Larson's instrument was capable of continuously recording blood pressure, pulse, and respiration; since it recorded these conditions simultaneously, it was called a polygraph. A later development provided for the recording of the psychogalvanic skin reflex (electrodermal response), the flow of current between two different parts of the body. Research is continuing on both the instrumentation and the psychological techniques necessary for its effective use by the police.

Even though the polygraph has been formally and successfully utilized in police intelligence and security investigation since 1924, there is still no complete agreement on its validity among leading psychologists. Furthermore, the results of a polygraph test are not always judicially acceptable. Despite these qualifications, the polygraph has proved an invaluable aid to police interrogation and the development of investigative leads.

Detection of narcotics and dangerous drugs. Portable field chemical kits have been developed for the detection of narcotics and dangerous drugs. These kits are designed to provide a practical method for initial tests in the field and are not intended to replace the more thorough laboratory examinations.

Drug tests. Tests for determining that a suspect has consumed a narcotic or dangerous drug must be performed in a forensic laboratory. These tests require the taking of a urine or blood sample from the suspect and, therefore, require close medical supervision. Many new techniques for the detection of both caches and small amounts of drugs are under development, ranging from the use of specially trained police dogs to electronic odour-detection equipment.

Alcohol tests. In view of many studies linking traffic fatalities with alcohol use, the determination of the amount of alcohol in a driver's blood has received increasing attention from police researchers in many countries.

For many years, the basic test for legally determining a driver's blood-alcohol content involved taking a blood sample and sending it to a laboratory for analysis. This method, although accurate, is not practical for law-enforcement purposes because many people object to having a blood sample taken, and test results are not immediately available. The development of the breath analyzer for the determination of blood-alcohol content has given the police a more practical tool. This instrument permits a precise amount of the suspect's breath to pass through a chemical solution (potassium dichromate and sulfuric acid), resulting in a colour density change proportional to the amount of alcohol in the air sample. The percentage of alcohol in a given air sample is determined and equated to a certain percentage of blood-alcohol content. In cases where the suspect is not conscious or requires serious medical attention, a blood test may be the only recourse. Blood-test and breath-analyzer techniques, however, are not accepted as evidence in all courts.

The breath analyzer

WEAPONS AND RELATED DEVICES

To contend with the wide diversity of criminal situations, police require a variety of weapons and restraining and protective devices. Science and technology have contributed to the development of increasingly practical and effective equipment of this type.

Standard weapons. Standard equipment varies. The London policeman carries only a small baton; the New York City policeman carries a large-calibre revolver and a baton; the Paris policeman carries a small-calibre semiautomatic pistol, a baton and a leaded cape; and the Cairo (Egypt) policeman carries a submachine gun or automatic rifle.

The most universal physical weapon and certainly the most frequently used weapon over the years has been the baton, nightstick, or billy. The baton has been the subject of study and research, and there now exists a family of batons which includes, in addition to the original wood or plastic baton, devices that contain firearms, aerosols or tear-gas cartridges, and electric-shock systems. Some batons are made to break at a certain degree of impact, in order to avoid serious injury.

Pistols, including the automatic pistol and the revolver, are the second most widely used police weapons. In the United States the revolver (.38 special to .45 calibre) is the primary pistol, whereas in Europe and Asia the automatic (.25 to .38 calibre 9 mm) finds extensive use.

Shotguns, rifles, and automatic weapons are usually included in the police arsenal but are issued only when essential.

Chemical and other nonlethal weapons. The most universally used chemical weapons are those that deliver compounds that cause either severe weeping or tearing at the eyes (lacrimators) or extreme discomfort or sickness (nauseators). The three most universally used chemicals are known by the initials CN and CS, both lacrimators, and DM, a nauseator. Among the systems used to dispense the chemicals are riot shotguns, fog generators, and grenades. The most commonly used new weapon is the handheld aerosol tear-gas projector (MACE). Of convenient size, it can be carried in a holster; it is easily aimed and is instantly effective. Its viable range is about 15 feet and it delivers about 40 one-second bursts.

Lacrimators and nauseators

Other nonlethal weapons include sound and light devices which cause confusion, pain, or temporary blindness.

Restraining and protective devices. Criminal restraining and police protective systems have been subjects of much study and research. The developments in restraining devices or systems include lighter, more effective handcuffs, with improved locking systems; lightweight thumb cuffs; pressure-twisting devices that are applied to wrists; lighter and stronger straitjackets; and lighter and stronger leg irons.

A recently developed restraining device is the disposable handcuff. It consists of a 22-inch-long nylon plastic strap with a slotted head at one end and a tapered tail at the other. A stainless steel barb in the head allows the tail to be pulled through the head in one direction and locks the strap so it cannot be withdrawn. To remove these plastic cuffs, the policeman merely cuts them off. This development is particularly useful in mass arrests, as in riot situations.

Significant developments in the technology of protective devices include heavy plastic shields to ward off rocks and other missiles; lightweight nylon clothing resistant to pistol bullets; bullet-proof shields; and bullet-proof armor made of nylon and ceramic materials, which can withstand high-powered rifle bullets.

LABORATORY PROCEDURES

Crime laboratories may be either specialized, such as those that handle principally documentary evidence, or they may be broad in coverage, as is the case with most national police laboratories. The fundamental purposes include the collection and preservation of physical evidence, its examination, and the preparation and presentation of accurate, objective testimony regarding it.

Personnel include specialists such as document examiners, generalists such as criminologists, and technicians and other supporting personnel. They are usually employed by a government unit and must pass rigid examinations.

Facilities and equipment. Crime laboratories usually are equipped with a wide array of scientific instruments and facilities. Instruments are constantly being developed and introduced, and the resources of individual laboratories vary. Some are equipped with expensive and highly sophisticated analytical instruments, such as the mass spectrograph, the electron microprobe, and the scanning electron microscope. Close cooperation among laboratories makes possible the sharing of specialized equipment when needed.

Value of photography

The largest category of laboratory equipment relates to photography. The usefulness of the medium extends from straightforward photography of, for example, car accidents and sites of crimes to specialized work of various kinds. Colour photography is increasingly used to obtain more realistic and informative representations of the crime scene. Ultraviolet and infrared photography aid in detecting stains, invisible writing, and alterations on documents. Radiography is also used to provide details that would otherwise be invisible. Photography through the microscope (photomicrography) is a common technique, a special case being the examination of bullets. Fingerprint photography is an area of specialization that may involve conventional methods or any of a number of special techniques to recover useful prints from different types of surfaces.

Among the more common facilities are chemical laboratories, with a range of reagents and analytical instruments, including spectrographs and chromatographs. Optical equipment includes a full range of magnifying devices, from magnifying glasses to specialized microscopes of various kinds, including stereomicroscopes and comparison microscopes.

Evidence and its examination. An investigation at the scene of a crime, particularly in crimes of violence, is the first phase of the laboratory function. Some jurisdictions have mobile crime laboratories, in which some of the regular laboratory tests can be performed.

Crime sites must be searched to locate pertinent physical evidence, which may range from latent fingerprints to bloodstains to pieces of a broken automobile headlight. Evidence must be marked for identification, preserved, and protectively packaged for transportation to the laboratory. In addition, records (written or photographic) must be made of each piece of evidence and its exact location with respect to the crime.

Certain laboratory procedures and tests are carried out in the field when time is limited or when there is a possibility that the evidence might be disturbed. Among such procedures or tests are: dusting for fingerprints, making casts of footprints and car tracks, making blood tests, and collecting specimens of organic and inorganic materials from the environment.

Classes of evidence. Evidence can be categorized according to potential value as proof. Certain types of evidence can be individualized or associated with a unique source, whereas other types cannot be pinpointed but can aid in identification because they fall into particular classes. Types of evidence that can be individualized are handwriting, typewriting, fingerprints, footprints, tire marks, impressions or casts of nonstandard items, striated markings on bullets, tool marks, objects with random fractures or tears, and substances that have undergone an alteration that makes them unique among others of the same class. Among evidence that only identifies a certain class are blood and other body fluids, narcotics, toxicological materials, fibres, soils, inks, and many kinds of materials produced in batches.

Examinations can be either unilateral or comparative in nature. Blood-typing and establishing the distance from which a weapon was fired are unilateral examinations, whereas studies of handwriting or spent bullets involve comparisons with other examples.

Pathological examinations. Although pathological examinations are not ordinarily conducted by a crime laboratory, they are part of forensic science. Definitive statements relating to the cause of death and time of death are made by the coroner or medical examiner of the jurisdiction in which the death took place. Observations by crime laboratory personnel of the conditions at a crime scene, plus toxicological analysis by a crime laboratory, are utilized in concluding whether a death was by accident, suicide, or murder. In most jurisdictions, an autopsy is conducted if a physician was not in attendance or if the cause of death is not obvious, as in cases involving poisoning, electrocution, asphyxiation, heart attack, or stroke. In cases involving identification of a decedent, crime laboratory personnel may work with a coroner or medical examiner to establish identity by such methods as comparison of fingerprints and handwriting.

Fingerprints, palm prints, and footprints. Fingerprints found at the scene of a crime can be very significant as evidence connecting an individual with a crime scene. Fingerprints can be either visible or latent. Visible prints —formed by dirt or blood, for example—or three-dimensional prints formed in soft matrices, can be photographed directly. Latent fingerprints, which are not ordinarily visible, can be brought out by dusting techniques when the surface is hard and by chemical techniques when the surface is porous. In dusting for fingerprints, a fine powder of contrasting colour is applied with a fine brush. The powder clings to the residual oils and fats in the print and the excess powder is removed with the brush. On porous surfaces (such as paper) fuming iodine, silver nitrate, or ninhydrin solutions are used to develop the latent fingerprints. The most effective developer of latent fingerprints is ninhydrin, which can reveal prints that are several years old.

Fingerprint "points"

Individual fingerprints are identified on the basis of agreements in a significant number of individualities, commonly known as "points." These are the bifurcations, ending ridges, and dots in the fingerprint pattern. If sufficient points are found with spatial relationship to other points, then a basis exists for identifying a fingerprint. It has been considered necessary to have 12 points to identify a fingerprint, but in current practice, a lesser number can be used under certain circumstances. Palm prints and footprints are identified in the same manner as fingerprints.

Document examination. The work of the "questioned document examiner" encompasses the examination of handwriting, typewriting, ink, paper, and typography and covers such problems as the identification of handwriting and typewriting, the sequence of events involved in the preparation and handling of a document, the alteration of a document, and the age of a document.

Handwriting, like all other evidence, is identified on the basis of individuality. Children are taught specific ways of forming letters, but they rapidly begin to inject their own individuality into letter forms and writing execution. By the time a person reaches maturity, his writing has acquired peculiarities which serve to identify him. Naturally, there is some variation in each person's writing from sample to sample, but any sample containing sufficient evidence of the individuality of its author can be identified.

A forged signature presents other problems. Simulated signatures based upon recollection contain a combination of the forger's own writing habits and his recollection of the victim's habits. In many cases, such simulations can be identified. On the other hand, when the perpetrator makes a careful drawing of the victim's signature or traces an authentic signature, the forgery can be exposed but cannot be identified with the handwriting of the perpetrator. Two different individuals making careful tracings of the same model signature would produce virtually identical drawings.

Typewriter identification

The make and model of typewriter used to prepare a document can be determined by examining typeface designs and comparing them with information contained in reference files and differentiation charts. The work of a specific typewriter can be identified if it contains sufficient identifying individuality, based on broken or tilting characters, badly aligned characters, characters that print more heavily on one side than on another, and rebounding characters. Mechanical conditions affecting the type basket or the escapement may also have identifying value. Since it is impossible to reinsert a document into a typewriter in its exact former position, it is possible, using a ruled screen, to determine if a document has been prepared at one sitting or has been re-inserted.

Ink comparisons provide class evidence that is frequently of value, but it is not possible to individualize ink; *i.e.*, identify it as coming from a unique source. Chemical tests of various kinds are used for ink comparisons.

Papers can be differentiated on the basis of fibre, filler, and sizing constituents. Fibres can be identified by differential staining and microscopic examination. Fillers can be identified by X-ray diffraction because they are crystalline substances. Chemical tests are used for the identification of sizing constituents. Research with neutron activation analysis, a technique involving chemical analysis by radioactivity, indicates that it may even be possible to identify paper by batches.

Questions concerning the sequence of writing—*i.e.*, whether the ink signature is over or under the typewritten portion of a document—can be important in proving or disproving the authenticity of a document. In this and related problems, such as deciphering and restoration of eradicated or erased writing, examinations are carried out using stereomicroscopes, ultraviolet and infrared techniques, or chemical tests, depending on the specific problem.

Ballistics techniques. Firearms examinations involve such problems as ascertaining the type of weapon and the specific weapon used to fire a bullet, determining the distance between the weapon and the target, and establishing whether or not a particular individual fired a weapon.

The identification of a particular weapon is based upon the comparison of minute striations, or striae, left on bullets as they pass through the rifled barrel. Each barrel leaves a different combination of striae on the side of bullets. A bullet recovered from the scene of a crime can be compared with a test bullet fired from a suspect weapon. The striated markings on both bullets are brought into juxtaposition under a comparison microscope. If there is agreement in the width, depth, and spacing of a significant number of striae in both bullets, an identification can be made. (The same principle is used in the identification of tool marks.) The extractor and ejector markings on rifle and automatic-pistol cartridges can be identified in the same manner. In some cases, the firing pins and breechblocks leave individual markings on the face of a cartridge.

The distance from which a weapon was fired can be determined, up to several feet, by examination of the form and size of the hole in a target, the extent of the burned area around the hole, the amount of embedded powder grains, and the presence of mercury or lead from primers. The effects vary with the weapon and the type of ammunition. Such information may be vital in determining whether or not a gunshot wound was self-inflicted. Infrared, X-ray, and chemical tests are used in determining the extent of powder residues.

The best means of determining whether or not an individual has recently fired a gun has been the subject of much controversy. In the dermal nitrate test, melted paraffin is used to pick up traces of nitrates and nitrites on the hand, after which a chemical test is made for such compounds. This method has been discredited because it can give incorrectly positive test results. Use of the neutron-activation technique to detect the presence on the hand of antimony and barium from cartridge primers is much more definitive.

Individualization of paint

Paint. Under certain circumstances, paint evidence can be individualized. If a jagged edge of a paint fleck found on a suspect can be matched with the edges of the matrix from which it came, a basis exists to connect a suspect with a crime scene. Multilayered paint chips can sometimes be individualized if the discrete layers possess sufficient individuality. Under other circumstances, paint must be considered class evidence. Examination of paint for class characteristics involves the use of chemical tests to determine constituency, as well as examination with instruments such as the infrared spectrophotometer, emission spectrograph, or electron microprobe.

Inorganic materials. Soil, safe insulation, and other inorganic evidence can also be compared to determine common origin. It is not possible to individualize the evidence, but the fact that the same type of mineral matter was found on a suspect and at a crime scene is usually considered significant. Inorganic materials can be compared in terms of metallic element constituents, density of constituents, and microscopic structure.

Blood, blood spots, and semen. Bloodstains are a natural consequence of many crimes of violence; examinations usually concern the position, extent, and nature of suspected bloodstains. In most cases, the blood is dry. Preliminary chemical tests indicate only that the stain in question may be blood. The Teichmann and Takiyama tests, both of which rely on a reaction with hemoglobin, are confirmatory tests for blood. Human blood can be differentiated from the blood of other species by the precipitin test, which involves the reaction between blood and antihuman rabbit serum. Blood can also be categorized as to group: A, B, O, or AB.

The shape, position, and extent of bloodstains often give valuable information about the circumstances surrounding the commission of a crime. Blood falling from different heights forms stains of different shapes, which are further modified by the angle at which the blood drop hits the surface. Rough estimates can be made as to the age of bloodstains.

Ultraviolet light and the acid phosphatase test are used for the presumptive identification of semen. Positive identification for semen relies on identifying spermatozoa in the suspected sample.

Hairs and fibres. Hairs and fibres found at the scene of a crime can also be important evidence. It is not possible to trace hair to a specific individual, but narrow classes having probative value can be established. Human hair can be distinguished microscopically from other animal hairs. Hairs from various portions of the body vary in structure. Microscopic examination can reveal if hair has been recently cut, burned, or split. Chemical tests can demonstrate the presence of bleached or tinted hair. Neu-

tron activation analysis has been used in recent years to identify specific trace elements in hair and to measure the quantities of such trace elements.

The various fibres used in clothing, rope, furniture, etc., are of vegetable, animal, or synthetic origin. They can be differentiated by microscopic study and by simple chemical tests for solubility and other properties. Colour comparisons can be conducted using a comparison microscope or by spectrophotometric evaluation. Because they are mass-produced, fibres cannot be individualized.

Drugs and poisons. Toxicological examinations concern the identification of drugs and poisons, either in the form of pills and powders or as materials recovered from body organs. Among drugs or narcotics or both, of interest to crime laboratories are marijuana, morphine, heroin, and other narcotic alkaloids, amphetamines, barbiturates, tranquillizers, and LSD. These substances must be differentiated from nonprescription drugs. In the case of pills, the exact shape, colour, and external marking may serve as aids in identification. Laboratory test methods for specific chemical substances include colour reaction tests, crystal tests, chromatography, and infrared spectrophotometry. Normally, several tests are performed for the sake of confirmation.

Examination of body organs for the presence of poisonous material of any type requires an involved extraction procedure with a series of tests to confirm or eliminate the presence of a number of toxic materials. In addition to the drugs previously mentioned, other possible poisonous substances include gaseous and liquid poisons, metallic and inorganic poisons, organic poisons, vegetable and animal poisons, and bacterial poisons. The extraction procedures serve to separate volatile and nonvolatile substances, metallic and nonmetallic substances, and alkaline and acid-soluble substances.

Voice identification. The use of the sound spectrograph for voice identification is a relatively new development. The sound spectrograph is an instrument that graphically presents the time, frequency, and intensity of speech sound waves. The graphic forms for a sound such as *you*, as spoken by one person, can be compared with those produced by the speech of a second person and thus differentiated. The accuracy of the technique in identifying individuals is still in some doubt among speech scientists, even though voicegraphs, or voiceprints, have been used in courtroom proceedings.

BIBLIOGRAPHY. INSTITUTE FOR DEFENSE ANALYSES, *Task Force Report: Science and Technology* (1967), a report to the President's Commission on Law Enforcement and Administration of Justice showing how the resources of science and technology can be and are employed to control crime and to solve related sociological problems; HAROLD K. BECKER and GEORGE T. FELKENES, *Law Enforcement* (1968), a bibliography containing resource articles pertaining to law enforcement and 44,000 references; PAUL LELAND KIRK and LOWELL W. BRADFORD, *The Crime Laboratory* (1965), a basic presentation serving as a guide in establishing a crime laboratory; W.H.B. SMITH, *Small Arms of the World*, 9th rev. ed. (1969), a compilation of data containing photographs, technical descriptions, and specifications as well as pertinent operational comments; LOIS VERMILLON, "Electronic Systems to Strengthen the Long Arm of the Law," *Electronics*, 43:115–118 (1970), a contemporary analysis of the impact of electronic equipment systems on law enforcement problems in the U.S.; DONALD O. SCHULTZ and LORAN A. NORTON, *Police Operational Intelligence* (1968), a presentation of basics serving as a guide to the establishment of police intelligence systems; GEORGE B. JOHNSON and HANS BERT LOCKHOVEN, *International Armament*, 2 vol. (1965), a contemporary, basic source for information relating to small arms, their specifications, and manufacture; G.J. LUTZ (ed.), *Forensic Science: A Bibliography of Activation Analysis Papers*, National Bureau of Standards Technical Note 519 (1970); *The Challenge of Crime in a Free Society* (1967), a report by the U.S. President's Commission on Law Enforcement and Administration of Justice with major findings and recommendations developed from the examination of every facet of crime and law enforcement; *Proceedings of the National Symposium on Law Enforcement Science and Technology* (1967–), the most current articles pertaining to the various aspects of research and development programs in the field of criminal justice; INSTITUTE FOR DEFENSE ANALYSES, *A National Program of Research, Development, Test and Evaluation on Law Enforcement and Criminal Justice* (1968), a specific study for the U.S. Dept. of Justice to use in developing a coordinated research and development program; INSTITUTE FOR POLICE STUDIES, DEPT. OF CRIMINOLOGY, *Project Sky Knight: A Demonstration in Aerial Surveillance and Crime Control* (1968), a project designed for the U.S. Dept. of Justice to demonstrate and evaluate the effectiveness of the application of helicopters to police work; NEW YORK STATE IDENTIFICATION AND INTELLIGENCE SYSTEM, *System Development Plan: A New Concept in Criminal Justice Information-Sharing* (1967), and LOCKHEED AIRCRAFT CORPORATION, MISSILES AND SPACE COMPANY, *State of California Information System Study* (1965), initial planning documents.

The following references are the most significant and comprehensive texts pertaining to the establishment, management and operation of a modern forensic laboratory: H.J. WALLS, *Forensic Science* (1968); W.R. HARRISON, *Suspect Documents* (1958); D.A. CROWN, *The Forensic Examination of Paints and Pigments* (1968); J.W. OSTERBURG, *The Crime Laboratory* (1968); ARNE SVENSSON and O. WENDEL, *Techniques of Crime Scene Investigation*, 2nd rev. ed. (Eng. trans. 1965); L.C. NICKOLLS, *The Scientific Investigation of Crime* (1956); P.L. KIRK, *Crime Investigation* (1953); and C.E. O'HARA and J.W. OSTERBURG, *An Introduction to Criminalistics* (1949).

(W.B.McM.)

Political Parties

Political parties are groups organized for the purpose of achieving and exercising power within a political system. They originated in their modern form in Europe and the United States in the 19th century, along with the electoral and parliamentary systems whose development is reflected in their own evolution. The term party has since come to be applied to all organized groups seeking political power, whether by democratic elections or by revolution.

In earlier, prerevolutionary, aristocratic and monarchical regimes, the political process unfolded within restricted circles in which cliques and factions, grouped around particular noblemen or influential personalities, were opposed to one another. The establishment of parliamentary regimes and the appearance of parties at first scarcely changed this situation. To cliques formed around princes, dukes, counts, or marquesses there were added cliques formed around bankers, merchants, industrialists, and businessmen. Regimes supported by nobles were succeeded by regimes supported by other elites, whose political power was derived from industrial and mercantile wealth and who organized themselves into so-called cadre or brokerage parties in order to acquire the political power that their newfound wealth put within their reach. These narrowly based parties were later transformed to a greater or lesser extent, for the 19th century saw the emergence in Europe and America of parties depending on mass support.

Ubiquity of parties in the 20th century

The 20th century saw the spread of political parties throughout the entire world. In Africa large parties have sometimes been formed in which a modern organization has a more traditional ethnic or tribal basis; in such cases the party leadership is frequently made up of tribal chiefs. In certain areas of Asia, membership in modern political parties is often determined largely by religious factors or by affiliation with ritual brotherhoods. Many political parties in the developing countries are partly political, partly military; they are equally capable of participating in the electoral process and of inspiring guerrilla or revolutionary action. Certain Socialist and Communist parties in Europe had the same tendencies several decades earlier.

These last-mentioned European parties have demonstrated an equal aptitude for functioning within multiparty democracies and as the sole political party in a dictatorship. Developing originally within the framework of liberal democracy in the 19th century, political parties have been used in the 20th century by dictatorships for entirely undemocratic purposes. The point, then, is that political parties are institutions that can function in widely divergent political situations.

TYPES OF POLITICAL PARTY

A fundamental distinction can be made between cadre parties and mass-based parties. The two forms coexist in

many countries, particularly in western Europe, where Communist and Socialist parties have emerged alongside the older conservative and liberal parties. Many parties do not fall exactly into either category but combine some characteristics of both.

Cadre parties. Cadre parties—*i.e.*, parties dominated by politically elite groups of activists—developed in Europe and America during the 19th century. Except in some of the states of the United States, France from 1848, and the German Empire from 1871, the suffrage was largely restricted to taxpayers and property owners, and, even when the right to vote was given to larger numbers of people, political influence was essentially limited to a very small segment of the population. The mass of people were limited to the role of spectators rather than that of active participants.

The cadre parties of the 19th century reflected a fundamental conflict between two classes: the aristocracy on the one hand and the bourgeoisie on the other. The former, composed of landowners, depended upon rural estates on which a generally unlettered peasantry was held back by a traditionalist clergy. The bourgeoisie, made up of industrialists, merchants, tradesmen, bankers, financiers, and professional people, depended upon the lower classes of clerks and industrial workers in the cities. Both aristocracy and bourgeoisie evolved its own ideology. Bourgeois liberal ideology developed first, originating at the time of the English revolution of the 17th century in the writings of John Locke, an English philosopher. It was then developed by French philosophers of the 18th century. In its clamouring for formal legal equality and acceptance of the inequities of circumstance, liberal ideology reflected the interests of the bourgeoisie, who wished to destroy the privileges of the aristocracy and eliminate the lingering economic restraints of feudalism and mercantilism. But, insofar as it set forth an egalitarian ideal and a demand for liberty, bourgeois classical liberalism expressed aspirations common to all men. Conservative ideology, on the other hand, never succeeded in defining themes that would prove as attractive, for it appeared to be more closely allied to the interests of the aristocracy. For a considerable period, however, conservative sentiment did maintain a considerable impact among the people, since it was presented as the expression of the will of God. In Catholic countries, in which religion was based upon a hierarchically structured and authoritarian clergy, the conservative parties were often the clerical parties, as in France, Italy, and Belgium.

Conservative and liberal cadre parties dominated European politics in the 19th century. Developing during a period of great social and economic upheaval, they exercised power largely through electoral and parliamentary activity. Once in power, their leaders used the power of the army or of the police; the party itself was not generally organized for violent activity. Its local units were charged with assuring moral and financial backing to candidates at election time, as well as with maintaining continual contact between elected officials and the electorate. The national organization endeavoured to unify the party members who had been elected to the assemblies. In general, the local committees maintained a basic autonomy and each legislator a large measure of independence. The party discipline in voting established by the British parties—which were older because of the fact that the British Parliament was long established—was imitated on the Continent hardly at all.

United States parties in the 19th century

The first United States political parties of the 19th century were not particularly different from European cadre parties, except that their confrontations were less violent and based less on ideology. The first United States form of the struggle between the aristocracy and the bourgeoisie, between conservative and liberal, was carried out in the form of the Revolutionary War, in which Great Britain embodied the power of the king and the nobility, and the insurgents that of bourgeoisie and liberalism. Such an interpretation is, of course, simplified. There were some aristocrats in the South and, in particular, an aristocratic spirit based on the institutions of slaveholding and paternalistic ownership of land. In this sense, the Civil War could be considered as a second phase of violent conflict between the conservatives and the liberals. Nevertheless, the United States was from the beginning an essentially bourgeois civilization, based on a deep sense of equality and of individual freedom. Federalists and Anti-Federalists, Republicans and Democrats—all belonged to the liberal family since all shared the same basic ideology and the same system of fundamental values and differed only in the means by which they would realize their beliefs.

In terms of party structure, United States parties in the beginning differed little from their European counterparts. Like them, the United States parties were composed of local notables. The ties of a local committee to a national organization were even weaker than in Europe. At the state level there was some effective coordination of local party organizations, but at the national level such coordination did not exist. A more original structure was developed after the Civil War—in the South to exploit the vote of the blacks and along the East Coast to control the votes of immigrants. The extreme decentralization in the United States enabled a party to establish a local quasi-dictatorship in a city or county by capturing all of the key posts in an election. Not only the position of mayor but also the police, finances, and the courts came under the control of the party machine, and the machine was thus a development of the original cadre parties. The local party committee came typically to be composed of adventurers or gangsters who wanted to control the distribution of wealth and to ensure the continuation of their control. These men were themselves controlled by the power of the boss, the political leader who controlled the machine at the city, county, or state levels. At the direction of the committee, each constituency was carefully divided, and every precinct was watched closely by an agent of the party, the captain, who was responsible for securing votes for the party. Various rewards were offered to voters in return for the promise of their votes. The machine could offer such inducements as union jobs, trader's licences, immunity from the police, and the like. Operating in this manner, a party could frequently guarantee a majority in an election to the candidates of its choosing, and, once it was in control of local government, of the police, the courts, and public finances, etc., the machine and its clients were assured of impunity in illicit activities such as prostitution and gambling rings and of the granting of public contracts to favoured businessmen.

The degeneration of the party mechanism was not without benefits. The European immigrant who arrived in the United States lost and isolated in a huge and different world might find work and lodging in return for his commitment to the party. In a system of almost pure capitalism and at a time when social services were practically nonexistent, machines and bosses took upon themselves responsibilities that were indispensable to community life. But the moral and material cost of such a system was very high, and the machine was often a purely exploitative mechanism, performing no services to the community whatever.

By the end of the 19th century the excesses of the machines and the bosses and the closed character of the parties led to the development of primary elections, in which party nominees for office were selected. The primary movement deprived party leaders of the right to dictate candidates for election. A majority of the states adopted the primary system in one form or another between 1900 and 1920. The aim of the system was to make the parties more democratic by opening them up to the general public in the hope of counterbalancing the influence of the party committees. In practice, the aim was not realized, for the committees retained the upper hand in the selection of candidates for the primaries.

The British Labour Party

In its original form the British Labour Party constituted a new type of cadre party, forming an intermediate link with the mass-based parties. It was formed with the support of trade unions and left-wing intellectuals. At the base, each local organization sent representatives to a district labour committee, which was in turn represented at the national congress.

The early (pre-1918) Labour Party was thus structured of many local and regional organizations. It was not possible to join the party directly; membership came only through an affiliated body, such as a trade union. It thus represented a new type of party, depending not upon highly political individuals brought together as a result of their desire to acquire and wield power but upon the organized representatives of a broader interest—the working class. Certain Christian Democrat parties—the Belgian Social Christian Party between the two world wars and the Austrian Popular Party, for example—had an analogous structure: a federation of unions, agricultural organizations, middle class movements, employers' associations, etc. After 1918, the Labour Party developed a policy of direct membership on the model of the continental Socialist parties, individual members being permitted to join local-constituency branches. The majority of its membership, however, continued to be affiliated rather than direct.

Mass-based parties. Cadre parties normally organize a relatively small number of party adherents. Mass-based parties, on the other hand, unite hundreds of thousands of followers, sometimes millions. But the number of members is not the only criterion of a mass-based party. The essential factor is that such a party attempts to base itself on an appeal to the masses. It attempts to organize not only those who are influential or well known or those who represent special interest groups but rather any citizen who is willing to join the party. If such a party succeeds in gathering only a few adherents, then it is mass based only in potential. It remains, nevertheless, different from the cadre-type parties.

Socialist mass-based parties

At the end of the 19th century the Socialist parties of continental Europe organized themselves on a mass basis in order to educate and to organize the growing population of labourers and wage earners, which industrialization was making increasingly large and which was becoming more important politically because of extensions of the suffrage, and to gather the money necessary for propaganda by mobilizing in a regular fashion the resources of those who, although poor, were numerous. Membership campaigns were conducted, and each member paid party dues. If its members became sufficiently numerous, the party emerged as a powerful organization, managing large funds and diffusing its ideas among an important segment of the population. Such was the case with the German Social Democratic Party, which by 1913 had more than 1,000,000 members.

Such organizations were necessarily rigidly structured. The party required an exact registration of membership, treasurers to collect dues, secretaries to call and lead local meetings, and a hierarchical framework for the coordination of the thousands of local sections. A tradition of collective action and group discipline, more developed among workers as a result of their participation in strikes and other union activity, favoured the development and centralization of party organization.

A complex party organization tends to give a great deal of influence to those who have responsibility at various levels in the hierarchy, resulting in certain oligarchical tendencies. The Socialist parties made an effort to control this tendency by developing democratic procedures in the choice of leaders. At every level those in responsible positions were elected by members of the party. Every local party group would elect delegates to regional and national congresses, at which party candidates and party leaders would be chosen and party policy decided.

The type of mass-based party described above was imitated by many non-Socialist parties. Some cadre-type parties in Europe, both conservative and liberal, attempted to transform themselves along similar lines. The Christian Democrat parties often developed organizations copied even more directly from the mass-based model. But non-Socialist parties were generally less successful in establishing rigid and disciplined organizations.

The first Communist parties were splinter groups of existing Socialist parties and at first adopted the organization of these parties. After 1924, as a result of a decision of the Comintern (the Third International, or federation of working class parties), all Communist parties were transformed along the lines of the Soviet model, becoming mass parties based on the membership of the largest possible number of citizens, although membership was and is limited to those who embraced and espoused the ideology of Marxism–Leninism.

Communist mass-based parties

The Communist parties developed a new structural organization: whereas the local committees of cadre and Socialist parties focussed their organizing efforts and drew their support from a particular geographical area, Communist groups formed their cells in the place of work. The work-place cell was the first original element in Communist party organization. It grouped together all party members who depended upon the same firm, workshop, or store or the same professional institution (school or university, for example). Party members thus tended to be tightly organized, their solidarity, resulting from a common occupation, being stronger than that based upon residence.

The work-place-cell system proved to be effective, and other parties tried to imitate it, generally without success. Such an organization leads each cell to concern itself with problems of a corporate and professional nature rather than with those of a more political nature. These basic groups, however, smaller and, therefore, more numerous than the Socialist sections, tend to go their separate way. It is necessary to have a very strong party structure and for party leaders to have extensive authority if the groups are to resist such centrifugal pressure.

This has resulted in a second distinctive characteristic of the Communist parties: a high degree of centralization. Although all mass-based parties tend to be centralized, Communist parties are more so than others. There is, in principle, free discussion, which is supposedly developed at every level before a decision is made, but afterward all must adhere to the decision that has been made by the central body. The splintering that has from time to time divided or paralyzed the Socialist parties is forbidden in Communist parties, which have generally succeeded in maintaining their unity. A further distinctive characteristic of Communist parties is the importance given to ideology. All parties have a doctrine or at least a platform. The European Socialist parties, which were doctrinaire before 1914 and between the two wars, later became more pragmatic, not to say opportunistic. But, in Communist parties, ideology occupies a much more fundamental place, a primary concern of the party being to indoctrinate its members with Marxism.

The Fascist model

The 1920s and '30s saw the emergence of Fascist parties that attempted, as do the Communist and Socialist parties, to organize the maximum number of members but that did not claim to represent the great masses of people. Their teaching was authoritarian and elitist. They thought that societies should be directed by the most talented and capable people—by an elite. The party leadership, grouped under the absolute authority of a supreme head, constituted such an elite. Party structure had as its goal the assurance of the obedience of the elite.

This structure resembled that of armies, which are also organized in such a way as to ensure, by means of rigorous discipline, the obedience of a large number of men to an elite leadership. The party structure, therefore, made use of a military-type organization, consisting of a pyramid made up of units that at the base were very tiny but that, when joined with other units, formed groups that got larger and larger. Uniforms, ranks, orders, salutes, marches, and unquestioning obedience were all aspects of Fascist parties. This similarity rests upon another factor; namely, that Fascist doctrine taught that power must be seized by organized minorities making use of force. The party thus made use of a militia intended to assure victory in the struggle for control over the unorganized masses.

Large parties built upon the Fascist model developed between the two wars in Italy and Germany, where they actually came into power. Fascist parties appeared also in most other countries of western Europe during this period but were unable to achieve power. The less-developed nations of eastern Europe and Latin America were equally infected by the movement. The victory of the Allies in

1945, as well as the revelation of the horrors of Nazism, stopped the growth of the Fascists and provoked their decline, but Fascist-type party organization and doctrine remains a potent means of exercising power.

PARTIES AND POLITICAL POWER

Whether they are conservative or revolutionary, whether they are a union of notables or an organization of the masses, whether they function in a pluralistic democracy or in a monolithic dictatorship, parties have one function in common: they all participate to some extent in the exercise of political power, whether by forming a government or by exercising the function of opposition, a function that is often of crucial importance in the determination of national policy.

The struggle for power. It is possible in theory to distinguish revolutionary parties, which attempt to gain power by violence (conspiracies, guerrilla warfare, etc.), from those parties working within the legal framework of elections. But the distinction is not always easy to make, because the same parties may sometimes make use of both procedures, either simultaneously or successively, depending upon the circumstances. In the 1920s, for example, Communist parties sought power through elections at the same time that they were developing an underground activity of a revolutionary nature. In the 19th century, liberal parties were in the same situation, sometimes employing the techniques of conspiracy, as in Italy, Austria, Germany, Poland, and Russia, and sometimes confining their struggles to the ballot box, as in Great Britain and France.

Revolutionary methods vary greatly. Clandestine plots by which minority groups seize the centres of power presuppose monarchies or dictatorships in which the masses of people have little say in government. But terrorist and disruptive activity can serve to mobilize citizens and to demonstrate the powerlessness of any government. At the beginning of the 20th century leftist trade unionists extolled the revolutionary general strike, a total stoppage of all economic activity that would paralyze society completely and put the government at the revolutionaries' mercy. Rural guerrilla activity has often been used in countries with a predominantly agrarian society; urban guerrilla warfare was effective in the European revolutions of the 19th century, but the development of techniques of police and military control has made such activity more difficult.

Functions of parties within a democracy

Revolutionary parties are less numerous than parties that work within the law: the contest at election time is the means normally used in the struggle for power. Such activity corresponds, moreover, to the original nature of political parties and involves three factors: the organization of propaganda, the selection of candidates, and the financing of campaigns. The first function is the most visible. The party first of all gives the candidate a label that serves to introduce him to the voters and to identify his position. Because of this party label the voters are better able to distinguish the candidates. The promises and declarations of individuals are seldom taken with too much seriousness, and it means more to indicate that one candidate is a Communist, another a Socialist, a third a Fascist, and a fourth a liberal. Finally, the party also furnishes the candidate with workers to raise funds, put up his posters, distribute his literature, organize his meetings, and canvass from door to door.

The function of selecting candidates is exercised in three ways. In cadre parties, candidates are selected by committees of the party activists who make up the party—the caucus system, as it is known in the United States. In general, local committees play essential roles in this regard. In some countries, however, the selection is centralized by a national caucus, as, for example, by the Conservative Party in Britain and the Union of Democrats in France. In mass-based parties, selection is made by members of the regional and national congresses according to apparently democratic procedures; in actual practice, the governing committees play an essential role, the local constituency members generally ratifying their choice. Thirdly, in the United States the mechanism of primary elections has established a system for selecting candidates by means of the votes of all party members or all voters within a particular electoral district.

The various processes of selecting candidates do not, however, differ significantly in their results, for it is almost always the party leaders who play the essential role. This introduces an oligarchical tendency into party politics, a tendency that has not been overcome by the congresses of the mass-based parties or the United States primaries, which provide only a partial limitation on the power of the governing committees.

An important aspect of the struggle for power between political parties is the financing of campaigns. Cadre parties always have in their committees some key figure having connections with businessmen who is responsible for collecting gifts from them. In mass-based parties, rather than looking for large sums of money from a few people, leaders gather smaller sums from a large number of people who usually give on a monthly or annual basis. This method has been viewed as one of the distinguishing characteristics of mass-based parties. Sometimes the law intervenes in the financing of elections and of parties. Laws often limit campaign expenses and attempt to restrict the resources of the parties, but they are generally inoperative because it is quite easy to circumvent them. In some countries the state contributes public funds to the parties. At first, such financial participation was limited to expenses for campaigns and was based on the uniform treatment of candidates (as in France), but in Sweden and Finland the state contributes to the general finances of parties.

Participation in power. Only the functions of parties in democratic regimes will be considered at this time. The role of the single party in a dictatorship will be analyzed separately.

Once a political party has achieved electoral victory, the question arises of how much influence the party is to have on the government. The influence of the party on members in elective office is frequently quite weak. It defines the general lines of their activity, but these lines can be quite hazy, and few decisions are taken in the periodic meetings between officeholders and their party. Each member of the legislature retains personal freedom of action in his participation in debates, in his participation in government, and, especially, in his voting. The party may, of course, attempt to enforce the party line, but parliamentary or congressional members cannot be compelled to vote the way the party wants them to. Such is the situation in the United States, within most of the liberal and conservative European parties, and within cadre parties in general.

Flexible and rigid parties

The question of how disciplined a party is, of the extent to which it will always present a united front, enables a distinction to be made between what may be termed rigid and flexible parties; that is, between those that attempt always to be united and disciplined, following what is most often an ideologically based party line, and those that, representing a broader range of interests and points of view, form legislatures that are assemblies of individuals rather than of parties.

Whether the parties operating within a particular system will be rigid or flexible depends largely on the constitutional provisions that determine the circumstances in which a government may continue in office. This is clearly illustrated by comparing the situation in the United States with that in Great Britain. In the United States the president and his government continue in office for the constitutionally defined period of four years, regardless of whether a majority in the legislature supports him or not. Since a united party is thus not crucial to the immediate survival of the government, both major parties are able to contain broad coalitions of interests, and votes on issues of major importance frequently split each party. In the United Kingdom the situation is quite different. There, government can continue in office only so long as it commands a majority in the legislature. A single adverse vote can result in the dissolution of Parliament and a general election. Party discipline and unity are thus of crucial importance, and this fact has far-reaching conse-

quences for the composition, organization, and policies of each party. The consequences of party disunity within such a constitutional framework are well illustrated by the weakness and instability of the governments of the Third and Fourth French republics.

The distinction between flexible and rigid parties applies equally to parties in power and to those forming the opposition. Votes of censure or of lack of confidence, votes on proposed legislation or on the budget, questions put to ministers or challenges made to them—in short, all the functions of an opposition party—are worked out differently in flexible and rigid party systems.

In flexible party systems the absence of strong discipline is often of great consequence to the opposition party because only rigid parties can constitute an opposition force sufficiently strong to counterbalance the strength of the party in power. At the same time, party discipline permits the opposition to present the public with an alternative to the majority party; the logical consequence of such a situation is Britain's "shadow cabinet," which accustoms the electorate to the idea that a new group is ready to take over the reins of government.

Parties provide, moreover, a channel of communication between opposition legislators and the public. The governing party performs a similar service for the government, although it is less necessary, since the government has at its disposal numerous means of communicating with the public. Opposition parties thus provide a means of expressing negative reaction to decisions of government and proposing alternatives. This role justifies the official recognition given to opposition parties, as is the case in Great Britain and Scandinavia.

Power and representation. It is difficult to envisage how representative democracy could function in a large industrialized society without political parties. In order for citizens to be able to make an intelligent choice of representative or president, it is necessary for them to know the real political orientation of each candidate. Party membership provides the clearest indication of this.

Individual candidates and the party

The programs and promises of each individual candidate are not too significant or informative, because most candidates, in their attempt to gain the most votes, try to avoid difficult subjects; they all tend to speak the same language; that is, to camouflage their real opinions. The fact that one is a Socialist, another a conservative, a third a liberal, and a fourth a Communist provides a far better clue as to how the candidate will perform when in office. In the legislature the discipline of the party limits the possibility that elected representatives will change their minds and their politics, and thus the party label acts as a sort of guarantee that there will be at least some correspondence between promise and performance. Parties make possible the representation of varying shades of opinion by synthesizing different positions into a stance that each representative adopts to a greater or lesser extent.

But parties, like all organizations, tend to manipulate their members, to bring them under the control of an inner circle of leaders that often perpetuates itself by cooptation. In cadre parties, members are manipulated by powerful committees containing cliques of influential party leaders. In mass-based parties, leaders are chosen by the members, but incumbents are very often re-elected because they control the party apparatus, using it to ensure their continuation in power.

Democratic political systems, while performing the function of representation, thus rest more or less on the competition of rival oligarchies. But these oligarchies consist of political elites that are open to all with political ambition. No modern democracy could function without parties, the oligarchical tendencies of which are best regarded as a necessary evil.

PARTY SYSTEMS

Party systems may be broken down into three broad categories: two-party, multiparty, and single-party. Such a classification is based not merely on the number of parties operating within a particular country but on a variety of distinctive features that the three systems exhibit. Two-party and multiparty systems represent means of organizing political conflict within pluralistic societies and are thus part of the apparatus of democracy. Single parties usually operate in situations in which genuine political conflict is not tolerated. This broad statement is, however, subject to qualification, for, although single parties do not usually permit the expression of points of view that are fundamentally opposed to the party line or ideology, there may well be intense conflict within these limits over policy within the party itself. And even within a two-party or a multiparty system, debate may become so stymied and a particular coalition of interests so entrenched that the democratic process is seriously compromised.

The distinction between two-party and multiparty systems is not as easily made as it might appear. In any two-party system there are invariably some tiny parties in addition to the two major parties, and there is always the possibility that such small parties might prevent one of the two main parties from gaining a majority of seats in the legislature. This is the case with regard to the Liberal Party in Great Britain, for example. Other countries do not fall clearly into either category; thus, Austria and the Federal Republic of Germany only approximate the two-party system. It is not simply a question of the number of parties that determines the nature of the two-party system; many other elements are of importance, the extent of party discipline in particular.

The multiparty system. In Anglo-Saxon countries there is a tendency to consider the two-party system as normal and the multiparty system as the exceptional case. But, in fact, the two-party system that operates in Great Britain, the United States, and New Zealand is much rarer than the multiparty system, which is found in almost all of western Europe.

In western Europe, three major categories of parties have developed since the beginning of the 19th century: conservative, liberal, and Socialist. Each reflects the interests of a particular social class and expounds a particular political ideology. After World War I other categories of parties developed that were partly the result of divisions or transformations of older parties. Communist parties began as splinter groups of Socialist parties, and Christian Democrat parties attempted to weld together moderate Socialists and conservatives and some liberals. Other distinctive types of party emerged in some countries. In Scandinavia, liberal rural parties developed in the 19th century, reflecting a long tradition of separate representation of the rural population. In many countries ethnic minorities formed the basis of nationalist parties, which then either joined existing parties or divided them.

The appearance of Socialism in the 19th century upset the earlier lines of battle between conservatives and liberals and tended to throw the latter two groups into a common defense of capitalism. Logically, this situation should have led to the fusion of conservatives and liberals into one bourgeois party that would have presented a united stand against Socialism. This is, in fact, what happened in Great Britain after World War I.

Influence of the electoral system on the number of parties

One of the most important factors determining the number of parties operating within a particular country is the electoral system (see ELECTORAL PROCESSES). Proportional representation tends to favour the development of multiparty systems because it ensures representation in the legislature for even small parties. The majority, single-ballot system tends to produce a two-party system, because it excludes parties that may gain substantial numbers of votes but not the majority of votes necessary to elect a representative within a constituency. The majority system with a second ballot favours a multiparty system tempered by alliances between parties. Such a system is very rare, found only in the German Empire (1871–1914) and in the French Third (1870–1940) and Fifth (since 1958) republics. Voters choose between the parties that did best in a first ballot. This leaves small parties at a disadvantge but, nevertheless, gives them opportunity to strengthen their role during the second balloting as long as they are willing to enter into alliances with the leading parties.

Another factor producing multiparty systems is the intensity of political conflicts. If, within a given political movement, extremists are numerous, then it is difficult for the moderates in that party to join with them in a united front. Two rival parties are likely to be formed. Thus, the power of the Jacobins among 19th-century French liberals contributed to the inability of the moderates to form one great liberal party, as was successfully achieved in Great Britain. Likewise, the power of the extremists among the conservatives was an obstacle to the development of a strong conservative party.

The distinction between the multiparty system and the two-party system corresponds largely to a distinction between two types of Western political regime. In a two-party situation the administration has, in effect, an assurance of a majority in the legislature, deriving from the predominance of one party; it has, therefore, a guarantee of continuance and effectiveness. Such a system is often referred to as majority parliamentarianism. In a multiparty situation, on the other hand, it is quite rare for one party to have a majority in the legislature; governments must, therefore, be founded on coalitions, which are always more heterogeneous and more fragile than a single party. The result is less stability and less political power. Such systems may be referred to as nonmajority parliamentarianism.

In practice, majority and nonmajority parliamentary systems do not coincide exactly with two-party systems and multiparty systems. For, if each of the two parties is flexible and does not control the voting patterns of its members (as is the case in the United States), the numerical majority of one of the parties matters little. It can happen, moreover, that one party in a multiparty system will hold an absolute majority of seats in the legislature so that no coalition is required. Such a situation is unusual but has occurred in West Germany, Italy, and Belgium at various times since 1945.

Ordinarily, however, a coalition will be the only means of attaining a parliamentary majority within the framework of the multiparty system. Coalitions are by nature more heterogeneous and more unstable than a grouping made up of one party, but their effectiveness varies greatly according to the discipline and organization of the parties involved. In the case of flexible parties that are undisciplined and that allow each legislator to vote on his own, the coalition will be weak and probably short-lived. The instability and weakness of governments is at its maximum in such situations, of which the Third French Republic provides a good example.

If, on the other hand, the parties involved in a coalition are rigid and disciplined, it is possible for a system quite similar to the two-party system to develop. This is often the case when two opposing alliances are formed, one on the left and one on the right, and when both are strong enough to endure through the legislative session. This type of coalition, referred to as bipolarized, introduces elements of the two-party system into a multiparty framework. A situation of this type has developed in Sweden, where conservative, liberal, and agrarian parties have been aligned against the Social Democrat Party, which eventually allied itself with the Communist Party (1970).

The system of bipolar alliances may be contrasted with the system of a centrist alliance. Rather than the parties on the right forming a centre–right coalition to oppose a centre–left coalition, there is the possibility that the centre–left and the centre–right will join forces and reject the extremes at both ends of the political spectrum. Such a situation occurred in Germany during the Weimar Republic, when the government rested on a majority formed of a coalition of Catholic Centrists and Social Democrats, with opposition coming from the Communists and the nationalists on the extreme left and right.

Benefits and disadvantages of coalition politics

Centrist coalitions all tend to give the average citizen a sense of political alienation. In rejecting both extremes, coalitions may well be isolating the radical, unstable elements, but the governing coalition may tend to be unresponsive to new ideas, uninspiringly pragmatic, and too ready to compromise. This situation gives rise to a more or less permanent breach between practical politics and political ideals. An advantage of bipolarization or of the two-party system is that the moderates of both sides must collaborate with those who are more extreme in their views, and the extremists must be willing to work with those who are more moderate; the pressure from the extremists prevents the moderates from getting bogged down, while collaboration with the moderates lends a touch of realism to the policies of the extremists.

The two-party system. A fundamental distinction must be made between the two-party system as it is found in the United States and as it is found in Great Britain. Although two major parties dominate political life in the two countries, the system operates in quite different ways.

The American two-party system. The United States has always had a two-party system, first in the opposition between the Federalists and the Anti-Federalists, then in the competition between the Republicans and the Democrats. There have been frequent third-party movements in the history of the country, but they have always failed. Presidential elections seem to have played an important role in the formation of this type of two-party system. The mechanism of a national election in so large a country has necessitated very large political organizations and, at the same time, relatively simplified choices for the voter.

American parties are different from their counterparts in other Western countries. They are not tied in the same way to the great social and ideological movements that have so influenced the development of political life in Europe during the last two centuries. There have been Socialist parties at various times in the history of the United States, but they have never challenged the dominance of the two major parties. It can be argued that the main reason for the failure of Socialist parties in America has been the high degree of upward mobility permitted by a rich and continually expanding economy. The consequence of this mobility has been that class consciousness has never developed in the United States in a manner that would encourage the formation of large Socialist or Communist parties.

In comparison with European political movements, therefore, American parties have appeared as two varieties of one liberal party, and within each party can be found a wide range of opinion, going from the right to the left.

The American parties have a flexible and decentralized structure, marked by the absence of discipline and rigid hierarchy. This was the structure of most of the cadre-type parties of the 19th century, a structure that most liberal parties have retained. Federalism and a concern for local autonomy accentuate the lack of rigid structure and the weakness of lines of authority in the parties. Organization may be relatively strong and homogeneous at the local level, but such control is much weaker on the state level and practically nonexistent on the national level. There is some truth to the observation that the United States has not two parties but 100—that is, two in each state. But it is also true that each party develops a certain degree of national unity for the presidential election and that the leadership of the president within his party gives the victorious party some cohesion.

Coalition voting patterns in the U.S. Congress

In voting, Republicans and Democrats are usually found on both sides. An alliance between liberal Republicans and Democrats against conservative Republicans and Democrats tends to develop. But neither bloc is stable, and the alignment varies from one vote to another. As a consequence, despite the existence of a two-party system, no stable legislative majority is possible. In order to have his budget adopted and his legislation passed, the president of the United States must patiently try to gather the necessary votes on every question; like Penelope, he has the wearisome task of beginning anew each day. The American two-party system is thus a pseudo-two-party system, because each party provides only a loose framework within which shifting coalitions are formed.

The British two-party system. Another form of the two-party system is operative in Great Britain and in New Zealand. The situation in Australia is affected somewhat by the presence of a third party, the Country Party.

A tight alliance between the Australian Liberal Party and the Country Party introduces, however, a rather rigid bipolarization with the Labour Party. The system thus tends to operate on a two-party basis. Canada also possesses what is essentially a two-party system, Liberals or Conservatives usually being able to form a working majority without the help of the small, regionally based parties.

Great Britain has had two successive two-party alignments: Conservative and Liberal prior to 1914 and Conservative and Labour since 1935. The period from 1920 to 1935 constituted an intermediate phase between the two. Britain's Conservative Party is actually a Conservative–Liberal Party, resulting from a fusion of the essential elements of the two great 19th-century parties. Despite the name Conservative, its ideology corresponds to political and economic liberalism. A similar observation could be made about the other major European conservative parties, such as the German Christian Democratic Party and the Belgian Social Christian Party.

Importance of rigid party lines in the British system

The British two-party system depends on the existence of rigid parties; that is, parties in which there is effective discipline regarding parliamentary voting patterns. In every important vote, all party members are required to vote as a bloc and to follow to the letter the directives that they agreed upon in common or that were decided for them by the party leaders. A relative flexibility may at times be tolerated, but only to the extent that such a policy does not compromise the action of the government. It may be admissible for some party members to abstain from voting if their abstention does not alter the results of the vote. Thus, the leader of the majority party (who is at the same time the prime minister) is sure to remain in power throughout the session of Parliament, and the legislation he proposes will be adopted. There is no longer any real separation of power between the executive and legislative branches, for the government and its parliamentary majority form a homogeneous and solid bloc before which the opposition has no power other than to make its criticisms known. During the four or five years for which a Parliament meets, the majority in power is completely in control, and only internal difficulties within the majority party can limit its power.

Since each party is made up of a disciplined group with a recognized leader who becomes prime minister if his party wins the legislative elections, these elections perform the function of selecting both the legislature and the government. In voting to make one of the party leaders the head of the government, the British assure him of a disciplined parliamentary majority. The net result is a political system that is at once stable, democratic, and strong; Many would argue that it is more stable, more democratic, and stronger than systems anywhere else in the world.

This situation presupposes that both parties are in agreement with regard to the fundamental rules of a democracy. If a Fascist party and a Communist party were opposed to one another in Great Britain, the two-party system would not last very long. The winner would zealously suppress the opponent and rule alone.

The system, of course, does have its weak points, especially insofar as it tends to frustrate the innovative elements within both parties. But it is possible that this situation is preferable to what would happen if the more extreme elements within the parties were permitted to engage in unrealistic policies. The risk of immobility is in fact a problem for any party in a modern industrial society, and not just for those in a two-party situation. The problem is related to the difficulties involved in creating new organizations capable of being taken seriously by an important segment of the population and in revitalizing long-standing organizations encumbered by established practices and entrenched interests.

Single-party systems. There have been three historical forms of the single-party system: Communist, Fascist, and that found in the developing countries.

The Communist model. In Communist countries the party is considered to be the spearhead of the urban working class and of other workers united with it (peasants, intellectuals, etc.). Its role is to aid in the building of a Socialist regime during the transitory phase between capitalism and pure Socialism, called the dictatorship of the proletariat. An understanding of the exact role of the party requires an appreciation of the Marxist conception of the evolution of the state. In countries based on private ownership of the means of production, the power of the state, according to the Marxist point of view, is used to further the interests of the controlling capitalists. In the first stage of revolution the power of the state is broken. Power, however, still has to be wielded to prevent counterrevolution and to facilitate the transition to Communism, at which stage coercion will no longer be necessary. Thus, the party, in effect, assumes the coercive functions of the state during the dictatorship of the proletariat or, to be more accurate, during the dictatorship of the party in the name of the proletariat.

Role of the party in Communist states

It is thus the party hierarchy and not the official state hierarchy that has the real power. The first secretary of the party is the most important figure in the government, and, whether the party leadership is in the hands of one man or several, the party remains the centre of political power.

In China the role of the party has been modified to some extent. The Great Cultural Revolution (1966–69) seems to have been a device used by Mao Tse-tung to reduce the power of the party, over which he was losing control. He sought support instead from the masses of fanatical youth and the army. Since then, the party has been gradually reorganized, but it is no longer as dominant as it was or as is usual within Communist states. The role of the army appears to be preponderant; moreover, it is an army that has been very politicized and that has always maintained close contact with the population.

In all Communist countries, the structure of the party has been determined largely by the need for it to govern firmly while at the same time maintaining its contact with the masses of the people. Party members are a part of the general public, of which they are the most active and most politically conscious members. They remain in contact with the masses by means of a network of party cells that are present everywhere. Party leaders are thus always "listening in on the masses," and the masses are always informed of decisions of party leaders, as long as the communication network is working in both directions.

The party is not only a permanent means of contact between the people and party leaders; it is also a propaganda instrument. Political indoctrination is essential to the survival of Communist parties, and many resources are devoted to it. Indoctrination is accomplished in training schools, by means of "education" campaigns, carefully censored newspapers, periodicals, and books, and the untiring efforts of militants, who play a role similar to that of the clergy in organized religion. The party is thus the guardian of orthodoxy and has the power to condemn and to excommunicate.

The Fascist model. Fascist parties in a single-party state have never played as important a role as Communist parties in an analogous situation. In Italy, the Fascist party was never the single most important element in the regime, and its influence was often secondary. In Spain the Falange never played a crucial role, and in Portugal the National Union was a very weak organization even at the height of dictator António Salazar's strength. Only in Germany did the National Socialist Party have a great influence on the state. But, in the end, Hitler's dictatorship was dependent on his private army, the SS (Schutzstaffel), which formed a separate element within the party and which was closed to outside influences, and on the Gestapo, which was a state organization and not an organization of the party. The Fascist party in the single-party state has a policing or military function rather than an ideological one.

After their rise to power, the Fascist parties in both Germany and Italy gradually ceased to perform the function of maintaining contact between the people and the government, a function that is usually performed by the party in a single-party situation. It was possible to ob-

serve a tendency for the party to close in upon itself while suppressing its deviant members. The renewal of the party was then assured through recruitment from youth organizations, from which the most fanatical elements, the products of a gradual selection process starting at a very early age, entered the party. The party tended, therefore, to constitute a closed order.

The single party in the developing countries. Some of the Communist parties in power in developing countries do not differ significantly from their counterparts in industrialized countries. This is certainly true of the Communist Party of the Democratic Republic of Vietnam and the Communist Party of North Korea. There are, however, countries in Africa, Asia, and Latin America in which the single party in power cannot be characterized in terms of a European counterpart. This observation applies to, for example, the Arab Socialist Union in Egypt, the Neo-Destour Party in Tunisia, and the National Liberation Front in Algeria, as well as many parties in black Africa.

Most of these parties claim to be more or less Socialist or at least progressive, while remaining far removed from Communism and, in some cases, ardent foes of Communism. President Nasser attempted to establish a moderate and nationalistic Socialism in Egypt. In Tunisia the Neo-Destour Party is more republican than Socialist and is inspired more by the example of the reforms in Turkey under Kemal Atatürk than by Nasserism. In black Africa, single parties often claim to be Socialist, but with few exceptions they rarely are in practice. But, while they may not be revolutionary, they are reform oriented.

Single parties in developing countries are rarely as well organized as Communist parties. In Turkey the Peoples' Republican Party was more a cadre party than a mass-based party. In Egypt it has been necessary to organize a core of professional politicians within the framework of a pseudoparty of the masses. In black Africa the parties are most often genuinely mass based, but the membership appears to be motivated primarily by personal attachment to the leader or by tribal loyalties, and organization is not usually very strong. It is this weakness in organization that explains the secondary role played by such parties in government.

Some regimes, however, have endeavoured to develop the role of the party to the fullest extent possible. The politics of Atatürk in Turkey were an interesting case study in this regard. It was also Nasser's goal to increase the influence of the Arab Socialist Union, thereby making it the backbone of the regime. This process is significant in that it represents an attempt to move away from the old-fashioned dictatorship, supported by the army or based on tribal traditions or on charismatic leadership, toward a modern dictatorship, supported by one political party. Single-party systems can institutionalize dictatorships by making them survive the life of one dominant figure.

The necessity of parties in democratic systems

It has often been said in the West that political parties are in a state of decline. Actually, this has been a long-standing opinion in certain conservative circles, arising largely out of a latent hostility to parties, which are viewed as a divisive force among citizens, a threat to national unity, and an enticement to corruption and demagoguery. In certain European countries—France, for example—right-wing political organizations have even refused to call themselves parties, using instead such terms as movement, union, federation, and centre. And it cannot be denied that to some extent the major European and American parties of the 1970s do appear old and rigid in comparison with their condition at the turn of the century or immediately following World War I. Even relatively new parties, such as the Christian Democratic parties of Germany and Italy (founded in 1945), seem somewhat lifeless.

In terms of size and number, however, political parties are not declining but growing. At the turn of the century they were confined mainly to Europe and North America; elsewhere they were quite weak or nonexistent. In the second half of the 20th century, parties are found practically everywhere in the world. And in Europe and North America there are generally far more people holding membership in parties than prior to 1914. Parties of the 1970s are larger, stronger, and better organized than those of the late 19th century. In the industrialized nations, especially in western Europe, parties have become less revolutionary and innovative, and this factor may explain the rigid and worn-out image that they sometimes present. But even this phenomenon is found only in a limited area and may, perhaps, pass.

The growth of parties into very large organizations may be responsible for the feelings of powerlessness on the part of many individuals who are involved in them. This is a problem experienced by people who find themselves part of any large organization, whether it be a political party, business enterprise, corporation, or union. The difficulties involved in reforming or changing political parties that have become large and institutionalized, coupled with the next-to-impossible task of creating new parties likely to reach sufficient strength to be taken seriously by the electorate, have resulted in much frustration and impatience with the party system. But it is difficult to imagine how democracy could function in a large industrialized country without political parties. In the modern world, democracy and political parties are two facets of the same reality, the inside and outside of the same fabric.

BIBLIOGRAPHY. ROBERT MICHELS, *Les Partis politiques: essai sur les tendances oligarchiques des democraties* (1914; Eng. trans., *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracies,* 1915), provides the first modern theory of political parties, stemming from the doctrines of the "elite," or "ruling class," of Gaetano Mosca. The book is based largely on the example of German social democracy. MAURICE DUVERGER, *Les Partis Politiques* (1951; Eng. trans., *Political Parties: Their Organization and Activity in the Modern State,* 2nd ed., 1964), gives a comparative and systematic study of political parties, emphasizing the experience of the European countries. JOSEPH G. LAPALOMBARA and MYRON WEINER (eds.), *Political Parties and Political Development* (1966), is a collection of contributions by various authors and is the best recent comparative explication of problems concerning political parties. D.W. RAE, *The Political Consequences of Electoral Laws* (1967), provides, on the basis of an examination of the balloting in numerous countries since 1945, an analysis of the relationship between the electoral system and political parties. It is a model of comparative study. Works about American parties are very numerous. Recommended are SAMUEL J. ELDERSVELD, *Political Parties: A Behavioral Analysis* (1964); and WILLIAM N. CHAMBERS and WALTER D. BURNHAM (eds.), *The American Party Systems* (1967). There are very few works that study European political parties comparatively. STANLEY HENIG and JOHN PINDER (eds.), *European Political Parties* (1969), however, provides much factual information in a concise and systematic format. On Great Britain, SAMUEL H. BEER, *Modern British Politics: A Study of Parties and Pressure Groups* (1965); and R.T. MCKENZIE, *British Political Parties: The Distribution of Power Within the Conservative and Labour Parties,* 2nd rev. ed. (1963), are particularly useful. D.E. APTER, *The Politics of Modernization* (1965); THOMAS L. HODGKIN, *African Political Parties* (1961); and A. MAHIOU, *L'Avènement du parti unique en Afrique noire* (1969), provide accounts of the operations of political parties in developing countries.

(M.Du./Ed.)

Political Philosophy

The central problem of political philosophy is how to deploy or limit public power so as to maintain the survival and enhance the quality of human life. Like all aspects of human experience, it is conditioned by environment and by the scope and limitations of mind; and the answers given by successive political philosophers to perennial problems reflect the knowledge and the assumptions of their times. Political philosophy, as distinct from the study of political and administrative organization, is more theoretical and normative than descriptive. It is inevitably related to general philosophy and is itself a subject of social anthropology, sociology, and the sociology of knowledge. As a normative discipline it is thus concerned with what ought, on various assumptions, to be and how this purpose can be promoted, rather than with a description of facts—although any realistic political theory is necessarily related to these facts. The politi-

cal philosopher is thus not concerned so much, for example, with how pressure groups work or how by various systems of voting decisions are arrived at, as with what the aims of the whole political process should be in the light of a particular philosophy of life.

There is thus a distinction between political philosophy, which reflects the world outlook of successive theorists and which demands an appreciation of their historical settings, and modern political science proper, which, insofar as it can be called a science, is empirical and descriptive. Political philosophy, however, is not merely unpractical speculation, though it may give rise to highly impractical myths: it is a vitally important aspect of life, and one that, for good or evil, has had decisive results on political action; for the assumptions on which political life is conducted clearly must influence what actually happens. Political philosophy may thus be viewed as one of the most important intellectual disciplines, for it sets standards of judgment and defines constructive purposes for the use of public power. Such consideration of the purposes for which power should be used are in a very real sense more urgent today than they have been in any earlier historical period, for mankind has at its disposal the power either to create a world civilization in which modern technology can immeasurably benefit the human race, or to destroy itself in pursuit of political myths. In this situation the scope for political philosophy is great, the clarification of its purpose and limitations urgent—an aspect, indeed, of civilization's survival.

Significance of political philosophy

Despite this unique aspect of the contemporary situation and although ancient political philosophies were formulated under very different conditions, their study still illuminates questions vital today. Questions concerning the aims of government, the grounds of political obligation, the rights of individuals against the state, the basis of sovereignty, the relation of executive to legislative power, and the nature of political liberty and social justice have been asked and answered in many ways over the centuries. They are all fundamental to political philosophy and demand answers in terms of modern knowledge and opinion.

The central purpose of this article is to describe how these questions have been asked and answered by representative and influential political philosophers, from Greco-Roman antiquity through the Middle Ages, early modern times, and the 19th and early 20th centuries. During so long a time span the historical context of these formulations has changed profoundly, and an understanding of the political philosophers selected demands some account of their background. Because of limitations of space, only political philosophers of outstanding importance have been at all fully described, although many minor figures also are briefly discussed.

The article is divided into the following sections:

I. The history of political philosophy in the West to the end of the 19th century
- Antiquity
  - Plato
  - Aristotle
  - Cicero and the Stoics
  - St. Augustine
- The Middle Ages
  - John of Salisbury
  - Aquinas
  - Dante
- The 16th to the 18th centuries
  - Machiavelli
  - Hobbes
  - Spinoza
  - Richard Hooker's adapted Thomism
  - Locke
  - Burke
  - Vico
  - Montesquieu
  - Rousseau
- The 19th century
  - Utilitarianism
  - Tocqueville
  - T.H. Green
  - Liberal nationalism
  - American constitutionalism
  - Anarchism and utopianism
  - Saint-Simon and Comte
  - Hegel
  - Marx and Engels

II. Political philosophy in the 20th century
- Political pragmatism
- Religious and Existentialist approaches
- Revolutionary doctrines
  - Lenin
  - Other Marxist approaches

III. Conclusion

I. The history of political philosophy in the West to the end of the 19th century

ANTIQUITY

Although in antiquity great civilizations arose in Egypt and Mesopotamia, in the Indus Valley, and in China, there was little speculation about the problems of political philosophy as formulated in the West and since predominant. The laws of Hammurabi of Babylon (*c.* 1750 BC) are rules propounded by the monarch as a representative of God on Earth and are mainly concerned with order, trade, and irrigation; the *Admonitions* of the Egyptian vizier Ptahhotep (*c.* 2300 BC) are shrewd advice on how to prosper in a bureaucracy; and the *Arthaśāstra* of Kauṭilya, grand vizier to the Indian Candragupta Maurya in the late 4th century BC, are Machiavellian precepts on how to survive under an arbitrary power. To be sure, the Buddhist concept of *dharma* (social custom and duty), which inspired the Indian emperor Aśoka in the 3rd century BC, implies a moralization of public power, and the teachings of Confucius in the 6th century BC are a code of conduct designed to stabilize society; but there is not, outside Europe, much speculation about the basis of political obligation and the purpose of the state, with both of which Western political philosophy is mainly concerned. An authoritarian society is taken for granted, backed by religious sanctions, and a conservative and arbitrary power is generally accepted.

In contrast to this overwhelming conservatism, paralleled by the rule of custom and tribal elders in most primitive societies, the political philosophers of ancient Greece question the basis and purpose of government; and, though they do not separate political speculation from shrewd observations that today would be regarded as empirical political science, they created the vocabulary of Western political thought.

Plato. The first elaborate work of European political philosophy is *The Republic* of Plato (*c.* 378 BC), a masterpiece of insight and feeling, superbly expressed in dialogue form and probably meant for recitation. Further development of Plato's ideas is undertaken in his *Statesman* and *Laws,* the latter prescribing the ruthless methods whereby they might be imposed. Plato grew up during the great war between Athens and Sparta in which Athens suffered defeat and, like many political philosophers, tried to find remedies for prevalent political injustice and decline. Indeed, *The Republic* is the first of the utopias, though not one of the more attractive; and it is the first classic attempt of a European philosopher to moralize political life.

Cast as a lively discussion between Socrates, whose wisdom Plato is recounting, and various leisured Athenians, Books V, VII–VIII, and IX of *The Republic* state the major themes of political philosophy with poetic power. Plato's work has been criticized as static and class bound, reflecting the moral and aesthetic assumptions of an elite in a slave-owning civilization and bound by the narrow limits of the city-state. The work is indeed a classic example of a philosopher's vivisection of society, imposing by relatively humane means the rule of a high-minded minority.

Epistemological assumptions in *The Republic*

The Republic is a criticism of current Hellenic politics —often an indictment. It is based upon a metaphysical act of faith, for Plato believes that a world of permanent Forms exists beyond the limitations of human experience and that morality and the good life, which the state should promote, are reflections of these ideal Forms (see PLATO). The point is best made in the famous simile of the cave, in which men are chained with their faces to the

wall and their backs to the light, so that they see only the shadows of reality. So constrained, they shrink from what is truly "real" and permanent and need to be forced to face it. This Idealistic doctrine, known misleadingly as Realism (in nontechnical language it is hardly realistic), pervades all Plato's philosophy: its opposite doctrine, Nominalism, declares that only particular and observed "named" data are accessible to the mind. On his Realist assumption, Plato, who was perhaps influenced by Indian thought, regards most ordinary life as illusion and the current evils of politics as the result of men pursuing brute instinct. It follows that

> unless philosophers bear kingly rule in cities or those who are now called kings and princes become genuine and adequate philosophers, and political power and philosophy are brought together . . . there will be no respite from evil for cities.

Only philosopher-statesmen can apprehend permanent and transcendant Forms and turn to "face the brightest blaze of being" outside the cave, and only philosophically minded men of action can be the saviours and helpers of the people.

Plato is thus indirectly the pioneer of modern beliefs that only a party organization, inspired by correct and "scientific" doctrines, formulated by the written word and interpreted by authority, can rightly guide the state. His rulers would form an elite, not responsible to the mass of the people. Thus, in spite of his high moral purpose, he has been called an enemy of the open society and the father of totalitarian lies. But he is also an anatomist of the evils of unbridled appetite and political corruption and insists on the need to use public power to moral ends.

Having described his utopia, Plato turns to analyze the existing types of government in human terms with great insight. Kingly government is the best but impracticable; in oligarchies the rule of the few and the pursuit of wealth divide societies—the rich become demoralized and the poor envious, and there is no harmony in the state. In democracy, in which the poor get the upper hand, demagogues distribute "a peculiar kind of equality to equals and unequals impartially," and the old flatter the young, fawning on their juniors to avoid the appearance of being sour or despotic. The leaders plunder the propertied classes and divide the spoils among themselves and the people until confusion and corruption lead to tyranny, a worse form of government. For the tyrant becomes a wolf instead of a man and "lops off" potential rivals and starts wars to distract the people from their discontent. "Then, by Zeus," Plato concludes, "the public learns what a monster they have begotten."

Plato's *Statesman* and *Laws*

In the *Statesman* Plato admits that, although there is a correct science of government, like geometry, it cannot be realized, and he stresses the need for the rule of law, since no man can be trusted with unbridled power. He then examines which of the current forms of government is the least difficult to live with, for the ruler, after all, is an artist who has to work within the limits of his medium. In the *Laws,* purporting to be a discussion of how best to found a polis in Crete, he presents a detailed program in which a state with some 5,000 citizens is ruled by 37 curators of laws and a council of 360. But the keystone of the arch is a sinister and secret Nocturnal Council to be "the sheet anchor of the state," established in its "central fortress as guardian." Poets and musicians will be discouraged and the young subjected to a rigid, austere, and exacting education. The stark consequence of Plato's political philosophy here becomes apparent. He had, nonetheless, stated, in the dawn of European political thought, the normative principle that the state should aim at promoting the good life and social harmony and that the rule of law, in the absence of the rule of philosopher-kings, is essential to this purpose.

Aristotle. Aristotle, who was a pupil in the Academy of Plato, remarks that "all the writings of Plato are original: they show ingenuity, novelty of view and a spirit of enquiry. But perfection in everything is perhaps a difficult thing." Aristotle was a scientist, not a prophet, and his *Politics* (*c.* 335–322 BC), written while he was teaching at the Lyceum at Athens, is only part of an encyclopaedic account of nature and society, in which he analyzes society as if he were a doctor and prescribes remedies for its ills. Political behaviour is here regarded as a branch of biology, as well as of ethics; in contrast to Plato, Aristotle was an empirical political philosopher. He criticizes many of Plato's ideas as impracticable, but, like Plato, he admires balance and moderation and aims at a harmonious city under the rule of law. The book is composed of lecture notes and is arranged in a confusing way—a quarry of arguments and definitions of great value but hard to master. The first book, though probably the last written, is a general introduction; Books II, III, and VII–VIII, probably the earliest, deal with the ideal state; and Books IV–VII analyze actual states and politics. The treatise is thus, in modern terms, a mixture of political philosophy and political science.

Like Plato, Aristotle naturally thinks in terms of the city-state, which he regards as the natural form of civilized life, social and political, and the best medium in which men's capacities can be realized. Hence his famous definition of man as a "political animal," distinguished from the other animals by his gift of speech and power of moral judgment. "Man, when perfected", he writes,

> is the best of animals, but when separated from law and justice he is the worst of all, since armed injustice is the most dangerous, and he is equipped at birth with the arms of intelligence and wit, moral qualities which he may use for the worst ends.

Since all nature is pervaded by purpose and since men "aim at the good," the city-state, which is the highest form of human community, aims at the highest good. Like sailors with their separate functions, who yet have a common object in safety in navigation, citizens, too, have a common aim—in modern terms survival, security, and the enhancement of the quality of life. In the context of the city-state, this high quality of life can be realized only by a minority, and Aristotle, like Plato, excludes those who are not full citizens or who are slaves; indeed, he says that some men are "slaves by nature" and deserve their status. Plato and Aristotle aim at an aristocratic and exacting way of life, reflecting, in more sophisticated forms, the ideas of the warrior aristocracies depicted by Homer.

Aristotle's view of the role of law in the city-state

Having stated that the aim of the city-state is to promote the good life, Aristotle insists that it can be achieved only under the rule of law.

> The rule of law is preferable to that of a single citizen; if it be the better course to have individuals ruling, they should be made law guardians or ministers of the laws.

The rule of law is better than that even of the best men, for "he who bids law rule may be deemed to bid God and reason alone rule, but he who bids men rule adds the element of the beast; for desire is a wild beast, and passion perverts the minds of rulers, even if they are the best of men." This doctrine, which distinguishes between lawful government and tyranny, survived the Middle Ages and, by subjecting the ruler to law, became the theoretical sanction of modern constitutional government.

Aristotle also vindicates the rule of custom and justifies the obligations accepted by members of society: the solitary man, he writes, "is either a beast or a God." This outlook at once reflects the respect for custom and solidarity that have promoted survival, in primitive tribal societies, even at the price of sacrificing individuals, and gives a theoretical justification for the acceptance of political obligation.

Like Plato, Aristotle analyzes the different kinds of city-states. While states are bound, like animals, to be different, he considers a balanced "mixed" constitution the best—it reflects the ideal of justice (*dikē*) and fair dealing, which gives every man his due in a conservative social order in which citizens of the middle condition preponderate. And he attacks oligarchy, democracy, and tyranny. Under democracy, he argues, demagogues attain power by bribing the electorate and waste accumulated wealth. But it is tyranny that Aristotle most detests; the arbitrary power of an individual above the law who is "responsible to no-one and who governs all alike with a view to his own advantage and not of his subjects, and therefore against their will. No free man can endure such a government."

The *Politics* contains not only a firm statement of these principles but also a penetrating analysis of how city-states are governed, as well as of the causes of revolutions, in which "inferiors revolt in order that they may be equal, and equals that they may be superior." The treatise concludes with an elaborate plan for educating the citizens to attain the "mean," the "possible," and the "becoming." The first implies a balanced development of body and mind, ability and imagination; the second, the recognition of the limits of mind and the range of limitations of talent; the third, an outcome of the other two, is the style and self-assurance that come from the resulting self-control and confidence.

While, therefore, Aristotle accepts a conservative and hierarchic social order, he states firmly that public power should aim at promoting the good life and that only through the rule of law and justice can the good life be attained. These principles were novel in the context of his time, when the great extra-European civilizations were ruled, justly or unjustly, by the arbitrary power of semidivine rulers and when other peoples, though respecting tribal custom and the authority of tribal elders, were increasingly organized under war leaders for depredation.

Cicero and the Stoics. Both Plato and Aristotle had thought in terms of the city-state. But Aristotle's pupil Alexander the Great swamped the cities of old Greece and brought them into a vast empire that included Egypt, Iran, and the Levant. Though the civilization of antiquity remained concentrated in city-states, they became part of an imperial power that broke up into kingdoms under Alexander's successors. This imperial power was reasserted on an even greater scale by Rome, whose empire at its greatest extent reached from central Scotland to the Euphrates and from Spain to eastern Anatolia. Civilization itself became identified with empire, and the development of eastern and western Europe was conditioned by it.

Emergence of universal philosophies

Since the city-state was no longer self-sufficient, universal philosophies developed that gave men something to live by in a wider world. Of these philosophies, Stoicism and Epicureanism were the most influential. The former inspired a rather grim self-sufficiency and sense of duty, as exemplified by the writings of the Roman emperor Marcus Aurelius; the latter, a prudent withdrawal from the world of affairs.

The setting for political philosophy thus became much wider, relating individuals to universal empire, thought of, as in China, as coterminous with the civilization itself. Its inspiration remained Hellenic; but derivative Roman philosophers reinterpreted it, and Roman legists enclosed the old concepts of political justice in a carapace of legal definitions, capable of surviving their civilization's decline.

Cicero lived in a time of political confusion during which the old institutions of the republic were breaking down before military dictators. His *De re publica* and *Laws* are both dialogues and reflect the classical sense of purpose: "to make human life better by our thought and effort." Cicero defined the *res publica* (commonwealth) as an association held together by law; he further asserted, as Plato had maintained with his doctrine of Forms manifest in the just city, that government was sanctioned by a universal natural law that reflected the cosmic order. Cicero expresses the pre-Christian Stoic attempt to moralize public power, apparent in the exacting sense of public responsibility shown by Hadrian and Marcus Aurelius in the 2nd century AD.

St. Augustine. With the conversion of the emperor Constantine (AD 312), when Christianity, long influential, became the predominant creed of the empire, and, under Theodosius (379–395), the sole official religion, political philosophy changed profoundly. St. Augustine's *City of God* (413–426), written when the empire was under attack by barbarians within and without, sums up and defines a new division between church and state and a conflict between "matter" and "spirit" resulting from the Fall of man and original sin.

St. Augustine, whose *Confessiones* are a record of a new sort of introspection, combined a classical and Hebraic dualism. From the Stoics and Virgil he inherited an austere sense of duty, from Plato and the Neoplatonists a contempt for the illusions of appetite, and from the Pauline and patristic interpretation of Christianity a sense of the conflict between Light and Darkness that reflects Zoroastrian and Manichaean doctrines emanating from Iran. In this context worldly interests and government itself are dwarfed by the importance of attaining salvation and of escaping from an astrologically determined fate and from the demons who embody the darkness. Life becomes illuminated for the elect minority by the prospect of eternal salvation or, for those without grace, shrivels under the glare of eternal fires.

St. Augustine regarded salvation as predestinate and the cosmic process as designed to "gather" an elect to fill the places of the fallen angels and so "preserve and perhaps augment the number of the heavenly inhabitants." The role of government and indeed of society itself becomes subordinated into a "secular arm," part of an earthly city, as opposed to the "City of God." The function of government is to keep order in a world intrinsically evil.

Augustine's view of the role of government

Since Christianity had long played the main role in defense of the veneer of a precarious urban civilization in antiquity, this claim is not surprising. Constantine came of crude Balkan origins, a soldier putting to rights a breakdown in government that would continue in the West with the abdication of the last Western emperor in 476, though in the East the empire would carry on with great wealth and power, centred in the new capital of Constantinople (Byzantium).

St. Augustine thus no longer assumed, as did Plato and Aristotle, that a harmonious and self-sufficient good life could be achieved within a properly organized city-state; he projected his political philosophy into a cosmic and lurid drama working out to a predestinate end. The normal interests and amenities of life became insignificant or disgusting, and the Christian Church alone exercised a spiritual authority that could sanction government. This outlook, reinforced by other patristic writings, would long dominate medieval thought, for with the decline of civilization in the West the church became more completely the repository of learning and of the remnants of the old civilized life.

THE MIDDLE AGES

The decline of ancient civilization in the West was severe; not, indeed, in technology, for the horse collar, the stirrup, and the heavy plow now came in; but political philosophy, like other intellectual interests, became elementary. In the Byzantine Empire, on the other hand, Justinian's lawyers in 529–533 produced the Codex Justinianus, the Digest, and the *Institutes,* which defined and condensed Roman law. The Byzantine *basileus,* or autocrat, had moral responsibility for guarding and harmonizing an elaborate state, a "colony" of heaven in which reason and not mere will ought to rule. And this autocracy and the orthodox form of Christianity were inherited by the Christianized rulers of the Balkans, of Kievan Russia, and of Muscovy.

In the West, two essential principles of Hellenic and Christian political philosophy were transmitted, if only in elementary definitions, in rudimentary encyclopaedias. Isidore of Seville in his 7th-century *Etymologiae,* for example, asserts that kings rule only on condition of doing right and that the rule reflects a Ciceronic law of nature "common to all people and mankind everywhere by natural instinct." Further, the barbarians respected the civilization they took over and exploited. When converted, they revered the papacy, and in 800 the Frankish Charlemagne even revived the Western Empire as holy and Roman. The idea of Christian empire coterminous with civilization thus survived in Western as well as Eastern Christendom.

John of Salisbury. After Augustine, not until the 12th-century Renaissance did another full-length speculative work of political philosophy appear in the West. The *Policraticus* of John of Salisbury (*c.* 1159) is the work of a man experienced in politics who became bishop of Chartres. Based on wide classical reading, it centres on

the ideal ruler, who represents a "public power." John admired Augustus and the Roman emperor Trajan, and, in a still predominantly feudal world, his book carried on the Roman tradition of centralized authority, though without its Byzantine autocracy. The prince, he insists, is he who rules in accordance with law, while a tyrant is one who oppresses the people by irresponsible power. This distinction, which derives from the Greeks, Cicero, and St. Augustine, is fundamental to Western concepts of liberty and the trusteeship of power.

John did not know Aristotle's *Politics,* but his learning is nevertheless remarkable, even if his political similes are unsophisticated. His favourite metaphor for the body politic is the human body: the place of the head is filled by the prince, who is subject only to God; the place of the heart by the senate; the eyes, ears, and tongue are the judges, provincial governors, and soldiers; and the officials are the hands. The tax gatherers are the intestines and ought not to retain their accumulations too long; and the farmers and peasants are the feet. John also compares a commonwealth to a hive and even to a centipede.

This vision of a centralized government, more appropriate to the memory of the Roman Empire than to a medieval monarchy operating in a still semifeudal world, is a landmark of the 12th-century revival of speculative thought and reflects the better organized monarchy that Henry II was then building up.

Aquinas. It is a far cry from this practical 12th-century treatise by a man of affairs to the elaborate justification of Christian kingship and natural law created by St. Thomas Aquinas in the 13th century, during the climax of medieval Western civilization. His political philosophy is only part of a metaphysical construction of Aristotelean range—for Aristotle had now been assimilated from Arabic sources and given a new Christian content, with the added universality of the Stoic and Augustinian world outlook. Aquinas' *Summa theologiae* purports to answer all the major questions of existence, including those of political philosophy. Like Aristotle, Aquinas thinks in terms of an ethical purpose. Natural law is discussed in the first part of the second book as part of the discussion of original sin and what would now be termed psychology, while war comes under the second part of the second book as an aspect of virtue and vice. Law is defined as "that which is regulation and measure." It is designed to promote the "felicity and beatitude" that are the ends of human life. Aquinas agrees with Aristotle that "the city is the perfection of community" and that the purpose of public power should be to promote the common good. The only legitimate power is from the community, which is the sole medium of man's well being. In his *De regimine* he compares society to a ship in need of a helmsman and repeats Aristotle's definition of man as a social and political animal. Again following Aristotle, he considers oligarchy unjust and democracy evil. Rulers should aim to make the "life of the multitude good in accordance with the purpose of life which is heavenly happiness." They should also create peace, conserve life, and preserve the state—a threefold responsibility. Here is a complete program for a hierarchical society within a cosmic order. It combines the Hellenic sense of purpose with Christian aims and asserts that, under God, power resides in the community, embodied in the ruler but only for so long as he does right. Hence the comment that "St. Thomas Aquinas was the first Whig"—a pioneer of the theory of constitutional government. The society he envisages, however, is medieval, static, hierarchical, conservative, and based on limited agriculture and even more limited technology. Nonetheless, Thomism remains the most complete and lasting political doctrine of the Catholic Church, since modified and adapted but not in principle superseded.

Aquinas' political views in the *Summa theologiae*

Dante. By the early 14th century the great European institutions, empire and papacy, were breaking down through mutual conflict and the emergence of national realms. But this conflict gave rise to the most complete political theory of universal and secular empire formulated in the medieval West. Dante's *De monarchia* (*c.* 1313), still in principle highly relevant, insists that only through universal peace can human faculties come to their full compass. But only "temporal Monarchy" can achieve this: "a unique princedom extending over all persons in time." The aim of civilization is to actualize human potentalities, and to achieve that "fullness of life which comes from the fulfillment of our being."

Monarchy, Dante argues, is necessary as a means to this end. The imperial authority of the Holy Roman emperor, moreover, comes direct from God and not through the pope. The empire is the direct heir of the Roman Empire, a legitimate authority, or Christ would not have chosen to be born under it. In subjecting the world to itself, the Roman Empire had contemplated the public good.

This high-flown argument, part of the political warfare between the partisans of the emperor and pope that was then affecting Italy, drives to essentials: that world peace can be secure only under a world authority. That Dante's argument was impractical did not concern this medieval genius, who was writing more the epitaph than the prospectus of the Holy Roman Empire; he was concerned, like St. Thomas, to create a political philosophy with a clear-cut aim and a universal view.

Out of the grand but impractical visions of the High Middle Ages in the 13th-century climax of Christian civilization there emerged by early modern times the idea of a well-governed realm, its authority derived from the community itself, with a program designed to ensure the solvency and administrative efficiency of a secular state. In spite of the decline of the civilization of antiquity in the West, the Greco-Roman sense of purpose, of the rule of law, and of the responsibility of power survived in Christian form.

THE 16TH TO THE 18TH CENTURIES

Machiavelli. In the thought of the Italian political philosopher Niccolò Machiavelli may be seen a complete secularization of political philosophy. Machiavelli was an experienced diplomat and administrator, and, since he stated flatly how the power struggle was conducted in Renaissance Italy, he won a shocking reputation. He was not, however, without idealism about the old Roman republic, and he admired the independent spirit of the German and Swiss cities. This idealism made him all the more disgusted with Italian politics, of which he makes a disillusioned and objective analysis. Writing in retirement after political disgrace, Machiavelli states firmly that,

Machiavelli's view of human nature

> Since this is to be asserted in general of men, that they are ungrateful, fickle, false, cowards, covetous, and as long as you succeed they are yours entirely: they will offer you their blood, property, life, and children . . . when the need is far distant; but when it approaches they turn against you.

And again,

> since the desires of men are insatiable, nature prompting them to desire all things and fortune permitting them to enjoy but few, there results a constant discontent in their minds, and a loathing of what they possess.

This view of human nature, already expressed by Plato and St. Augustine, is here unredeemed by Plato's doctrine of form and illusion or by St. Augustine's dogma of salvation through grace. Machiavelli accepts the facts and advises the ruler to act accordingly. The prince, he states, must combine the strength of the lion with the cunning of the fox: he must always be vigilant, ruthless, and prompt, striking down or neutralizing his adversaries without warning. And when he does an injury it must be total. For "men ought to be either well treated or crushed, because they can avenge themselves of lighter injuries, of more serious ones they cannot." Moreover, "irresolute princes who follow a neutral path are generally ruined." He advises that it is best to come down at the right moment on the winning side and that conquered cities ought to be either governed directly by the tyrant himself residing there or destroyed. Princes, furthermore, unlike private men, need not keep faith: since politics reflects the law of the jungle, the state is a law unto itself, and normal moral rules do not apply to it.

Machiavelli had stated with unblinking realism how, in fact, tyrants behave; and, far from criticizing their conduct or distinguishing between the just prince who rules

by law and the tyrant whose laws are in his own breast, he considers that the successful ruler has to be beyond morality since the safety of and expansion of the state are the supreme objective. In this myopic view, the cosmic visions of Aquinas and Dante are disregarded, and politics becomes a fight for survival. Within his terms of reference, Machiavelli made a convincing case, although as an experienced diplomat he might have realized that dependability in fact pays and that systematic deceit, treachery, and violence usually bring about their own nemesis.

Hobbes. The 17th-century English political philosopher Thomas Hobbes, who spent his life as a tutor and companion to great noblemen, was a writer of genius with a greater power of phrase than any other English political philosopher. He was not, as he is sometimes misrepresented, a prophet of "bourgeois" individualism, advocating free competition in a capitalistic free market. On the contrary, he was writing in a preindustrial, if increasingly commercial, society and did not much admire wealth as such but rather "honours." He was socially conservative and anxious to give a new philosophical sanction to a hierarchical, if businesslike, commonwealth in which family authority was most important.

Philosophical influences on Hobbes

Philosophically, Hobbes was influenced by nominalist scholastic philosophy, which had discarded Thomist metaphysics and had accepted a strict limitation of mind. He therefore based his conclusions on the rudimentary mathematical physics and psychology of his day and aimed at practical objectives—order and stability. He believed that the fundamental physical law of life was motion and that the predominant human impulses were fear and, among those above the poverty level, pride and vanity. Men, Hobbes argued, are strictly conditioned and limited by these laws, and he tried to create a science of politics that would reflect them. "The skill of making, and maintaining Common-wealths," therefore, "consisteth in certain Rules, as doth Arithmetique and Geometry; not (as Tennis play) on Practise onely: which Rules, neither poor men have the leisure, nor men that have had the leisure, have hitherto had the curiosity, or the method to find out."

Hobbes ignores the classical and Thomist concepts of a transcendent law of nature, itself reflecting divine law, and of a "chain of being" whereby the universe is held harmoniously together and, following Descartes's practical method of investigation, states plainly that power creates law, not law power. For law is only law if it can be enforced, and the price of security is one supreme sovereign public power. For, without it, such is the competitive nature of men, once more than subsistence has been achieved they are actuated by vanity and ambition, and there is a war of all against all. The true law of nature is self-preservation, he argues, which can be achieved only if the citizens make a compact among themselves to transfer their individual power to the "leviathan" (ruler), who alone can preserve them in security. Such a commonwealth has no intrinsic supernatural or moral sanction: it derives its original authority from the people and can command loyalty only so long as it succeeds in keeping the peace. He thus uses both the old concepts of natural law and contract, often invoked to justify resistance to authority, as a sanction for it.

Hobbes, like Machiavelli, starts from an assumption of basic human folly, competitiveness, and depravity, and contradicts Aristotle's assumption that man is by nature a "political animal." On the contrary, he is naturally antisocial; and, even when men meet for business and profit, only "a certain market-fellowship" is engendered. All society is only for gain or glory, and the only true equality among men is their power to kill each other. Hobbes sees and desires no other equality. Indeed, he specifically discouraged "men of low degree from a saucy behaviour towards their betters."

Significance of the *Leviathan*

The *Leviathan* horrified most of his contemporaries; Hobbes was accused of atheism and of "maligning the Human Nature." But, if his remedies were tactically impractical, in political philosophy he had gone very deep by providing the sovereign nation-state with a pragmatic justification and directing it to utilitarian ends.

Spinoza. The 17th-century Dutch philosopher Benedict de Spinoza also tried to make a scientific political theory, but it was more humane and more modern. Hobbes assumes a preindustrial and economically conservative society, but Spinoza, a Portuguese Jew born in Amsterdam, assumes a more urban setting. Like Hobbes, he is Cartesian, aiming at a scientific basis for political philosophy; but, whereas Hobbes was dogmatic and authoritarian, Spinoza desired toleration and intellectual liberty, by which alone human life achieves its highest quality. Spinoza, reacting against the ideological wars of religion and skeptical of both metaphysics and religious dogma, was a scientific humanist who justified political power solely by its usefulness. If state power breaks down and can no longer protect him or if it turns against him, frustrates, or ruins his life, then any man is justified in resisting it, since it no longer fulfills its purpose. It has no intrinsic divine or metaphysical authority.

In *Tractatus Theologico-Politicus* and the *Tractatus Politicus* Spinoza develops this theme. He intends, he writes, "not to laugh at men or weep over them or hate them, but to understand them." In contrast to St. Augustine, he glorifies life and holds that governments should not try to "change men from rational beings into beasts or puppets, but enable them to develop their minds and bodies in security and to employ their reason unshackled." The more life is enjoyed, he declares, the more the individual participates in the divine nature. God is immanent in the entire process of nature, in which all creatures follow the laws of their own being to the limit of their powers. All are bound by their own consciousness, and man creates his own values.

It seems that Spinoza thought good government approximated to that of the free burgesses of Amsterdam, a city in which religious toleration and relative political liberty had been realized. He is thus a pioneer of a scientific humanist view of government and of the neutrality of the state in matters of belief.

Richard Hooker's adapted Thomism. While out of the breakup of the medieval social order there emerged the humanist but sceptical outlook of Machiavelli, then the scientific humanist principles of Descartes, Hobbes, and Spinoza, from which the utilitarian and pragmatic outlook of modern times derives, another influential and politically important strain of political philosophy also emerged. During the Reformation and Counter-Reformation, Protestant and Catholic dogmatists denounced each other and even attacked the authority of princes who, from interest or conviction, supported one side or the other. Political assassination became endemic, for both Protestant and Catholic divines declared that it was legitimate to kill an heretical ruler. Appeal was made to rival religious authority as well as to conscience. Men would resist authority and suffer execution rather than risk damnation, and in the resulting welter Hobbes and Spinoza advocated a sovereign state as the remedy. But other political philosophers salvaged the old Thomist concept of a divine cosmic order and of natural and human laws sanctioning the state. They also put forth the classical and medieval idea of the derivation of public power from the commonwealth as a whole and the responsibility of princes to the law. When Hobbes wrote that might makes right, he outraged such critics, who continued to assert that public power was responsible to God and the laws and that it was right to resist a tyrant who declared that the laws were in his own breast. This political theory was most influentially developed in England, where it inspired the constitutionalism that would also predominate in the United States.

Richard Hooker, an Anglican divine who wrote *The Lawes of Ecclesiasticall Poline* (1593–97), reconciled Thomist doctrines of transcendent and natural law, binding on all men, with the authority of the Elizabethan Anglican Church, which he defended against the Puritan appeal to conscience. Society, he argued, is itself the fulfillment of natural law, of which human and positive law are reflections, adapted to society. And public power is not something personal, for it derives from the community under law. Thus,

> The lawful power of making laws to command whole politic societies of men belongeth so properly unto the same entire societies, that for any prince . . . to exercise the same of himself . . . is no better than mere tyranny.

Such power can derive either directly from God or else from the people. The prince is responsible to God and the community; he is not, like Hobbes's ruler, a law unto himself. Law makes the king, not the king law.

Hooker, indeed, insisted that "the prince has a delegated power, from the Parliament of England, together with the convocation (of clergy) annexed thereto . . . whereupon the very essence of all government doth depend." This is the power of the crown in parliament in a balanced constitution. Hence an idea of harmonious government by consent. The Thomist medieval universal harmony had been adapted to the nation-state.

Locke. It was John Locke, politically the most influential English philosopher, who further developed this doctrine. His *Two Treatises of Government* (1690) were written to justify the Glorious Revolution of 1688–89, and his *Letter Concerning Toleration* (1689) was written with a plain and easy urbanity, in contrast to the baroque eloquence of Hobbes. Locke was a scholar, physician, and man of affairs, well-experienced in politics and business. As a philosopher he accepted strict limitations for mind, and his political philosophy is moderate and sensible, aimed at a balance among executive, judicial, and legislative powers, although with a bias toward the last.

His first *Treatise* was devoted to confuting the Royalist doctrine of patriarchal divine right by descent from Adam, an argument then taken very seriously and reflecting the idea of government as an aspect of a divinely ordained chain of being. If this order were broken, chaos would come about. The argument was part of the contemporary conflict of the ancients and the moderns.

Locke's view of the role and limits of public power

Locke tried to provide an answer by defining a limited purpose for political power, which purpose he considered to be "a right of making laws with penalties of death, and consequently all less penalties, for the regulating and preserving of property, and of employing the force of the community in execution of such laws, and in the defense of the commonwealth from foreign injury, and all this only for the public good." The authority of government derives from a contract between the rulers and the people, and the contract binds both parties. It is thus a limited power, proceeding according to established laws and "directed to no other end but the peace, safety, and public good of the people."

Whatever its form, government, to be legitimate, must govern by "declared and reasoned laws," and, since every man has a "property" in his own person and has "mixed his labour" with what he owns, government has no right to take it from him without his consent. It was the threat of attack on the laws, property, and the Protestant religion that had roused resistance to James II. Locke is expressing the concerns and interests of the landed and moneyed men by whose consent James's successor, William III, came to the throne, and his commonwealth is strictly conservative, limiting the franchise and the preponderant power to the propertied classes. Locke was thus no democrat in the modern sense and was much concerned to make the poor work harder. Like Hooker, he assumes a conservative social hierarchy with a relatively weak executive power and defends the propertied classes both against a ruler by divine right and against radicals. In advocating toleration in religion he was more liberal: freedom of conscience, like property, he argued, is a natural right of all men. Within the possibilities of the time, Locke thus advocated a constitutional mixed government, limited by parliamentary control of the armed forces and of supply. Designed mainly to protect the rights of property, it was deprived of the right of arbitrary taxation or imprisonment without trial and was in theory responsible to all the people through the politically conscious minority who were thought to represent them.

Though he was socially conservative, Locke's writings are very important in the rise of liberal political philosophy. He vindicates the responsibility of government to the governed, the rule of law through impartial judges, and the toleration of religious and speculative opinion. He is an enemy of the totalitarian state, drawing on medieval arguments and deploying them in practical, modern terms.

Burke. The Irishman Edmund Burke, while elaborating Whig constitutional doctrine expressed with such common sense by Locke, wrote with more emotion and took more account of time and tradition. While reiterating that government is responsible to the governed and distinguishing between a political society and a mere mob, he thought that governments were trustees for previous generations and for posterity. He made the predominant political philosophy of the 18th-century establishment appear more attractive and moral, but he wrote no great single work of political philosophy, expressing himself instead in numerous pamphlets and speeches.

In his early *Vindication of Natural Society* Burke is critical of the sufferings imposed by government, but his "Thoughts on the Cause of the Present Discontents" defines and defends the principles of the Whig establishment. He invoked a transcendent morality to sanction a constitutional commonwealth, but he detested abstract political theories in whose name men are likely to vivisect society. He set great store by ordered liberty and denounced the arbitrary power of the Jacobins who had captured the French Revolution. In his *Reflections on the Revolution in France* (1790) and *An Appeal from the New, to the Old Whigs* (1791), he discerned in the doctrine of sovereignty of the people, in whose name the revolutionaries were destroying the old order, another and worse form of arbitrary power. No one generation has the right to destroy the agreed and inherited fabric of society, and "Neither the few nor the many have the right to govern by their will". A country is not a mere physical locality, he argued, but a community in time into which men are born, and only within the existing constitution and by the consent of its representatives can changes legitimately be made. Once the frame of society is smashed and its law violated, the people become a "mere multitude told by the head," at the mercy of any dictator who can seize power. He was realistic in predicting the consequences of violent revolution, which usually ends up in some kind of dictatorship. Burke, in sophisticated accents, spoke for the ancient and worldwide rule of custom and conservatism and supplied a needed romanticism to the calculating good sense of Locke.

Burke's concept of an obligation to posterity

Vico. The political philosophies hitherto surveyed contained little idea of progress. In antiquity the idea of cyclic recurrence predominated, and even 18th-century Christians believed that the world had been created in 4004 BC and would end in the Second Coming of Christ and a judgment. The 14th-century Arab philosopher of history Ibn Khaldūn of Tunis, in the *Muqaddimah* to his *Kitāb al-ʿibar*, had pioneered a vast sociological view of the historical process; but in western Europe it was a neglected Neapolitan philosopher, Giambattista Vico (1668–1744), who first interpreted the past in terms of the changing consciousness of mankind. His *Scienza nuova* (1725; revised edition 1744) interpreted history as an organic process involving language, literature, and religion and attempted to reveal the mentality or ethos of earlier ages: the age of the gods, the heroic age, and the human age, its climax and decadence. These ages recur, and each is distinguished by mythology, heroic poetry, and rational speculation respectively. In contrast to the legalistic, contractual, and static political philosophies then prevalent, Vico had discerned new horizons.

Montesquieu. This sort of vision was developed and elegantly popularized by the cosmopolitan French savant Charles-Louis de Secondat, Baron de Montesquieu, whose work *The Spirit of the Laws* (Eng. trans., 1750) won immense influence. It was an ambitious treatise on human institutions and a pioneer work of anthropology and sociology. Believing in an ordered universe—for "how could blind fate have produced intelligent beings?" —Montesquieu examined the varieties of natural law, varying customs, laws, and civilizations in different

environments. He made the pedestrian good sense of Locke seem provincial, although he admired him and the British constitution. Unfortunately, he overemphasized the separation of executive, judicial, and legislative powers, considerable in Locke's day but by his own time tending to be concentrated in the sovereignty of Parliament. This doctrine much influenced the founders of the United States and the early French Revolutionaries.

Rousseau. The revolutionary romanticism of the Swiss-French philosopher Jean-Jacques Rousseau may be interpreted in part as a reaction to the analytic rationalism of the Enlightenment. He was trying to escape the aridity of a purely empirical and utilitarian outlook and attempting to create a substitute for revealed religion. Rousseau's *Émile* (1762) and *Du contrat social* (1763) proved revolutionary documents, and his posthumous *Considérations sur le gouvernement de Pologne* ("Considerations on the Government of Poland") contains desultory but often valuable reflections on specific problems.

There had been radical political slogans coined in medieval peasant revolts and in the 17th century, as in the Putney debates (1647) in the Cromwellian army, when a Puritan officer declared that "the poorest hee that is in England hath a life to live as the greatest hee," but the inspiration of these movements had been religion. Now Rousseau proclaimed a secular egalitarianism and a romantic cult of the common man. His famous sentence, "man is born free, but he is everywhere in chains," called into question the traditional social hierarchy: hitherto, political philosophers had thought in terms of elites, but now the mass of the people had found a champion and were becoming politically conscious.

Rousseau's idea of the general will

Rousseau was a romantic, given to weeping under the willows on Lake Geneva, and the *Social Contract and Discourses* is hypnotically readable, a flaming protest by one who found the hard rationality of the 18th century too exacting. But man is not, as Rousseau claims, born free. Man is born into society, which imposes restraints on him. Casting about to reconcile his artificial antithesis between man's purported natural state of freedom and his condition in society, Rousseau utilizes the old theories of contract and transforms them into the concept of the "general will." This general will, a moral will that aims at the common good and in which all participate directly, reconciles the individual and the community by representing the will of the community as deriving from the will of moral individuals, so that to obey the laws of such a community is in a sense to follow one's own will, assuming that one is a moral individual.

Similar ideas to that of the general will became accepted as a basis for both the social-democratic welfare state and for totalitarian dictatorships. And, since the idea was misapplied from small village or civic communities to great sovereign nation-states, Rousseau was also a prophet of a nationalism that he never advocated. Rousseau himself wanted a federal Europe. He never wrote the proposed sequel to the *Du contrat social* in which he meant to deal with international politics, but he declared that existing governments lived in a state of nature, that their obsession with conquest was imbecilic, and that "if we could realize a European republic for one day, it would be enough to make it last for ever" (*Political Writings* I, pp. 365–388). But, with a flash of realism, he thinks the project impracticable, owing to the folly of men.

The incursion of this revolutionary romantic into political philosophy changed the climate of political opinion, for it coincided with the breakdown of the old dynastic order and the emergence first of the middle classes and then of the masses to political consciousness and power.

That the concept of general will was vague only increased its adaptability and prestige: it would both make constitutionalism more liberal and dynamic and give demagogues and dictators the excuse for "forcing people to be free" (that is, forcing people to follow the general will, as interpreted by the ruling forces). Rousseau could inspire liberals, such as the 19th-century English philosopher T.H. Green, to a creative view of a state helping people to make the best of their potential through a variety of free institutions. It could also play into the hands of demagogues claiming to represent the general will and bent on molding society according to their own abstractions.

THE 19TH CENTURY

Utilitarianism. A major force in the political and social thought of the 19th century was Utilitarianism, the doctrine that the actions of governments should be judged simply by the extent to which they promoted the "greatest happiness of the greatest number." The founder of the Utilitarian school was Jeremy Bentham, an eccentric Englishman trained in the law.

Bentham judged all laws and institutions by their utility thus defined. "The Fabric of Felicity," he wrote, "must be reared by the hands of reason and Law."

Bentham's *Fragment, on government* (1776) and *Introduction to the Principles of Morals and Legislation* (1789) elaborated a Utilitarian political philosophy. Bentham was an atheist and an exponent of the new laissez-faire economics of Adam Smith and David Ricardo, but he inspired the spate of legislation that, after the Reform Bill of 1832, had tackled the worst consequences of 18th-century inefficiency and of the Industrial Revolution. His influence, moreover, spread widely abroad. At first a simple reformer of law, Bentham attacked notions of contract and natural law as superfluous, "The indestructible prerogatives of mankind," he wrote, "have no need to be supported upon the sandy foundation of a fiction." The justification of government is pragmatic, its aim improvement and to release the free choice of individuals and the play of market forces that will create prosperity. Bentham thought men far more reasonable and calculating than they are and brushed aside all the Christian and humanist ideas rationalizing instinctive loyalty and awe. He thought society could advance by calculation of pleasure and pain, and his *Introduction* even tries to work out "the value of a lot of pleasure and pain, how now to be measured." He compared the relative gratifications of health, wealth, power, friendship, and benevolence, as well as those of "irascible appetite" and "antipathy." He also thought of punishment purely as a deterrent, not as retribution, and graded offenses on the harm they did to happiness, not on how much they offended God or tradition.

Bentham's calculation of pain and pleasure

If Bentham's psychology was naïve, that of his disciple James Mill was philistine. Mill postulated an economic man whose decisions, if freely taken, would always be in his own interest, and he believed that universal suffrage, along with Utilitarian legislation by a sovereign parliament, would produce the kind of happiness and well-being that Bentham desired. In his *Essay on Government* (1828) Mill thus shows a doctrinaire faith in a literate electorate as the means to good government and in laissez-faire economics as a means to social harmony.

This Utilitarian tradition was humanized by James Mill's son, John Stuart Mill, the most influential of mid-Victorian liberals.

Whereas James Mill had been entirely pragmatic, his son tried to enhance more sophisticated values. He thought that civilization depended on a tiny minority of creative minds and on the free play of speculative intelligence. He detested conventional public opinion and feared that complete democracy, far from emancipating opinion, would make it more restrictive. Amid the dogmatic and strident voices of mid-19th-century nationalists, utopians, and revolutionaries, the quiet, if sometimes priggish, voice of mid-Victorian liberalism proved extremely influential in the ruling circles of Victorian England.

Accepting democracy as inevitable, J.S. Mill expressed the still optimistic and progressive views of an intellectual elite. Without complete liberty of opinion, he insisted, civilizations ossify. The quality of progress results not merely from the blind forces of economic competition but from the free play of mind. The worth of the state in the long run is only the worth of the individual's composing it, and without men of genius society would become a "stagnant pool." This militant humanist, unlike his father,

was aware of the dangers of even benevolent bureaucratic power and declared that a state that "dwarfs its men" is culturally insignificant.

Mill also advocated the legal and social emancipation of women, holding that ability was wasted by mid-Victorian conventions. He believed that the masses could be educated into accepting the values of liberal civilization, but he defended private property and was as wary of rapid extensions of the franchise as of bureaucratic power.

Tocqueville. Mill's friend Alexis de Tocqueville, whose *De la démocratie en Amérique* (*Democracy in America*) appeared in 1835–40, was a French civil servant also concerned with maintaining the standards and creativeness of civilization in face of the rising tide of mass democracy. Since the United States was then the only large-scale democracy extant, Tocqueville decided to go there, and as a result of his visit wrote a classic account of early 19th-century American civilization. "We cannot," he wrote, "prevent the conditions of men from becoming equal, but it depends upon ourselves whether the principle of equality will lead them to servitude or freedom, to knowledge or barbarism, to prosperity or wretchedness." He feared the possible abuse of power by centralized government, unrestrained by the power of the old privileged classes, and thought it essential to "educate democracy" so that, although it would never have the "wild virtues" of the old regimes, it would have its own dignity, good sense, and even benevolence. Tocqueville greatly admired American representative institutions and made a penetrating analysis of the new power of the press. He realized, as few people then did, that the United States and Russia would become world powers, and he contrasted the freedom of the one and the despotism of the other. He also foresaw that under democracy education would be respected more as a ladder to success than for its intrinsic content and might thus become mediocre. He was alive to the dangers of uniform mediocrity but believed, like Mill, that democracy could be permeated by creative ideas.

T.H. Green. This kind of humanism was given a more elaborate philosophical content by the English philosopher T.H. Green, whose *Lectures on the Principles of Political Obligation* (1895; reprinted from *Philosophical Works*, vol. 2, 1885) greatly influenced the Liberals in the British governments of the period 1905–15. Green, like J.S. Mill and Tocqueville, wished to extend the minority culture to the people and even to use state power to "hinder hindrances to the good life." He had absorbed from Aristotle, Spinoza, Rousseau, and Hegel an organic theory of the state. The latter, by promoting the free play of spontaneous institutions, ought to help individuals both to "secure the common good of society [and] enable them to make the best of themselves."

While hostile to the abuse of landed property, Green was not a Socialist. He accepted the idea that property should be private and unequally distributed and thought the operation of the free market the best way to benefit the whole society; for free trade would, he thought, diminish the inequalities of wealth in a common prosperity. But Green would have extended the power of the state over education, health, housing and town planning, and the relief of unemployment—a new departure in Liberal thought. These recommendations are embedded in the most elaborate and close-knit intellectual construction made by any modern British political philosopher, and they laid the foundation of the British welfare state.

Green's writing on social reform

Liberal nationalism. Whereas Green shirked the extension of liberal and constitutional principles into international affairs, the Italian patriot and revolutionary prophet Giuseppe Mazzini made it his vision and became the most influential prophet of liberal nationalism. In his *The Duties of Man* and *Essays* he envisaged a harmony of free peoples—a "sisterhood of nations," in which the rule of military empires would be thrown off, the destruction of clerical and feudal privileges accomplished, and in which the emancipated peoples would be regenerated by means of education and universal suffrage. This vision inspired the more idealistic aspects of the Italian Risorgimento (national revival or resurrection) and of nationalistic revolts in Europe and beyond. Though, in fact, fervid nationalism proved destructive, Mazzini advocated a united Europe of free peoples, in which national singularities would be transcended in a pan-European harmony. This sort of liberal democratic idealism was catching, and even if it frequently inspired Machiavellian policies, it also inspired President Woodrow Wilson of the United States, who had he not been thwarted by domestic opposition, might well have made a Mazzinian type League of Nations a success. Moreover, the Europe of the Common Market owes much to the apparently impractical liberal idealism of Mazzini (see NATIONALISM).

American constitutionalism. The United States was founded by men deeply influenced by republicanism, by Locke, and by the optimism of the French Enlightenment. George Washington, John Adams, and Thomas Jefferson all concurred that laws, rather than men, should be the final sanction and that government should be responsible to the governed. But the influence of Locke and the Enlightenment was not entirely happy. John Adams, who followed Washington as president, prescribed a constitution with a balance of executive and legislative power checked by an independent judiciary. The federal constitution, moreover, could be amended only by a unanimous vote of the states. Anxious to safeguard state liberties and the rights of property, the founding fathers gave the federal government insufficient revenues and coercive powers, as a result of which the constitution was stigmatized as being "no more than a Treaty of Alliance." Yet the federal union was preserved. The civil power controlled the military, and there was religious toleration and freedom of the press and of economic enterprise. Most significantly, the concept of natural rights had found expression in the Declaration of Independence and was to influence markedly political and legal developments in the ensuing decades, as well as inspire the French Declaration of the Rights of Man.

Natural rights in American constitutional thought

Anarchism and utopianism. While a liberal political philosophy within a framework of capitalistic free trade and constitutional self-government dominated the greatest Western powers, mounting criticism developed against centralized government itself. Radical utopian and anarchist views, previously expounded mainly by religious sects, became secularized in such works as William Godwin's *Political Justice* (1793), Robert Owen's *New View of Society* (1813), and Pierre-Joseph Proudhon's voluminous and anticlerical writings (see ANARCHISM).

The English philosopher William Godwin, an extreme individualist, shared Bentham's confidence in the reasonableness of mankind. He denounced the wars accepted by most political philosophers and all centralized coercive states. The tyranny of demagogues and of "multitudes drunk with power" he regarded as being as bad as that of kings and oligarchs. The remedy, he thought, was not violent revolution, which produces tyranny, but education and freedom, including sexual freedom. His was a program of high-minded, atheistic anarchy.

The English Socialist Robert Owen, a cotton spinner who had made a fortune, also insisted that bad institutions, not original sin or intrinsic folly, caused the evils of society, and he sought to remedy them by changing the economic and educational system. He thus devised a scheme of model cooperative communities that would increase production, permit humane education, and release the naturally benevolent qualities of mankind.

The French moralist and advocate of social reform Pierre-Joseph Proudhon attacked the "tentacular" nation-state and aimed at a classless society in which major capitalism would be abolished. Self-governing producers, no longer slaves of bureaucrats and capitalists, would permit the realization of an intrinsic human dignity, and federation would replace the accepted condition of war between sovereign states. Proudhon tried to transform society by rousing the mass of the people to cooperative humanitarian consciousness.

Saint-Simon and Comte. Another revolt against the prevalent establishment, national and international, was made by the French social philosopher Henri de Saint-Simon. Saint-Simon wanted to develop the Industrial

Revolution so as to ameliorate the condition of the poorest class. This would be achieved not through political revolution, but through a government of bankers and administrators who would supersede kings, aristocrats, and politicians. If France were suddenly deprived of three thousand leading scientists, engineers, bankers, painters, poets, and writers, he argued, the result would be catastrophic; but if all the courtiers and bishops and 10,000 landowners vanished, the loss, though deplorable, would be much less severe. Saint-Simon also demanded a united Europe, superseding the warring nation-states, with a European parliament and a joint development of industry and communication. He also invented a synthetic religion appropriate to a scientific phase of history, with a cult of Newton and the great men of science.

Comte's "religion of humanity"

Saint-Simon's disciple Auguste Comte went further. His *Course of Positive Philosophy* (1830–42) and *System of Positive Polity* (1851–54) elaborated a "religion of humanity," with ritual, calendar, and a priesthood of scientists, and secular saints, including Julius Caesar, Dante, and Joan of Arc. Society would be ruled by bankers and technocrats and Europe united into a Western republic. This doctrine, backed by pioneering sociology, won much influence among intellectuals. Yet Comte, like Saint-Simon, tackled the essential questions: how to deploy the power of modern technology for the benefit of all mankind; how to avoid wars between sovereign states; and how to fill the void left by the waning of Christian beliefs.

Hegel. Whereas the utopian reformers had discarded metaphysical arguments, the German philosopher G.W.F. Hegel claimed to apprehend the totality of the cosmos by speculative cognition. Like Vico, he saw the past in terms of changing consciousness, but he viewed the historical process as one of "becoming" rather than as one of eternal recurrence. Hegel had no adequate historical data for his intuitions, since the whole of world history was even less known then than it is today, but his novel sweep and range of theory proved an intoxicating substitute for religion. He divided world history into four epochs: the patriarchal Eastern empire, the brilliant Greek boyhood, the severe manhood of Rome, and the Germanic phase after the Reformation. The "Absolute," like a conductor, summons each people to their finest hour, and neither individuals nor states have any rights against them during their historically determined period of supremacy. Many felt some sense of anticlimax, however, when he claimed that the Prussian state embodied the hitherto highest self-realization of the "Absolute" (see HEGELIANISM). Not since St. Augustine had so compelling a drama been adumbrated. Hegel's drama, moreover, culminates in this world, for "the state is the divine idea as it exists on earth."

Marx and Engels. Hegel was a conservative, but his influence on the revolutionaries Karl Marx and his collaborator Friedrich Engels was profound. They inherited the Hegelian claim to understand the "totality" of history and life as it progressed through a dialectic of thesis, antithesis, and synthesis. But, whereas Hegel envisaged a conflict of nation-states, Marx and Engels thought that the dynamism of history was generated by inevitable class conflict economically determined. This was an idea even more dynamic than Hegel's and more relevant to the social upheavals that were a consequence of the Industrial Revolution. Marx was a formidable prophet whose writings lead up to an apocalypse and redemption. A deeply learned humanist, his ideal was the fullest development of the human personality. But, whereas Plato was concerned with an elite, Marx cared passionately for the elevation of whole peoples.

The Marxist credo was all the more effective as it expressed with eloquent ferocity the grievances of the poor, while prophesying retribution and a happy ending. For the state, once captured by the class-conscious vanguard of the proletariat, would take over the means of production from the capitalists, and a brief "dictatorship of the proletariat" would establish a truly communist society. The state would then wither away and man at last become "fully human" in a classless society.

The powerful slogans of Marx and Engels were a natural result of the unbridled capitalism of laissez faire, but politically they were naïve. In classical, medieval, and humanistic political philosophy the essential problem is the control of power, and to imagine that a dictatorship, once established, will wither away is utopian. As even Marx's fellow revolutionary the Russian anarchist M.A. Bakunin observed,

Marx's naïveté regarding the dictatorship of the proletariat

> The revolutionary dictatorship of the doctrinaires who put science before life would differ from the established state only in external trappings. The substance of both are a tyranny of the minority over the majority—in the name of the many and the supreme wisdom of the few.

The revolutionaries would vivisect society in the name of dogmas and "destroy the present order, only to erect their own rigid dictatorship among its ruins." For a full account of Marxist philosophy, see MARXISM.

II. Political philosophy in the 20th century

Nineteenth-century European civilization had been the first to dominate and pervade the whole world and to create a new self-sustaining productivity in which all eventually might share. But, as Saint-Simon had pointed out, this civilization had a fatal flaw. The rule of law, accepted within the politically advanced states, had never been achieved among them. Heavily armed nations and empires remained in a Hobbesian "posture of war," and classical and medieval ideals of world order had long been discarded. Within states, also, laissez-faire capitalism had exacerbated class conflicts, while the decline of religious belief had undermined traditional solidarity. And in 1914, when a general European war broke out, the peoples, contrary to the hopes of cosmopolitan revolutionaries, rallied behind their national governments. When the victorious powers failed to promote world order through the League of Nations, a second global conflict followed, during which were developed weapons so destructive as to threaten life everywhere.

In the aftermath of these catastrophes and the worldwide revulsion they occasioned, not least against the European colonial powers, three mainstreams of mid-20th-century political philosophy may be discerned.

In liberal-constitutional states, with modified, managerial capitalism and various degrees of public welfare, a political pragmatism has emerged, still maintaining the Aristotelian distinction between the rule of law and government by consent, on the one hand, and tyranny on the other. Second, there has been a reaffirmation of religious or quasi-religious values appealing to conscience and the inner man, expressed most persuasively in Existentialist writings. Third, revolutionary ideas have also developed, most of them along Marxist lines. Other revolutionary doctrines appeal to anarchist traditions and are elaborated with neo-Marxist and neo-Freudian insights. Within these categories many shades of opinion are expressed, and only a sampling of representative views is presented here.

POLITICAL PRAGMATISM

The first, pragmatist approach has been most powerfully asserted in the United States and Great Britain. The American writer Lewis Mumford, for example, has advocated a militant humanism, defending people against the alienations of megalapolitan life and attacking mechanization and materialism. Like the Greek philosophers and like Tocqueville, whom he admires, Mumford declares, "In the end, all our contrivances have but one object; the continued growth of human personalities and the cultivation of the best life possible." The American philosopher and educationist John Dewey, on the other hand, sought to counteract the dehumanization of industrial mass society by a freer form of education, liberating the personality.

Both these writers criticize the existing structure of society and its modified capitalism, but try to work within it. Another humanist, the English philosopher Bertrand Russell, was more radical. Russell carried into political philosophy an artistocratic individualism, campaigning for toleration, sexual freedom, compassion, and common sense. He broadcast elite values to a mass society and

attacked materialism, crass bureaucracy, and war. He twice went to prison in pacifist protest and was obsessed with the universal menace of nuclear weapons. He denounced warlike political theories: "Remember your humanity," he said, "and forget the rest." On political tactics often inept, Russell won wide influence as a man of principle, concerned to adapt archaic institutions to the changed environment of mankind.

The Austrian-born British philosopher Karl Popper has demonstrated the pretensions of the 19th-century determinist philosophies such as those of Hegel and Marx, while an English historian and philosopher, Sir Isaiah Berlin, has ridiculed the idea of a supposedly objective march of history. Berlin also rejects the Marxist belief that all values are conditioned by the place men occupy on the "moving stair of time." Marx, he points out, was as romantic as Hegel in envisaging a "world which moves from explosion to explosion in order to fulfil the great cosmic design." Moral values, he insists, are not just a "subjective gloss unworthy of consideration on the great hard edifice of historical construction." No single formula can be found, Berlin argues, whereby the various objectives of men can be harmoniously realized. There are many human goals, which may well be in conflict with one another.

This empirical, pluralist, and liberal political philosophy has much in common with the approach of the Frenchman Émile Durkheim and the Englishman Graham Wallas, both founding fathers of modern sociology. Statesmen and political philosophers, they contend, should not play the part of prophets but rather confine themselves to investigating social patterns and the ideas that are part of them. Ways might thus be found of promoting the survival and vitality of a given society in its particular setting.

Scientific humanism of Graham Wallas

Graham Wallas was concerned to adapt constitutional societies by consent. He wanted to nationalize many essential means of production, including transport and communications, and through increased taxation strengthen social democracy by greater economic and social equality. He was not a revolutionary but a reformer, who understood the precariousness of civilization and the dangers of nationalism, which could only bring, he prophesied, centuries of warfare and regression. He advocated a worldwide and constitutionalist scientific humanism, inspired by the idea of the solidarity of the whole species, for "the master task of civilized mankind is to promote the conditions leading to the good life."

Other political sociologists who accepted the established order did not expect to improve it. The Italian Vilfredo Pareto, and Gaetano Mosca, a Sicilian-born lawyer, set themselves not to state what they wanted but to record what occurs in society. Pareto's *Mind and Society* (1916) is an elaborate, quasi-mathematical classification of nonlogical political myths. Its form is daunting, but its insights are penetrating, especially a hilarious dissection of Rousseau's *General Will,* of which, Pareto concludes, "the intrinsic logico-experimental value . . . is zero." Ranging sardonically over history, Pareto insists that elites will always manipulate society, power merely shifting from one set of rulers to another.

Mosca, in *The Ruling Class* (1939), analyzed how political myths are exploited. He also concluded that elites everywhere are bound to rule and that the least bad government occurs when abuse of power is checked by legal means; that is, by the rule of law. Mosca admired the liberal constitutionalism of the 19th century, although he was aware of its precariousness and limitations. He argued that there is no total explanation of history, which has always been the unpredictable outcome of competing and interacting interests. One thing is certain, nevertheless: in various forms there will always be a struggle for predominance. Mosca's views, more clearly set out than Pareto's, have a salutary realism.

The American philosopher and critic James Burnham also analyzed shifts of power. In *The Managerial Revolution* (1941) he propounded a theory of bureaucratic revolution: the rulers of the new society, the class with power and privilege, will be the bureaucratic managers of "super states." In *The Machiavellians, Defenders of Freedom* (1943), he reinterprets Machiavelli and cites Mosca as a modern Machiavellian. Following Pareto's idea of the "circulation of elites," he asserts that, when a ruling class becomes inadequate, frivolous, or bored, loses confidence in itself and its myths, and becomes irresolute in deploying necessary force, new elites are bound to take over—as in the managerial revolution of the 20th century.

RELIGIOUS AND EXISTENTIALIST APPROACHES

In the second, religious and quasi-religious group of political philosophies, the Catholic hierarchy has reiterated its ancient neo-Thomist doctrine of original sin and redemption. Pope Leo XIII, in the encyclicals *Inscrutabili Dei Consilio* (1878), *Immortale Dei* (1885), and *Rerum Novarum* (1891), dismissed all anthropocentric political philosophies as new versions of old heresies. The world, "through an insatiable craving for things perishable," was "rushing wildly upon the straight road to destruction." Society is intelligible only in the light of the Christian revelation and a future life:

> exclude the idea of futurity and forthwith the very notion of what is good and right would perish, nay, the whole scheme of the universe would become a dark and unfathomable mystery.

Such is the human condition that visionary innovations are fruitless, and "venomous" teachings can only bring "death-bearing fruit." Society, as St. Augustine had declared, if organized without God, can only be a present hell. Hierarchy, authority, and censorship can alone "control the excesses of the unbridled intellect, which unfailingly end in the oppression of the untutored multitude." Property is essential to the family, on which the social order depends, and inequality is inherent in all human societies. Only a harmonious Christian commonwealth can assuage the consequences of sin, and within that social order the state should therefore encourage Christian trade unions and promote the welfare of the poor. Thus, the papacy, maintaining its monopoly of revelation, tried to come to terms with the demands of industrial civilization.

During the rise of 19th-century nationalism and of Communist, Fascist, and Nazi dictatorships, and in face of the increasing dominance of governments and large-scale industry in all mass societies, the importance of individual responsibility with regard to moral issues was emphasized by a divergent group of thinkers who have come to be described as Existentialists. Søren Kierkegaard (died 1855), a Danish Neo-Kantian, declared that "truth is subjectivity" and that only by means of inward revelation can man know God. Jean-Paul Sartre, a brilliant French Existentialist, tried to come to terms with dialectical Materialism. His *Existentialism and Humanism* (1948) comprises an affirmation of human dignity. "If," he writes, "I have excluded God the Father, there must be someone to invent values." Man, who has abandoned God, "must liberate himself by some practical commitment," for only then can he become fully human. Sartre's elaborate *L'Etre et le néant* (1943; Eng. trans., *Being and Nothingness,* 1956) is at once Cartesian and laborious, complete with a "key" to its special and pedantic terms. It investigates the loneliness of the human condition, attitudes to others, love, masochism, indifference, desire, and hate. This intense introspection is even more vividly expressed in his fiction and drama. The Algerian-born Albert Camus in *The Myth of Sisyphus* (1942) and *The Rebel* (1951) also agonizes brilliantly over the current human condition, and in *Man in Revolt* he discards hope of pragmatic improvement.

Sartre's view of human dignity

REVOLUTIONARY DOCTRINES

The third stream of contemporary political philosophy is Marxist–Leninist totalitarian and neo-Marxist anarchist. Many of Marx's original insights into the socio-economic process and its effect on ideas are now generally accepted. His prophecies, on the other hand, have not been fulfilled. The proletarian revolution, for example, came not in an economically advanced country but in one of the most backward; and the state, far from withering away or

Doctrinal assumptions of latter-day Marxists

being diminished by inexorable economic trends, has in fact become more powerful both in Communist and in social-democratic countries. Those who have accepted the total Marxist revelation as superseding all else have had thus to adapt and revise it. Hence, much tortuous and artificial debate has ensued. All Marxists of legitimate descent accept the Hegelian position that one can get beyond empirical knowledge and perceive the historically revealed installments of a total explanation. They also start from Marx's 19th-century belief that economically determined conflicts among feudal, bourgeois, and proletarian classes are the dynamic of history and that the rule of law is not a safeguard for the whole society against arbitrary power but merely the expression of class interest.

Lenin. The first and by far the most significant interpretation of Marx's doctrine as realized in the Soviet Union was made by Lenin and developed by Stalin and is entirely authoritarian. According to Marx and Engels, the revolution could occur only after the bourgeois phase of production had "contradicted" the tsarist order, but Lenin determined to take advantage of the opportunities provided by World War I and settle accounts directly with the "accursed heritage of serfdom, of Asiatic barbarism . . . an insult to mankind," and in 1917 he engineered a coup that secured the support of the peasantry and the industrial workers. He also adopted the revolutionary theorist Leon Trotsky's idea of a "permanent revolution" from above by a small revolutionary elite, using the existing Russian bureaucracy.

Already in *What Is To Be Done?* (1902), Lenin had argued that an educated elite must direct the proletarian revolution, and when he came to power he dissolved the recalcitrant constituent assembly and ruled through a "revolutionary and democratic dictatorship supported by the state power of the armed workers." In asserting the need for a self-constituted elite of professional revolutionaries to seize power, Lenin reverted to Marx's program in *The Communist Manifesto*, rather than conforming to the fated pattern of economic development worked out in *Das Kapital*.

In 1921 he further adapted theory to the times. His new economic policy sanctioned the development of a class of prosperous "kulak" peasantry to keep the economy viable. For Lenin thought always in terms of world revolution; and in spite of the failure of the Marxists in central Europe and the defeat of the Red armies in Poland, he died in the expectation of a global sequel. Thus, in *Imperialism, the Last Stage of Capitalism* (1917), he had extended the class war into an inevitable conflict between European imperialism and the colonial peoples involved. He had been influenced by the English historian J.A. Hobson's *Imperialism, a Study* (1902), which alleged that decadent capitalism was bound to turn from glutted markets at home to exploit the toil of "reluctant and unassimilated peoples."

But, as observed by classical, medieval, and modern constitutionalist political philosophers, authoritarian regimes suffer the tensions of all autocracies. Marx himself might have thought that such planned autocracies had made the worst of his revelation, which was to lead to an unplanned, classless utopia.

Other Marxist approaches. Many Marxist revisionists tend toward anarchism, stressing the Hegelian and utopian elements of his theory. The Hungarian György Lukács, for example, and the German Herbert Marcuse, who fled from the Nazis to the United States, have won some following among those in revolt against both authoritarian "peoples' democracies" and the diffused capitalism and meritocracy of the managerial welfare state. Lukács' "History and Class Consciousness" (published in Hungarian, 1922), a neo-Hegelian work, claims that only the intuition of the proletariat can properly apprehend the totality of history. But world revolution is contingent, not inevitable, and Marxism is an instrument, not a prediction. Lukács renounced this heresy after residence in the Soviet Union under Stalin, but he maintained influence through literary and dramatic criticism. After Khrushchev's denunciation of Stalin, Lukács advocated peaceful coexistence and intellectual rather than political subversion. In *The Meaning of Contemporary Realism* (trans. 1963), he again relates Marx to Hegel and even to Aristotle, against the Stalinist claim that Marx made a radically new departure. Lukács' neo-Marxist literary criticism can be tendentious, but his neo-Hegelian insights, strikingly expressed, have appealed to those anxious to salvage the more humane aspects of Marxism and to promote revolution, even against a modified capitalism and social democracy, by intellectual rather than by political means.

Marcuse's assessment of revolutionary prospects

Marcuse also reached back to the more utopian Marx. Now that most of the proletariat has been absorbed into a conformist managerial capitalism or has been regimented into bureaucratic peoples' democracies, freedom, argues Marcuse, is in retreat. In Western affluent societies most employers and workers are equally philistine, dominated by the commercialized mass media, or "cogs in a culture machine." The Soviet Union has reverted to an even more philistine monolithic repression, distorting art and literature. This enslavement of man by his own industrial productivity has been clinched by the colossal power of governments, which renders the old brief and brisk class warfare a romantic, impracticable idea. Marcuse attacked all establishments and transferred the redeeming mission of the proletariat to a fringe of alienated minorities—radical students and the exponents of the "hippie" way of life—as well as to Viet Cong guerillas and Black Power militants. Such groups, he declared, could apparently form liberating elites and destroy the managerial society. Thus reappeared the old Marxist–Hebraic pattern of redemption through struggle by a chosen people.

The Italian Communist Antonio Gramsci deployed a vivid rhetorical talent in attacking existing society. Like Marcuse, Gramsci was alarmed that the proletariat was being assimilated by the capitalist order and flung himself into the current controversy between social democrats and the radical revolutionaries. He took his stand on the already obsolescent Marxist doctrine of irreconcilable class war between bourgeois and proletariat. He aimed to unmask the bourgeois idea of liberty and to replace parliaments by an "implacable machine" of workers' councils, which would destroy the current social order through a dictatorship of the proletariat. "Democracy," he wrote, "is our worst enemy. We must be ready to fight it because it blurs the clear separation of classes."

Not only would parliamentary democracy and established law be unmasked, but culture, too, would be transformed. A workers' civilization, with its great industry, large cities, and "tumultuous and intense life," would create a new civilization with new poetry, art, drama, fashions, and language. Gramsci insisted that the old culture should be destroyed and that education should be wrenched from the grip of the ruling classes and the church and run under the direction of workers' councils.

But this militant revolutionary was also a utopian. He turned bitterly hostile to Stalin's regime, for he believed, like Engels, that the dictatorship of the workers' state would wither away. "We do not wish," he wrote, "to freeze the dictatorship." Following world revolution, a classless society would emerge, and mankind would be free to master nature instead of being involved in a class war.

Since World War II, Gramsci's notions have enjoyed a minor revival. They appeal to the fringe of revolutionaries who admire Marcuse and detest the embourgeoisement of an idealized proletariat. But, in a civilization in which, if total war can be avoided, material prospects are good, the destruction of the old culture out of rage, envy, and naïve idealism appears to be a pointless program. Like Marcuse's doctrine, it is a cry of pain, typical of the 1920s in Italy.

III. Conclusion

The history of political philosophy from Plato until the present day makes plain that modern political philosophy is still faced with the basic problems defined by the Greeks. The need to redeploy public power in order to maintain the survival and enhance the quality of human

life, for example, has never been so essential. And if the opportunities for promoting well-being are now far greater, the penalties of the abuse of power are nothing less than the destruction or gross degradation of all life on the planet.

In these circumstances it is of no great importance that some analytical philosophers have declared themselves neutral; they have at least often discredited pretentious metaphysical myths. On the empirical evidence, constitutionalism and the rule of law, with the ancient classic, medieval, and humanist traditions behind them, have proved themselves a more successful response to the environment than tyranny and repression. In the current and more sophisticated view, there are no shortcuts to the millennium. As Mosca points out, utopian ideas become

> dangerous when they succeed in bringing a large mass of intellectual and moral energies to bear upon an end that can never be achieved, and that in the day of purported achievement can mean nothing more than the triumph of the worst people and distress and disappointment for the good.

There will perhaps always be a struggle for pre-eminence in any society, and public laws are necessary to regulate it. Too much cannot be hoped of government, and the best society is that in which tyranny and caprice of power are prevented and in which men are free to create diverse and spontaneous institutions within the framework of law. Only within such a framework of a tolerably well-organized constitutionalism, gradually extended to relations between states, can the swiftly mounting opportunities provided by applied science be taken and the pattern of social life adjusted, so that the human species, instead of being thwarted and deformed by its institutions, can realize its full potentialities.

BIBLIOGRAPHY. The following editions of standard works and commentaries may be consulted.

Classical: PLATO: *The Republic*, trans. with an introduction by A.D. LINDSAY (1957); *The Statesman* (Loeb edition, 1925); *The Laws*, ed. by A.E. TAYLOR (1934). *The Politics of Aristotle*, trans. with introduction by ERNEST BARKER is still the best edition (1946). For the social background of Greek thought, see MAURICE BOWRA, *The Greek Experience* (1957). CICERO, *On the Commonwealth* (*De re publica*) is available in translation by G.H. SABINE and S.B. SMITH (1929). For *De legibus*, see the Loeb edition (1928, reprinted 1966). ST. AUGUSTINE, *The City of God* has been translated by JOHN HEALEY, with an introduction by ERNEST BARKER (1931). For the background, see PETER BROWN, *Augustine of Hippo* (1967) and *The World of Late Antiquity* (1971). C.N. COCHRANE, *Christianity and Classical Culture* (1940), is worth consultation; and H. OSBORN TAYLOR, *The Medieval Mind*, 4th ed., vol. 2 (1925), is still valuable for patristic and medieval thought.

Medieval: The most convenient edition of John of Salisbury is *The Statesman's Book of John Salisbury*, trans. and selected by J. DICKSON (1927). For St. Thomas Aquinas, see the *Selected Political Writings*, ed. by A.P. D'ENTREVES (1959). The full texts are in *The "Summa Theologiae" of St. Thomas Aquinas*, 2nd ed., 22 vol., trans. by the FATHERS OF THE ENGLISH DOMINICAN PROVINCE (1921–32). For DANTE'S *Monarchia*, see P.H. WICKSTEED'S translation in *The Latin Works of Dante* (1904), and that by D. NICHOLL (1954). See also A.P. D'ENTREVES, *Dante As a Political Thinker* (1952). The best general survey is R.W. and A.J. CARLYLE, *A History of Medieval Political Theory in the West*, 6 vol. (1903–36). A.L. SMITH, *Church and State in the Middle Ages* (1913); and C.H. MCILWAIN, *The Growth of Political Thought in the West, from the Greeks to the End of the Middle Ages* (1932), are still valuable.

Sixteenth–eighteenth centuries: NICCOLO MACHIAVELLI, *The Prince* and *Discourses on the First Decade of Titus Livius* are edited with an introduction by MAX LERNER (1950); his *Literary Works* are edited and translated by J.R. HALE (1961). See also ALLAN H. GILBERT, *Machiavelli's Prince and Its Forerunners* (1938); HERBERT BUTTERFIELD, *The Statecraft of Machiavelli* (1962); and FELIX RAAB, *The English Face of Machiavelli* (1964). THOMAS HOBBES, *The Leviathan, or the Matter, Forme and Power of a Commonwealth, Ecclesiastical and Civil* is edited by A.D. LINDSEY (1950); see also *The Complete Works*, ed. by WILLIAM MOLESWORTH, 11 vol. (1839–45). K.V. THOMAS, "The Social Origins of Hobbes's Political Thought," in *Hobbes Studies*, ed. by K.C. BROWN (1965), is the best account. JOHN BOWLE, *Hobbes and His Critics* (1951), describes Hobbes's political impact and the contemporary reaction to it. SPINOZA'S *Tractatus Theologico-Politicus* (in part) and the full *Tractatus Politicus* have been edited by A.G. WERNHAM (1958). See also STUART HAMPSHIRE, *Spinoza* (1951). For general background, see ALFRED NORTH WHITEHEAD, *Science and the Modern World* (1925). RICHARD HOOKER, *The Laws of Ecclesiastical Polity*, ed. by JOHN KEBLE, 3 vol. (1836), was revised by R.W. CHURCH and F. PAGET (1888). See also C.J. SISSON, *The Judicious Marriage of Mr. Hooker and the Birth of The Laws of Ecclesiastical Polity* (1940). JOHN LOCKE, *Two Treatises on Civil Government* has been edited by PETER LASLETT (1960); *The Second Treatise of Civil Government and A Letter Concerning Toleration* by J.W. GOUGH (1947). For EDMUND BURKE, see his *Complete Works* (Bohn's Classics, 1853) and *Selected Writings*, ed. by R.J.S. HOFFMAN and P. LEVACK (1949). See also PHILIP MAGNUS, *Edmund Burke: A Life* (1939); and ALFRED COBBAN, *Edmund Burke and the Revolt Against the Eighteenth Century*, 2nd ed. (1960). GIOVANNI BATTISTA VICO, *The New Science* is translated by T.G. BERGIN and M.A. FISCH (1961). See HENRY P. ADAMS, *The Life and Writings of Giambattista Vico* (1935); and BENEDETTO CROCE, *The Philosophy of Gianbattista Vico*, trans. by T.G. COLLINGWOOD (1964); also CHARLES LOUIS DE SECONDAT, *Oeuvres choisis*, trans. by J. CHAIX-RUY (1946). MONTESQUIEU, *The Spirit of the Laws* is translated by THOMAS NUGENT, with an introduction by FRANZ NEUMAN (1949). For the French Enlightenment generally, see J.B. BURY, *The Idea of Progress* (1921); and CARL BECKER, *The Heavenly City of the Eighteenth Century Philosophers* (1933). The best translation of JEAN-JACQUES ROUSSEAU'S *Social Contract*, with an introduction, is that of ERNEST BARKER (1947). See also *The Political Writings of Jean Jacques Rousseau*, ed. by C.E. VAUGHAN, 2 vol. (1915). WILLIAM H. BLANCHARD, *Rousseau and the Spirit of Revolt* (1967); and J. MCMANNERS, *The Social Contract* and *Rousseau's Revolt Against Society* (1968), are both illuminating. DAVID HUME, "Of the Original Contract" is available in *Social Contract*, ed. by BARKER (*op. cit.*)

The nineteenth century: JEREMY BENTHAM, *A Fragment on Government* and *An Introduction to the Principles of Morals and Legislation* are edited by WILFRED HARRISON (1948). See also JOHN PLAMENATZ, *The English Utilitarians*, 2nd rev. ed. (1948), particularly for Mill. J.S. MILL, *On Liberty* and *Considerations on Representative Government* are edited by R.B. MCCALLUM (1946). ALEXIS DE TOCQUEVILLE, *De la démocratie en Amérique* (Eng. trans., *Democracy in America*), ed. by PHILIPS BRADLEY, 2 vol. (1945), is the best edition. For a good abridgment, see H. REEVES's translation, ed. by H.S. COMMAGER (1946). *L'Ancien Régime* was edited by G.W. HEADLAM (1904). AUGUSTE COMTE, *Système de politique positive (System of Positive Polity)* is translated by J.H. BRIDGES, 4 vol. (1875–77). See J.S. MILL, *Auguste Comte and Positivism* (1865); and EDWARD CAIRD, *The Social Philosophy and Religion of Comte* (1885). BASIL WILLEY in *Nineteenth Century Studies* (1949), has a good essay. *Hegel's Philosophy of Right* is translated by T.M. KNOX (1942); his *Philosophy of History* by J. SIBREE, rev. ed. (1944). HERBERT MARCUSE, *Reason and Revolution: Hegel and the Rise of Social Theory*, 2nd ed. (1960), makes a new interpretation. ROBERT OWEN, *A New View of Society and Other Writings* are edited by G.D.H. COLE (1927). PIERRE-JOSEPH PROUDHON, *Qu'est-ce que la propriété? (What Is Property? An Enquiry into the Principles of Right and Government)* is edited by J.A. LANGLOIS (1966). See his *Oeuvres complètes*, ed. by C. BOUGLE and H. MOYSSET, 15 vol. (1923–36)—also D.W. BROGAN, *Proudhon* (1934). KARL MARX, *A Contribution to the Critique of Political Economy* was translated by N.T. STONE, 2nd rev. ed. (1904). See also *The Communist Manifesto*, ed. by HAROLD J. LASKI (1948); *Capital: A Critique of Political Economy*, trans. by E. and C. PAUL (1928); *Capital: The Process of Capitalist Production As a Whole* and *The Process of Circulation of Capital*, trans. by ERNEST UNTERMANN (1909); *The Holy Family* (1956); and *Selected Writings in Sociology and Social Philosophy*, ed. by T.B. BOTTOMORE and M. RUBEL (1964). Studies of Marx particularly recommended are FRANZ MEHRING, *Karl Marx: Geschichte seines Lebens* (1933; Eng. trans., *Karl Marx: The Story of His Life*, 1948); and ISAIAH BERLIN, *Karl Marx: His Life and Environment*, 4th ed. reprinted with corrections (1960). JOHN PLAMENATZ, *German Marxism and Russian Communism* (1954), is illuminating. FRIEDRICH ENGELS' works in translation include *The Condition of the Working Classes in England in 1844* (1958); *The Anti-Dühring: Herr Eugen Dühring's Revolution in Science* (1959); and *The Origin of the Family, Private Property and the State* (1884). The standard biography is by GUSTAV MAYER, *Friedrich Engels*, 2 vol. (1934; Eng. trans., 1936). GEORGES SOREL, *Reflectons sur la violence* (1912; *Reflections on Violence*), was translated by T.E. HULME (1915), and by J. ROTH (1950). See also *Les Illusions du progrès*

(1908; 5th ed., 1947); and P.L. PERRON, *Les Idées sociales de George Sorel* (1935).

The twentieth century: ISAIAH BERLIN, *Historical Inevitability* (1954) and *Two Concepts of Liberty* (1958); EMILE DURKHEIM, *The Rules of Sociological Method*, trans. by S.A. SOLVAY and J.H. MUELLER, ed. by C.E.G. CATLIN (1938); VILFREDO PARETO, *Trattato di sociologia generale*, 2nd ed. (1923; Eng. trans., *The Mind and Society*, ed. by A. LIVINGSTON, trans. by A. BONGIONNO and A. LIVINGSTON, 4 vol., 1935); GAETANO MOSCA, *Elementi di scienza politica*, 2nd ed. (1923; Eng. trans., *The Ruling Class*, ed. by A. LIVINGSTON and trans. by H.D. KHAN, 1939); LEO XIII, *The Pope and the People: Select Letters and Addresses on Social Questions* (1929); JEAN-PAUL SARTRE, *L'Être et le néant (Being and Nothingness: An Essay in Phenomenological Ontology)*, trans. and ed. by H.E. BARNES (1957); *Existentialism and Humanism*, trans. and ed. by PHILIP MAIRET (1948); LENIN, *The Development of Capitalism in Russia* (1899); *Collected Works*, 4th ed. (1960), vol. 3, *What Is To Be Done?*, vol. 5, *The State and Revolution*, vol. 25, *Imperialism: The Highest Stage of Capitalism* (1916); see also his *Selected Works*, 12 vol. (1936–39). GEORGE LICHTHEIM, *Marxism: An Historical and Critical Study*, 2nd rev. ed. (1964), is a good short account. See also HERBERT MARCUSE, *Reason and Revolution: Hegel and the Rise of Social Theory*, 2nd ed. (1960), *Eros and Civilization* (1955), *Soviet Marxism: A Critical Analysis* (1958), and *One Dimensional Man* (1964); ALASDAIR MACINTYRE, *Marcuse* (1970); GYORGY LUKACS, *Goethe und seine Zeit* (1947; *Goethe and His Age*, trans. by ROBERT ANCHOR, 1968); *Der historische Roman* (1955; *The Historical Novel*, trans. by H. and G. MITCHELL, 1962); and JAMES JOLL, *The Anarchists* (1965). For A. Gramsci, see *The Modern Prince and Other Writings*, trans. and introduced by L. MARKS (1957); J.M. CAMERETT, *Antonio Gramsci and the Origins of Italian Communism* (1967); A. POZZOLINI, *Antonio Gramsci: An Introduction to His Thought* (1970).

General histories of political philosophy: G.H. SABINE, *A History of Political Theory*, 3rd ed. (1961), is the most comprehensive survey. W.A. DUNNING, *A History of Political Theories*, vol. 1, *Ancient and Medieval*, vol. 2, *From Luther to Montesquieu*, vol. 3, *From Rousseau to Spencer* (1902–20, reprinted 1936–38), is still valuable. See also KARL POPPER, *The Open Society and Its Enemies*, 5th rev. ed., vol. 1, *The Spell of Plato* and vol. 2, *The High Tide of Prophecy: Hegel, Marx and the Aftermath* (1966). ERNEST BARKER, *The Principles of Social and Political Theory* (1951), analyses essential problems. JOHN BOWLE, *Politics and Opinion in the Nineteenth Century* (1954, paperback 1964); and WILLIAM EBENSTEIN, *Modern Political Thought: The Great Issues*, 2nd ed. (1960), are valuable, as is HAROLD D. LASSWELL, *The Future of Political Science* (1963).

(J.E.B.)

Political Power

Power, conceived broadly as man's influence over man, has been a deep and self-conscious human concern for more than 2,000 years. From Aristotle, Machiavelli, and Marx to modern political writers and activists, political philosophy and political conflict have revolved basically around who should have power over whom. But despite over two thousand years of conjecture, and at least occasional attempts at systematic analysis, scientific understanding of power phenomena has emerged only in the 20th century. The topic has been as elusive and formidable as it is vital.

The concept of power is employed widely in all the social sciences. Anthropologists study power, authority, and leadership in small communities ranging from primitive tribes to peasant villages and urban neighbourhoods. Sociologists investigate power as a ubiquitous aspect of social stratification, involving classes, elites, status groups, legal systems, social movements, and organizations, such as bureaucracies and parties. Psychiatrists and psychologists probe human motivations toward power and human reactions to its use. Economists analyze power in the market—monopolistic, duopolistic, oligopolistic, and perfectly competitive—as well as the influence of international firms and economic forces. Management specialists from various disciplines investigate the processes of power and control within organizations, while social psychologists concentrate on the same phenomena in small groups. Lastly, political scientists frequently take the study of power to be the very focus of their discipline. Having traditionally approached the topic through an examination of the formally authoritative institutions of power, such as constitutions and governments, they moved on to immediately related agencies, such as political parties and interest groups. Subsequently, political scientists have investigated power phenomena in virtually any significant setting, including families, occupations, and voluntary associations. The analysis of power remains a basic concern of social sciences.

CONCEPTS OF POWER

Analyses of power in the social sciences

In very fundamental terms, analysts of power have confronted three initial problems: (1) the concept of power itself; *i.e.*, what should be meant by the term power; (2) what aspects and types of power and power structures need to be distinguished; and (3) how power, so defined and elaborated, can be located and measured. The first two problems only will be described in this article; the third is a problem of social-scientific methodology.

Meaning and characteristics of power. The etymology of the term power is traced back to a Latin root (*potere*, *potesse*, or *posse*) essentially meaning "to be able to," or having the capacity to do something. One common meaning of "power"—now a rather ambiguous term—is, therefore, the "ability to act," "capacity to produce a result," or simply "strength." In politics, this meaning became more specialized, referring to the ability to affect the actions of others; that is, to produce alterations in their behaviour.

Political power is frequently distinguished from other forms of power, such as economic, social, and religious. The power that is political is defined in terms of an institution such as the government or the state. Political power then becomes power wielded in relation to the government or the state, either by it, or to affect it, or both. It is power directly or indirectly of and over government or the state.

A less common way of viewing political power is to regard it as power employed in relation to collective actions, governmental or other. Collective acts require the cooperation of more than one person and are a means by which the goals of collectivities are realized. As the American sociologist Talcott Parsons saw it, for example,

> The political process then is the process by which the necessary organization [for collective action] is built up and operated, the goals of action are determined and the resources requisite to it are mobilized.

Political power thus becomes defined in terms of actors' effects on this process, sometimes broadly interpreted to include all kinds of effects and sometimes more narrowly limited in terms of legitimacy and other considerations.

A third view of political power is stated or implied by those who define what is "political" in terms of power. Though at first it might appear that "political power" would then be a redundant expression, this is not true. Rather, political power is basically defined as "power over the allocation and distribution of power." Political power is thus power over power; the more a particular power relation affects other power relations, the more political it is. The power of the president of the United States over the Senate majority leader, for example, presumably includes much more political power than that of a local machine "boss" over a party hack even though the actual amount of direct control between the first pair may be less than that between the second pair.

Working definition of power

In the remainder of this article, unless otherwise indicated, power will be defined as a relationship in which the behaviour of one or more persons alters the behaviour of another or others. Political power will be construed as power over the allocation and distribution of power in some social unit (nation, city, organization) that must always be specified.

General aspects of power relationships. Various aspects of power have been found to be of critical importance in virtually all power relations. One is simply the matter of the actors involved. Any power relationship minimally involves two actors—one who influences and another who is influenced. Often there are several of each along with intermediaries who link the original influencer and ultimate influencee in a power chain.

Actors involved. A crucial feature of any power analysis is the choice of actors and actor characterizations. The power structure of a city, for example, may look quite different depending on who the main actors in the analysis are. They could be the formal governmental leaders such as the mayor, the councilmen, or the sheriff, on the one hand, or they could be the "old political hands" and the "young Turks"; the "Establishment" and the "People"; or liberals, radicals, and conservatives. When analyses of political power seem to disagree, it is frequently because they focussed on different political actors. Conclusions about political power thus must be interpreted in light of the adequacy of the choice and description of the actors selected.

Since power is a behavioral relationship among actors, description of the specific actions of all of them is important. Few actors influence everything another actor does, and few are affected by everything done to them. Studies of political power generally stress certain kinds of action on the part of the influencer (called his "influential behaviour") and certain kinds of response by the actor influenced (usually called the "scope" of the relationship). To paraphrase the American political scientist Harold Lasswell, power descriptions attempt to answer the question, "Who does what to whom with what effect?"

Recognition. In many settings, the more ample and accurate that an actor's perception is of the power relations around him, the greater his own power is likely to be. For this reason, many power struggles centre about control over information. Also, it often seems that an actor who newly recognizes that he is being influenced by another actor tends to strive to reduce that influence if possible. Hence the attractiveness to the influencer of "manipulation"—a power relation that is intended and recognized by the influencer, but is unrecognized by the actor who is influenced. Some analysts consider that manipulative power is rising in modern societies. One of the best means of destroying manipulative power is to make the person influenced aware that he is being manipulated.

Rapidity and power sequencing. Timing plays a prime role in political power. Two of the several facets of timing are of particular importance. One can be called "rapidity" or "response lag," the other, "power sequencing" (see below). Rapidity refers to the time that elapses between the influential behaviour of the influencer and the response (scope) of the one influenced. Many power conflicts seem to be exacerbated by discrepant interpretations of rapidity. University students, for example, attempting to influence their school's administration, may regard four months as an inordinately long response time, probably indicative of stalling. Administrators and faculty, on the other hand, may view such a response lag as extremely rapid—about as fast as the institution can be expected to react—and a clear manifestation of good faith.

The element of rapidity is extremely important to some types of organizations. The response lag is shortened when power is patterned in certain ways—especially hierarchically, unilaterally, and directly—and when certain types of power, such as absolute authority, are employed. Military units in combat, for example, require great rapidity of response and correspondingly adopt a hierarchical and authoritative power structure. Surgeons performing operations have similar needs and work within similarly concentrated and authoritative structures. In other types of organizations, however, rapidity of response is not as crucial. Filibustering and certain other forms of stalling in legislatures, for example, can be viewed as deliberate, legally permitted efforts to lengthen the response lag so as to deter the influencer from his attempt.

"Power sequencing"

Many power relations are not one-shot affairs. In the second facet of timing, "power sequencing," an original influence attempt may be followed by refusal, partial compliance, or evasion on the part of the one intended to be influenced; then another influence attempt is made, followed by another response, and so on. During such exchanges both parties often engage in contingency planning. The government, for example, must not only pass legislation, but it must also indicate what is to happen if citizens violate the legislation, what appeals can be made, how the violation is to be discovered. Actors vary significantly in the richness of their power repertoires, in the sophistication of the influence sequences they can handle. Some actors lapse quickly into violence and other fairly extreme responses, while others display great ingenuity in constructing alternative tactics. Some political systems expect quicker compliance and less sequencing than others. Explosions often occur when the sequencing expectations of the participants in a power relation are conspicuously divergent.

Resistance and cost. In attempting to compare the power of various actors, resistance and cost must be considered. Though very difficult and only imprecisely worked out, these important notions refer crudely to how hard or easy it is to get the influenced actor to respond (his resistance) and how much of a sacrifice, if any, it will be for the influencer to engage in the influential behaviour (his cost). Two groups may each get three bills through a legislature, but if one set of bills is noncontroversial or trivial while the others are faced with strong opposition but important, the difference in resistance shows the second lobby exerted greater power.

Indirect power. Some power relations are direct and some are indirect. In an indirect relation, intermediaries come between the original influencer and the ultimate influencee. Such indirect relations have not been well studied, but they seem to offer certain advantages and disadvantages. Indirect relations, for example, are less likely to be recognized. Thus, when governments or political parties wish to influence public opinion but know that a direct communication from them would be suspect, they frequently work through various agents. Indirect relations—very long chains of them, in fact—are inevitable in any large and complicated power structure. Such complicated patterns of power are part of the basic fabric of all modern societies.

Some indirect relations, on the other hand, often prove to be relatively fragile, offering opportunities to intermediaries to take power into their own hands. Indirectness masks the influencer's ability to observe compliance just as it masks recognition by the actor influenced, sometimes to the point that no one accurately perceives the true nature of the power patterns. Responsibility is then hard to assess.

The mask of power in indirectness

Reciprocity. A further important aspect of power relations is called reciprocity. Often power relations get reversed, so the original influencer is in turn influenced by the influencee. Such reciprocity phenomena appear to be especially significant. Democracy, for example, exhibits a relatively high incidence of reciprocal relations in the political power structure, in which the government influences the people but the people also influence the government. Several basic kinds of reciprocity can be observed, and students of political power investigate the causes and consequences of each.

Types of power. *Bases for classification.* Probably the most common classification of power would divide it into economic, social, and political categories—an essentially institutional classification. Political power pertains to the state or government, economic power to the financial and production establishments, and social power to the remaining institutions. Similar, but more elaborate, classifications that have been suggested would include religious, military, judicial, informational, and international power along with economic, social and political power.

Power and values

From another perspective, these various types of power can be regarded as referring to control over different kinds of valued resources. Political scientists who have pursued this line of thought most explicitly list eight types of human values: power, respect, rectitude, affection, well-being, wealth, skill, and enlightenment. Since power can be based on control over any of these values, there are eight basic types of power ("forms of influence"): political power, councillorship, mentorship, personal influence, violence, economic power, expertness, and advisory influence, respectively.

Another approach is to classify power relations in terms of types of behaviour—that is, according to kinds of influential behaviour (the behaviour of the influencer), or of scope (the behaviour of the one influenced). The former requires a classification of the types of behaviour an influencer can use to affect the others' responses; the latter requires a classification of the main dimensions along which an influencee can be affected.

Amitai Etzioni, a political sociologist, has analyzed complex organizations using a typology of power based on influential behaviour. The kind of question he is concerned with is essentially, "What kinds of things can actors do to influence other actors," and he discerns three main types of power: coercive (the use or threat of physical sanctions), remunerative (the use of material resources as rewards and punishments), and normative (the use of symbolic resources as rewards and punishments). Etzioni argues that although most organizations employ all three types of power, they tend to stress one more than the others. The military thus stress coercive power, business firms stress remunerative power, and churches stress normative power. He contends further that the basic type of power used by an organization is associated with differences in such factors as organizational goals, leadership characteristics, the kind and degree of consensus within the organization, recruitment patterns, and processes of training and socialization.

The last fundamental approach to classifying power relations focusses on the scope of the relationship. This approach poses the question: "In what ways can any actor's behaviour be affected by other actors?" This in turn requires a model of the state of an actor at any given time. In the same way that alterations in a physical object can be determined only by reference to regular descriptive categories, such as colour, shape, and hardness, in order to determine changes in influenced actors in a power relation, it is necessary to have categories for describing the various kinds of change that can occur.

Many typologies of power are based on such an analysis, either explicitly or implicitly. The English philosopher Bertrand Russell, for example, suggested three main types of power: direct physical power over the bodies of individuals, rewards and punishments, and influence on opinions. Within this framework, he noted further subtypes, such as priestly, kingly, naked, revolutionary, and economic power. But the basic typology centres on the scope of the power relationship, suggesting that human actors can be influenced physically, or through rewards and punishments, or through opinion change. More recent thought distinguishes physical power, in which actors are influenced merely as physical entities, from psychological power, in which behavioral effects are produced through first influencing the mind. Psychological power is divided into cognitive and motivational categories. Cognitive power (influencing an actor by changing his perceptions) mainly consists of threats of punishment and promises of reward. Motivational power (influencing an actor by altering his drives) consists of the inculcation of ethics and the generation of desires.

Problems of power classification. One outstanding problem with all attempts to classify power is that the categories developed are not mutually exclusive in an empirical sense. In other words, a given concrete power relationship rarely falls exclusively into any single category. If the government imprisons a man, it is simultaneously using physical power, cognitive power, and motivational power.

Power and authority

One of the most important types of power is authority. The interpretations of authority are many, but two conceptions seem uppermost. One views authority as compulsive power—power which brooks no refusal, which is binding on the actor influenced. He does not pause to ask why; when confronted with an authoritative imperative, he obeys. The other main notion of authority is more limited and is based on legitimacy. People learn conceptions of what constitutes proper, ethically desirable power relations. The actual relations that conform to these conceptions of what ought to be are said to have legitimacy for the perceiver. He is under strong moral pressure to obey legitimate influence assertions. Authority is a power relation involving the legitimate power of one role over another, for authority is not vested in particular idiosyncratic individuals, but in roles, offices, and positions. Virtually all political systems endeavour to establish the authority of crucial roles and institutions. It is uncertain, however, how long pure authority endures if it is not reinforced by other types of power.

Power structures. Theoretically, for any clearly specified social unit, such as a nation, community, or organization, and for any appropriate time period, one can designate the set of actors in the unit and the set of power relations among those actors. Together, these constitute the power structure of the unit.

Assuming sufficient information about actors and their power relations, structural analysis is concerned with three main questions: (1) how to provide a summary description of the power structure of the unit; (2) how to explain the particular structural configurations found; and (3) how to ascertain the consequences of those structural patternings for other variables such as the distribution of scarce resources, the unit's effectiveness at performing certain tasks or functions, and the morale of the actors in the unit.

Terms describing power structure

The analysis here will be confined to the first problem, that of describing the power structure of the unit. Many terms in ordinary language offer useful distinctions: oligarchy, autocracy, polyarchy, democracy, totalitarianism, plutocracy, anarchy, dictatorship, pluralism. The Greek endings "-archy" and "-cracy," from *archōn* ("ruler") and *kratos* ("rule, power"), have been attached to scores of stems to form structural power terms. Examination of such terms brings out three important aspects of power structures. The first and most basic is the distribution of power within the unit; that is, how widely shared the existing power relations are among the members. Does power tend to be concentrated in a few hands or dispersed? Aristotle, for instance, distinguished three fundamental types of power structures based on the distribution of power. These were, in their pure forms, kingship, aristocracy, and democracy; and in their corrupt forms, dictatorship, oligarchy, and ochlocracy (mob rule). They refer, respectively, to rule by one actor, rule by a few actors, and rule by many or all actors.

The second aspect of power structures, *who* rules, is reflected in ordinary language. The type of actor in the system with greatest power is in a plutocracy, the rich; in a gerontocracy, the aged; and in a timocracy, the propertied class.

The third structural aspect commonly noted in ordinary language describes the extent of the ruling element's power—"total," as in totalitarian systems, or limited and fragmented, as in pluralistic, democratic societies.

POWER IN VARIOUS SETTINGS

There is a vast lore about political power as it is exercised in various settings, little of which is organized into coherent theory and much of which is either contradictory or false. Nevertheless, enough is known about power at different levels of social organization to summarize the dynamics of its exercise at three levels: in groups, organizations, and communities.

Small groups. The small group is the setting for much of political life. Cabinets, party-executive committees, supreme courts, legislative committees, revolutionary juntas, town-finance committees, bureaucratic staffs, personal entourages, party cells, and even friendships all come to mind. In addition to its intrinsic importance, the small group is the building block from which larger political organizations are constructed. Hence, they are sometimes called "primary groups."

The most visible political actors are those occupying prominent organizational roles, such as the chief executive of the government, the leader of the opposition, or the spokesman for an interest group. Organizational roles, however, only set rough possibilities and limits for the actual exercise of power, which is always personal. The particular person playing the role must activate the potential power inherent in his organizational authority,

operating within the context of a small group much of the time. Some analysts, in fact, assert that most significant political decisions are made in small groups. An understanding of the dynamics of small groups is thus essential to an understanding of power.

Many definitions of the small group have been offered. All seem to revolve not about absolute size but around the notion of direct, face-to-face interaction. In the small group, each member knows and interacts directly with every other member. In contrast to the relatively large, indirect and impersonal organization, the small group tends to be relatively personal, direct, and intimate. Individual style, personal liking and disliking, and face-to-face interaction skills, which are much less important in the larger context, are crucial in the small group.

Leadership roles in small groups

The face-to-face interaction that characterizes the small group brings the emotional, affective or expressive features of relationships into greater prominence. Experimental studies, in fact, have suggested a bifurcation of group leadership along these lines. An instrumental, or task, leader moves the group toward realization of its manifest goals. Such leadership, however, is likely to generate tension and dissatisfaction. The instrumental leader, moreover, is usually somewhat distant from other group members. Hence an expressive leader also frequently emerges, acting to assuage the tensions and conflicts arising from pressure toward task completion.

Compelling emotional ties are frequently formed in the small group context—ties not merely to specific members but to the group itself as a symbolic entity. The member's need for acceptance and solidarity, his need for the cooperation of others in achieving many personal goals—all are disproportionately met in small-group settings. Satisfying these needs gives such groups great power over the individual. It is no accident that political organizations that make extremely difficult demands on their members are especially likely to be organized into relatively small, tight, and totally involving cells. Similarly, some argue that the more crucial the political decision, the smaller the unit handling it and the simpler its role structures. It has even been suggested that, as a general principle, other things being equal, the control of a social unit over its members is inversely related to its size.

Though the small groups of a political organization can be a major source of organizational power, supporting the formal structure in critical ways, they can also be a challenging and disruptive force. If individual members who oppose the larger unit can band together in a small group, they can frequently reinforce each other and blunt the otherwise potent pressures to conform. Extreme deviance such as criminality, attempts at coups and cabals, die-hard resistance, or even "family circles" of mutual protection against party informers (as in the Soviet Union) is almost always incubated in a supportive small-group atmosphere. Awareness of this potential for power in small groups has led many totalitarian regimes to try to atomize interpersonal relations below the level of the state, except in carefully controlled regime-supporting organizations such as the Hitler Youth of Nazi Germany, or the Pioneers of Stalin's Russia.

Organizations. Most positions of great political power are obtained by moving into established roles. Any new power holder creates for himself only a tiny fraction of the enormous web of power relations he commands when he moves into an existing structure of authority and power resources.

The modern world is distinctly an organizational world; modern power is organizational power, and modern man is, unavoidably, an organization man. Complex organizations are indeed built up from small groups, but those complex organizational arrangements in turn frequently determine what sort of small groups will be formed —where, when, how and among whom. Contemporary political power, in particular, is most conspicuously revealed in organizational settings and in the clash of organizational actors.

Basically, organizations can be regarded as devices for coordinating the behaviour of individuals and groups. Bureaucracies coordinate the behaviour of officials, presumably to implement governmental policies. Political parties concert the behaviour of partisans in order to increase their chances of gaining control over the government. By organizing and coordinating their actions, individuals and small groups can accomplish goals that otherwise would be unlikely or impossible.

Goals and power. The power position and activities of an organization are fundamentally related to its goals. Other things being equal, the greater the consensus within the organization on goals and means, the less the need for the internal exertion of power. The more the members spontaneously want to coordinate their behaviour and agree on how to do so, the less need there is to provide organizational influences toward that end. For this reason, many organizations expend much effort on maximizing consensus in areas of activity regarded as critical. The two main instruments for obtaining consensus are recruitment (or, more broadly, the selection, promotion, and release of personnel) and socialization (the teaching of attitudes and behaviour to members). The more the organization selects people who already have an appropriate viewpoint or the more it trains existing members to adopt such an outlook, the less it needs to use force, authority, reward, punishment, and other forms of power to induce appropriate behaviour.

The role of consensus on goals and means

Theories of political power assign a central position to distinctions among crucial types of goals. Probably the most basic distinction is that between power itself as a goal and other substantive goals. "Political man" has often been regarded as the actor dominated by the pursuit of power in the same way that "economic man" is dominated by the pursuit of pecuniary profit. The detailed implications of such an assumption have never been elaborated for political man as carefully as they have for economic man. The main proposition that emerges from such a concept may be that the goal of power itself is in many respects "unbounded," a goal never fully satisfied. On the other hand, many substantive goals can be reached and extinguished; they do not motivate the political actor to seek power across the board but only in certain limited areas.

Distribution of power. Much research indicates that it is usually misleading to think of the power structure of an organization. Structural arrangements differ according to activity. The distribution of power over personnel policies may vary considerably from that over public relations. If so, it is important to ascertain whether power is more likely to be concentrated in some areas than in others, and whether an individual actor in the organization can transfer some kinds of power to other areas more easily than he can transfer other kinds. Thus far, the answers to both questions seem to be affirmative and similar. There appears to be a remarkably general tendency for most organizations—political, business, social —to concentrate power most narrowly in three basic areas: (1) control over major resource allocations; (2) control over high-level personnel decisions; and (3) control over important interactions with critical external actors, such as main competitors or allies.

Concentrations of power in organizations

Other analytic structures of organizations. Power is not the only analytic structural aspect of an organization. Patterns of prestige, communication, wealth, friendship, and so on can also be distinguished. A major theoretical concern is to comprehend the relationships among these analytic structures. The most conspicuous finding to date would seem to be the tendency toward isomorphism, or similarity, among these various structures. It appears difficult, for example, to maintain a highly democratic power structure in the face of a highly concentrated prestige or status structure. Great inequalities of wealth seem to be incompatible with marked decentralization of power. Extremely limited and centralized communication structures apparently are conducive to similarly concentrated power structures. Conversely, scattered communication patterns are generally incompatible with centralized, rapid, decisive power reactions. The reason for these agglutinative tendencies is obvious: prestige, wealth, and information are all resources that can contribute to an actor's power in most situations. Power patterns are formed in

part by the distribution of these other relations, just as they in turn are partly shaped by existing power patterns.

Communities. Power in various types of communities—villages, towns, cities, and metropolitan areas—has been extensively studied by researchers in two broad fields. Political scientists and sociologists have conducted dozens of studies, usually of fairly sizable communities located in the United States or, occasionally, in Europe; and anthropologists have made hundreds of case studies of peasant and primitive villages scattered around the world, especially in nontemperate-zone climates. The political science and sociological studies usually focus primarily and explicitly on power and stratification, while the anthropological literature is usually concerned with power peripherally and casually. Because both literatures suffer serious methodological difficulties, comparative conclusions are extremely doubtful.

The discussion that follows pertains almost entirely to studies conducted in the United States by political scientists and sociologists. Most of the literature has been concerned with prevailing power structures of communities, especially with the internal distribution of power. Of paramount interest have been questions of whether communities differ in their distributions of power, and if so, why. Somewhat less effort has also been directed toward understanding the consequences of differences in the distribution of community power for community policies.

Power structures of American communities

Fragmentation of formal power. Despite sharp methodological controversies, studies of power in American communities have revealed some surprising agreements in their findings. In contrast to the situation in many other countries, formal political power in most American communities is highly fragmented. The numerous boards, agencies, districts, councils, committees, associations, and offices share important blocks of community authority. In addition, there are federal, state, and local divisions of power. To get any single positive action through such a fragmented structure minimally requires agreement from many diverse governmental entities. The structure of formal authority thus seems to discourage major community initiatives and to provide an advantage to those favouring the status quo.

This fragmentation of formal power has been at least partly compensated for by informal power arrangements, such as the "machine," political parties, popular movements of various kinds, and informal leadership groups to "get things done." Community studies have noted, however, that the "reform movement" in urban America may have undermined these informal mechanisms to the point that establishing effective policy on basic problems has become structurally almost impossible to achieve in many communities. In any event, the fragmentation of formal power in municipal America has been widely observed.

Political–economic split. Many, but by no means all, studies have found some sort of "bifurcation of power" in American communities. Many investigations depict two main constellations of power in communities—political and economic. Sometimes one seems to be dominant and the other supporting, and sometimes the two appear as virtually independent realms. In the latter instance, economic interests remain largely aloof from politics, at least refraining from direct intervention; while politicians shun any major incursions into economic affairs. Several relatively sophisticated analyses suggest that important community initiatives are most likely to come from the staffs of governmental and civic organizations or their professional associates, while business leadership often plays a prestigious supporting role that gives it a marked veto power over many proposed projects or initiatives.

Power of elites. Very few investigations have found power within any issue area to be widely distributed. The bulk of the populace seems generally to have little influence on and little interest in most public policy. At best, the picture is one of competition among different small elites of power, with popular pressure being exerted only sporadically and very indirectly; perhaps most of the popular pressure comes largely through the reactions that power holders themselves, rightly or wrongly, anticipate from the electorate. The main controversy in the community-power literature has been over the degree of elite competition that typically exists. Little disagreement has occurred over the relative quiescence and ineffectiveness of the mass of the populace. At best, activated minorities are seen to predominate; at worst, a small power elite is perceived to rule.

Such studies frequently are based on the examination of issues that have arisen in the community. Critics contend, however, that an elite may be so powerful that it permits only the issues acceptable to it to surface in the community. In such instances, therefore, the elite does not have to display its power by intervening. The elite, in other words, is most effective when it can keep threatening items off the political agenda. One effective way to do this is through socialization—through spreading a political culture, much of it unrecognized by those who accept it, that supports and enhances the power of the dominant elite.

Factors shaping community power structures

If the essential correctness of each of the various studies of community power be assumed and the total literature then inspected for interesting empirical findings, certain conclusions emerge. Associations between the distribution of power and the following factors are suggested:

1. The relationship between community size and power does not seem to be linear, for there are many indications that the largest cities of the country tend to have relatively less concentrated power structures. Pluralistic, less concentrated patterns also seem to be found more often in communities that are growing less rapidly in size, while more concentrated patterns are more likely in rapidly growing communities.
2. It has frequently been asserted that economic diversity in a community is positively associated with a greater distribution of power. Certainly, several egregious examples of one-industry cities dominated by a single family have been presented. Other, more comparative, research, however, casts doubt on this assertion, so that any definitive conclusions must await better evidence.
3. At least one regional variation has been regularly reported: cities in the South seem to have more concentrated power structures than similar cities in other areas.
4. It has also been alleged, and is theoretically plausible, that the more heterogeneous the population of the community, the more pluralistic its power structure. This is especially true when cleavages have crystallized into established, competitive social groupings. Again, however, strong support for this contention is lacking.
5. Existing data seem to indicate that nonpartisan elections, city-manager forms of government, and other "reform" oriented arrangements are associated with more concentrated power patterns, while partisan elections, mayoral systems, and more institutionalized political conflict are linked with pluralistic patterns.
6. Finally, there seems to have been a movement in the direction of reduced concentration of power. Perhaps because of the increasing urbanization of the country, perhaps because of changes in its political culture, structures of community power seem to be somewhat more pluralistic than at the turn of the century. Some writers also note a tendency for power holders to be less "blatant," arrogant or obvious in their use of power.

BIBLIOGRAPHY. The literature on power is enormous. General treatments include: BERTRAND RUSSELL, *Power: A New Social Analysis* (1938); CHARLES E. MERRIAM, *Political Power: Its Composition and Incidence* (1934); HAROLD D. LASSWELL and ABRAHAM KAPLAN, *Power and Society* (1950); ROBERT M. MACIVER, *Power Transformed: The Age-Slow Deliverance of the Folk and Now the Potential Deliverance of the Nations from the Rule of Force* (1964); and ADOLF A. BERLE, *Power* (1969).

Many of the most valuable articles on the concept of power are assembled in RODERICK BELL, DAVID V. EDWARDS, and R. HARRISON WAGNER (comps.), *Political Power: A Reader in Theory and Research* (1969). One additional article that is essential is JAMES G. MARCH, "The Power of Power," in DAVID EASTON (ed.), *Varieties of Political Theory* (1966). An attempt at logical explication is INGMAR PORN, *The Logic of Power* (1970).

A few works offering explicit discussions of power are: FELIX E. OPPENHEIM, *Dimensions of Freedom* (1961); DORWIN CARTWRIGHT (ed.), *Studies in Social Power* (1959); MAX WEBER, *From Max Weber: Essays in Sociology*, ed. by H.H. GERTH and C.W. MILLS (1947); and FREDERICK FREY, "Developmental Aspects of Administration," in J.P. LEAGENS and C.P. LOOMIS (eds.), *Behavioral Change in Agriculture* (1971).

Basic works on power in small groups are: SIDNEY VERBA, *Small Groups and Political Behavior* (1961); TERENCE HOPKINS, *The Exercise of Influence in Small Groups* (1964); GEORGE C. HOMANS, *The Human Group* (1950); BARRY E. COLLINS and HAROLD S. GUETZKOW, *A Social Psychology of Group Processes for Decision-Making* (1964); ROBERT T. GOLEMBIEWSKI, *The Small Group: An Analysis of Research Concepts and Operations* (1962); JOHN W. THIBAUT and HAROLD H. KELLEY, *The Social Psychology of Groups* (1959); DORWIN CARTWRIGHT and ALVIN ZANDER (eds.), *Group Dynamics: Research and Theory*, 2nd ed. (1960); PAUL HARE *et al.* (eds.), *Small Groups* (1955); and JAMES D. BARBER, *Power in Committees: An Experiment in the Governmental Process* (1966).

Among the most significant works in community power studies are: NELSON W. POLSBY, *Community Power and Political Theory* (1963); ROBERT A. DAHL, *Who Governs? Democracy and Power in an American City* (1961); EDWARD BANFIELD, *Political Influence* (1960); ROBERT E. AGGER, DANIEL GOLDRICH, and BERT SWANSON, *The Rulers and the Ruled: Political Power and Impotence in American Communities* (1964); FLOYD HUNTER, *Community Power Structure* (1953); ANDREW MCFARLAND, *Power and Leadership in Pluralist Systems* (1969); ELIHU KATZ and PAUL F. LAZARSFELD, *Personal Influence: The Part Played by People in the Flow of Mass Communications* (1955); and TERRY CLARK (ed.), *Community Structure and Decision-Making: Comparative Analyses* (1968).

The literature on power in organizations is extremely large. A good starting point is the compendium by JAMES G. MARCH and HERBERT A. SIMON, *Organizations* (1958); or JAMES G. MARCH (ed.), *Handbook of Organizations* (1965). Other major titles are: CHESTER BARNARD, *The Functions of the Executive* (1938); HERBERT A. SIMON, *Administrative Behavior* (1957); ROBERT A. GORDON, *Business Leadership in the Large Corporation* (1961); WILBERT MOORE, *The Conduct of the Corporation* (1962); RICHARD S.F. EELLS, *The Government of Corporations* (1962); ROBERT L. KAHN and ELISE BOULDING (eds.), *Power and Conflict in Organizations* (1964); JAMES D. THOMPSON *et al.* (eds.), *Comparative Studies in Administration* (1960); HAROLD LEAVITT (ed.), *The Social Science of Organizations: Four Perspectives* (1963); AMITAI ETZIONI, *A Comparative Analysis of Complex Organizations* (1961); ROBERT L. PEABODY, *Organizational Authority* (1964); PETER M. BLAU and W. RICHARD SCOTT, *Formal Organizations: A Comparative Approach* (1962); and PHILIP SELZNICK, *Leadership in Administration: A Sociological Interpretation* (1957).

(F.W.F.)

Political Science

Political science is most generally understood to mean the systematic study of government processes by the application of scientific methods of analysis. More narrowly and more traditionally, it has been thought of as the study of the state and of the organs and institutions through which the state functions. In most countries, political science is thought to be a single discipline, but the plural form has been used in France, as in the name of the École Libre des Sciences Politiques (now Institut d'Études Politiques de l'Université de Paris), founded in 1871—although there is also an Association Française de Science Politique. Speculation about political subjects is not unknown in ancient non-Western cultures, but most students agree that the roots of political science are to be found in the earliest sources of Western thought, especially in the works of Aristotle, who is recognized by many as the founder of political science.

THE NATURE AND SCOPE OF POLITICAL SCIENCE

Although political science may be distinguished from political philosophy, the distinctions are unsatisfactory inasmuch as they lack categorical rigour. In the most usual distinction, political philosophy is thought to be concerned primarily with the study of political ideas, often within the context of their times. It is strongly normative in its thrust and disposition and rationalistic in its method. Political science, however, concerns itself with institutions and behaviour, eschews normative judgments as much as possible, and attempts to derive principles from objective facts with as much quantification as the evidence will allow. Political philosophy thus speculates about the place and order of values, the principles of political obligation (why men should or should not obey political authority), and the nature of such terms as right, justice, and freedom. Political science, on the other hand, seeks to establish by observation (and, if possible, by measurement) the existence of uniformities in political behaviour and to draw correct inferences from these data. The stated differences between political philosophy and political science are less than is sometimes supposed, for the most empirical scholar in both the social and natural sciences makes use of unproved postulates, hunches, and intuitions; and the most rationalistic philosopher employs conceptions that embody empirical statements.

Some theorists, however, who believe it possible to develop a completely value-free science of politics insist that the distinction between political philosophy and political science is not faint but vivid. It is their opinion that only a few hundred philosophers and political theorists have ever contributed to systematic speculation about and study of politics. It was said in 1966 by one of the exponents of this view that "probably two out of every three political scientists who have ever lived are alive and practicing today." In this view, political philosophy is believed to be addicted to obscurity and opacity of statement—an effort to bespeak the unspeakable—and it is thought that its task should be a more modest one, namely, to grope with the grammar of philosophical statements. This would require political philosophers to put their intelligence to the elucidation of the language of politics and to expose the difficulties placed by language in the consideration of matters of fact. One of those interested in this approach was the English political philosopher T.D. Weldon in *The Vocabulary of Politics* (1953), but few have followed his initiative.

Scientific status of the subject

The question as to whether political science is a science is largely inconsequential, because the problem is primarily one of definition. If the term science is to be applied to any body of systematically organized knowledge based on facts ascertained by empirical methods and described by as much measurement as the material allows, then political science is a science, just as are the other social disciplines. If, on the other hand, the term science is to be limited to those disciplines in which the scholar can control the materials to be studied and can perform experiments that others can reproduce under the same conditions and in which predictability is possible, then the label is less appropriate, although not entirely misapplied. The American political economist Thorstein Veblen denied that political science was anything more than a "taxonomy of credenda," and, more recently, a British writer, Bernard Crick, has said that the hope of creating an artificial science of politics on natural principles, although not originally and uniquely American, has been largely generated and sustained by two aspects of the American culture—an agreement on liberal doctrine that has made politics less a matter of serious doctrinal splits than of mere disagreement among partisans of the same creed, and a general national preoccupation with technology.

DEVELOPMENTS IN THE 19TH AND EARLY 20TH CENTURIES

Influence of early sociology. The origins of contemporary political science are to be found in the enthusiasm for the creation of social science that was widespread in the 19th century, an enthusiasm stimulated by the rapid growth of the natural sciences. It might be said that one starting point for the development of modern political science is the work of the Comte Henri de Saint-Simon, a notable Utopian Socialist, who in 1813 suggested that morals and politics could become "positive" sciences; that is, disciplines whose authority to command belief would rest not upon subjective preconceptions but upon objective evidence. With him worked the mathematician and philosopher Auguste Comte, the two collaborating in the publication in 1822 of the *Plan of the Scientific Operations Necessary for the Reorganization of Society*, which

argued, among much else, that politics would become social physics and that the purpose of social physics was to discover unchanging laws of progress. Out of this collaboration emerged the law of the three stages through which knowledge had to pass—the theological, the metaphysical, and the positive—that Comte was to establish as the theme of the science of social physics, a study he came to name sociology. An intellectual connection between political science and sociology was thus early established in schemes of political and social regeneration and reform, although political science was thought to be limited to only one form of association in society, namely, the state. Comte thought that the principal methods for the study of social phenomena were observation, experiment, and abstraction. Although one might have thought that politics could not be an experimental science, Comte was of the view that political experimentation did take place whenever there was a change in the life of the state, intended or not. It must be said, however, that even on this account there is no close similarity to experimentation in chemistry and physics, in which all the variables can be controlled.

In the search for more objective methods of inquiry into political and other social phenomena in the 19th century, contributions to the explanation of the state were supplied by several new intellectual disciplines. Because political science deals with some aspects of human behaviour, for example, it is closely allied to other social sciences that also deal with human behaviour. Long before the development of scientific inquiries in the 19th century, numerous theories of the state had drawn inspiration from the human being as model, as in the *Policraticus* of John of Salisbury (1159), in which the physiology of the body and that of the state are compared; or in *The Republic* of Plato, in which the elements of the human personality prefigure the class structure of the state; and in Rousseau's *Contrat Social*, in which the political order is animated by a general will (will being a human attribute). The positivism of the 19th century, however, brought new approaches to the study of the state, although the older ones continued to coexist with them.

Sociological analyses of the state

Among those following Comte was the Polish-born sociologist Ludwig Gumplowicz, who built a sociology on Comtean foundations but who owed much also to Darwin, to the social Darwinist Herbert Spencer, and to contemporary anthropology. In Gumplowicz' view, the earliest forms of group life were small hordes bound by consanguinity, which developed into matriarchies and patriarchies. He supposed the existence of a social-evolutionary process characterized by conflicts between autonomous groups and by conflicts of interest within those groups. The product of this process was the state, founded on force and maintained by power.

Gumplowicz and several other 19th-century political sociologists anticipated 20th-century concerns in political science with the significance of groups, the nature of interests, the role of parties as interest groups, and the social context within which political events occur.

Still another 19th-century writer with some precedental connection to the political science of the 20th was the Italian Vilfredo Pareto. Although he lived and wrote in both the 19th and 20th centuries, he may be counted in the earlier period because of his advocacy of the "logico-experimental" approach to sociology, which involved observation and logical inference. Pareto had no direct influence on the development of political science, but in two respects his sociology had implications for the developing discipline. First, his was a psychological sociology, and much of his concern was with the influence of beliefs, attitudes, opinions, and sentiments in shaping social life. This anticipates the 20th-century approach of many political scientists who regard psychology as the most important adjunct of the scientific study of politics. Second, Pareto thought of society as a system always tending toward equilibrium, and the conception of politics as a system was to mark much of academic political science after World War II (see below). Also to be mentioned in the context of important 19th-century sociological theories is the work of a Swedish political scientist, Rudolf Kjellen, whose systematic treatment of the state as a fusion of organic and intellectual moral elements in an ambience of geographical determinism led to a theory of politics that he termed geopolitics.

Juristic influences. In addition to sociology, another discipline with which the development of political science has been historically related is law. A close connection between the state and law was made in the 16th century in the French political philosopher Jean Bodin's theory of sovereignty, which supposed the necessary existence in each state of an authority to make the law. This theory gave rise to numerous juristic theories of the state, especially among German publicists of the 19th century who sought to define the nature of federation and empire. Although the doctrine of sovereignty made most sense as a statement of the power of monarchs to make the law in simple unitary (nonfederal) systems, earnest writers forced the facts of federation and empire to fit the theory and often ignored facts that did not. The Bodinian view invested sovereignty with attributes of singularity and omnipotence—that is, sovereignty was viewed as being indivisible and absolute—but theoretical efforts were made in Germany before 1871, notably by Georg Waitz, to establish federal theory on a division of sovereignty between the centre and the member states. After 1871, the theory of Max von Seydel, namely, that since sovereignty is indivisible, it cannot be divided and must rest in a single location—either in the constituent members of a federal system or in the centre—supplanted that of Waitz. Georg Jellinek, an Austrian, attempted to resolve the paradox that so-called sovereign states are in fact limited by constitutions and laws and by membership in the family of nations by arguing that the restrictions are autolimitations. That is to say, since by definition there is no power above the state, if in fact there are limitations it must be the state that created them. The state therefore continues to be omnipotent. In France, juristic theories of the state also persisted strongly because the teaching of political science was conducted in the law schools as "constitutional law." Efforts were made by some professors, however, to broaden the legal approach, notably by Léon Duguit, who attempted a sociological positivistic treatment of juridical rules for the limitation of the state, and Maurice Hauriou, who contributed a theory of institutions. In England the emergence of political science as a subject was recognized in the establishment of the London School of Economics and Political Science in 1895 and by the founding of a separate chair of politics at Oxford in 1912.

The doctrine of sovereignty

Developments in the United States. The enthusiastic development of social sciences in the 19th century, stimulated as it had been by the rapid growth of the natural sciences, reinforced an existing interest in politics in the United States and created a generation of distinguished American political scientists. There had, in fact, been much interest in the teaching of political subjects in American colleges and universities well before the 19th century. Political science in the United States, however, free of any connection with moral philosophy, fusion with history, or submergence in political economy, may be said to date from 1880, when John W. Burgess, after studying at the École Libre des Sciences Politiques, succeeded in establishing a separate school of political science at Columbia University. Although political-science faculties increased in numbers after 1900, the growth was uneven, and in some major institutions separate departments were not created until after World War I.

The development of American political science in the last quarter of the 19th century was influenced by the experience of numerous scholars who had done graduate work at German universities in which political science was taught as *Staatswissenschaft* ("science of the state") in an ordered, structured, and analytical organization of concepts, definitions, comparisons, and inferences. To modern readers the work of these men often seems somewhat formalistic and institutional in tone and focus. It did represent, however, an effort to establish an autonomous discipline, separate from history, moral philosophy, and political economy. Some of the new American political

scientists, such as Woodrow Wilson and Frank Goodnow, also showed in their writing an awareness of such new intellectual currents as theories of evolution and turned their attention to an examination of American institutions that, in the United States Constitution, had originally been based on the admiration of much of the 18th-century world for the harmonious perfections of mechanics. For Wilson, it was to be Darwin, not Newton, who would provide the inspiration for a transformation in American political science from the study of static institutions to the study of social facts, more truly in the positivist temper, less in the analytic tradition, and more oriented toward factual realism.

Bentley's early work on group processes

Little-noticed at the time he wrote *The Process of Government* (1908), Arthur F. Bentley was to have a celebrated influence on the development of American political science in the 1930s and the 1950s. Although his influence was not immediately felt, he sounded certain themes that became the orthodoxy of the new science of politics after World War II. First, he rejected all metaphysics and normative formulations as "spooks" and "mind-stuff" and insisted that the proper study of politics was observable fact, in imitation of the natural sciences. Second, his basic concept was the "group," rejecting thereby all previous formulations of the subject matter of political science that centred on the "state." Third, the "raw material of government," he thought, was the activity of men and the processes through which this activity flowed in legislation, administration, and adjudication. Behaviour and process were to become the focus of much of the interest in political science in the 1950s, after Bentley's single work on politics was given new currency just before the outbreak of World War II. He was considerably before his time in presaging the end of preoccupation with institutional and descriptive analysis. He acknowledged the leadership of Gumplowicz, who, in his view, had "taken the most important step toward bringing out clearly the nature of the group process," discarding the "individual as a causal factor in society," and insisting "that all social movements are brought about by group interaction."

Although the effort of Bentley to develop an objective, value-free analysis of politics had no initial consequence, other movements toward this goal were more immediately successful. The principal impetus was provided by what became known as the Chicago school in the mid-1920s and thereafter. The leading figure in this movement was Charles E. Merriam, who in 1925 published *New Aspects of Politics*, a book that argued for a reconstruction of method in political analysis, urged the greater use of statistics in the aid of empirical observation and measurement, and postulated that out of the converging interests of politics, medicine, psychiatry, and psychology might come "intelligent social control." The basic political datum for Merriam at this stage of his thinking was "attitude"; hence his reliance upon the insights of psychology for a better understanding of politics. These ideas were not entirely new, since Graham Wallas, an Englishman, had said in *Human Nature in Politics* (1908) that a new political science should be based upon quantitative methods and that serious attention should be given to the psychological elements ("human nature") in political activity, including nonrational acts and the exploitation in political life of subconscious nonrational inferences. The American political scientist and journalist Walter Lippmann had expressed much the same view in *Public Opinion* (1922). One of those in the Chicago group who carried the connection between politics and psychology quite far was Harold Lasswell, in his *Psychopathology and Politics* (1930). In *Power and Personality* (1948) he fused the Freudian categories of the earlier work with subsequent writings on power.

Influential works of Merriam and Lasswell

These two leading expositors of the Chicago school, Merriam and Lasswell, both published books at about the same time that gave a central place to the phenomenon of power in the empirical study of politics. Merriam published *Political Power* in 1934 and Lasswell *Politics: Who Gets What, When, How* in 1936. Merriam undertook to show how power came into being, to describe what he called the credenda, miranda, and agenda of authority (which he tended to equate with power), the techniques of power holders, the defenses available to those over whom power is wielded, and the dissipation of power. Lasswell's 1936 work was a naturalistic description of "influence and the influential." Although both were cast in the empirical mode, the second was more successful in this regard than the first, which tended to be abstract and rhetorical. A truly empirical work of the Chicago school that had considerable significance in the development of academic political science was Charles E. Merriam and Harold F. Gosnell's work, published in 1924, on *Non-Voting, Causes and Methods of Control*, which used sampling methods and survey data. Since then, certainly one of the most successful achievements in empirical political science has been the study of voter behaviour and election results. Although members of the Chicago school insistently professed an interest in value-free political science, they were characterized by two normative predilections—their acceptance of the values of the democratic system and their earnest attempts to improve it through their writings.

By 1945 political science in the United States was much more than the concern for institutions, law, formal structures of public government, procedures, and rules that it had earlier been. There was by then also a considerable body of writing on processes—on the dynamics as well as the statics of public governance. There were works on pressure groups and lobbies, on the "invisible government" operating behind public authority, on actual bureaucratic processes as distinguished from the rules of administrative procedure, on bosses and political parties, and on ethnic influences in the behaviour of the electorate. All of these works signified a turning away from formality and the development of a growing interest in the factual realities of political behaviour.

Developments in Europe. Outside the United States the development of political science was less quantitative and behavioural. In France the main tradition in the study of political institutions was still legal rather than sociological. Constitutional law and history were abundant, and politics in the main was viewed from legal perspectives, though not entirely so. As early as 1913 one writer, André Siegfried, had introduced the geographical and historical study of elections in *Tableau politique de la France de l'Ouest sous la troisième République*, and the development of French sociology in the works of Émile Durkheim and others had applications in political science.

In England, although there was little development of a quantitative science of politics, a substantial contribution was made toward the formulation of political philosophies that centred attention upon the group basis of politics. Particularly notable were the works of J.N. Figgis, G.D.H. Cole, Ernest Barker, and Harold Laski. In Sweden, Herbert Tingsten in his work *Political Behaviour: Studies in Election Statistics* (1937) gave currency in the title to what was to be the main development in political science after World War II.

CONTEMPORARY POLITICAL SCIENCE

Behavioralism. In American political science since the end of World War II, the behavioral persuasion has been the dominant one. A former president of the American Political Science Association has attributed the rapid development of the behavioral approach to six causes: the inspiration of the Chicago school; the immigration to the United States in the 1930s of large numbers of European scholars (particularly Germans) with backgrounds in European sociology, who stressed the relevance of sociology to politics; the movement of many political scientists into administrative and political positions during World War II; the influence of foundation support in the encouragement of research in political behaviour; the increasing development of the survey method in certain political studies, such as voter behaviour; and the missionary work of the Social Science Research Council under leadership sympathetic to behavioralism.

Etiology of the behavioral approach

Although the term behavioralism has been freely used in political-science writings, there is in fact confusion as to

whether it is a field of study, a method, or an approach. One American political scientist, Heinz Eulau, in *The Behavioral Persuasion in Politics* (1963), has said that the behavioral persuasion "is concerned with what man does politically and the meanings he attaches to his behavior," and he has suggested that researchers cannot afford to get tangled up in problems of definition. Another American, Robert Dahl, has said that it is a mood or even "the scientific outlook." The term behavioral, then, may be merely a term having distinctiveness, weight, and value for a certain time only, since it seems primarily to signify that phase in the quarter century after World War II during which there was a significant revival of interest in empirical studies in politics, a movement strong enough to establish at least a partnership with the traditional approaches, although some of its advocates have gone so far as to say that their science has made traditional approaches outdated.

Systems analysis. Contemporary political science does not display as much coherence as earlier modes of inquiry (coherence was one of their principal virtues), but it does exhibit a refreshing complexity, of both concept and method. Although much effort has been spent on the production of empirically derived evidence in a scatter of fields and a diffusion of subjects and although critics have scoffed at studies that they have said only prove the obvious, the product of 25 years of behavioralism has been voluminous and often stimulating in quality. Common in this output is the assumption that politics is process, the ceaseless interaction of individuals and groups on each other in a flow of activities in and around public governance. The most commanding concept devised to fit this flow with form has been the concept of the system. In this view the focus of political science is not the individual in solitude nor the multitude nor even the group. The American political scientist David Easton's *The Political System* (1953), a seminal work in empirical political theory, conceived of the political system as part of the total social system. For Easton the political system comprised all those activities having to do with the formulation and execution of social policy; that is, "policy-making process." The identifying criterion of political behaviour for Easton was the "authoritative allocation of values" for the society, by which he distinguished his sense of the subject matter of political science from that of Lasswell, who had argued that political science was concerned with the distribution and content of patterns of value throughout the society. In this conception of system, which is not unrelated to the conception of system in physics and biology, linkages are found between the system and its environment. Inputs (demands) flow into the system and are converted into the outputs (decisions and actions) that constitute the authoritative allocation of values for the society; that is, the distribution of rewards in wealth, power, and status that the system may provide.

Variants of systems theory

There are various versions of systems theory, although all suppose the existence of the system and more or less consciously pattern it after systems models in other disciplines. There has thus been developed a view of systems, based upon engineering models and ideas, that rejects the view of Easton that there can be a general theory of political systems. Another theorist, drawing on cybernetics, has viewed the political system as a communications net. Although expositors of systems concepts generally suppose they are working with scientific theory, critics have said that they are merely creating new taxonomies. What is certain is that systems theory has introduced new and—to more traditional scholars in the field—unusual terms into the vocabulary of politics. Thus, instead of such terms as the state and sovereignty, the new idiom speaks of systems, inputs and outputs, feedbacks, circular loops, networks, legitimacy symbols, information storage and retrieval, political socialization, interest articulation and aggregation, cluster blocs, zero-sum games, macropolitics and micropolitics, and much else drawn from the concepts and languages of other sciences and from statistics, with hopeful assertions about the possibility of making politics predictable.

Although decision making as a research field in political science may be analytically regarded as part of the process by which inputs are converted into outputs in systems theory, it also has an origin independent of systems theory. Interest in game theory provided a stimulant to studies of decision making, which already had an established place in the lore and lexicon of politics and administration. In decision theory the paradigm is the rational actor, and the research problem is the ascertainment of the most efficient means to effect given goals, which may or may not be rational. Investigators into decision making are generally less concerned with the system as a whole than with the activity that they regard as the most important part of it, namely, the way in which decisions are or should be made according to rational calculation. A body of writing on decision making by judges has earned a sub-classification as "jurimetrics."

Interest groups, elites, and political parties. Studies of interest groups, elites, and political parties also have their own independent origins, although they, too, have been brought within the framework of systems analysis. Interest groups and political parties, for example, have been described as agencies for the articulation and aggregation of interests, which in turn provide inputs (demands) for the political system to convert into outputs (decisions and actions). But interest-group analysis antedates the advent of behavioralism. The modern concern for the subject starts perhaps with studies of prohibition and other pressure groups during the 1920s. More generalized and theoretical treatments of interest groups and political parties resulted in part from a revived interest in the work of Bentley in the 1930s and 1950s. The study of elites was begun at least as early as 1936 (Lasswell's "influentials"), and it came to the forefront in the 1950s in community studies made by political scientists in Atlanta, Chicago, New York, New Haven, and elsewhere. Such studies had long been familiar in sociology, but they acquired special significance in political science, because democratic values seemed threatened by the possible existence of elites. Despite declarations of their concern to establish a value-free science, American political scientists have thus tended to cleave to the traditional democratic ethic, in which elites are presumed to have no place.

Analysis of political attitudes and voting behaviour. What some have called the behavioral revolution had its greatest successes in the analysis of public opinion, political attitudes, and electoral behaviour. Especially in the period after World War II, the refinement of statistical techniques in public-opinion polling, the analysis of voter behaviour, and the development of new research concepts have brought the study of opinions and attitudes closest to the goal of the scientific outlook and some considerable distance from the mark made by Merriam in the 1920s. The Survey Research Center at the University of Michigan has become an important national centre for the collection of data on elections and voter behaviour.

CURRENT TRENDS

The effort to establish a value-free, objective political science in the United States in the two decades after World War II has won what is doubtless a permanent place for the scientific study of politics, but it it has also bred a critical reaction. In the late 1960s opponents of "scientism" rejected what they felt was the increasing subjection of spontaneity and human values to determinisms in every aspect of life and argued that political science was an example of the pervasiveness of technology and of a search for rationality in a social complex that might be irrational and out of control. Although political science had developed skillful and sensitive techniques for the quantification of data termed political, there is no orthodoxy on the scope of political science nor on the delimitation of separate areas of research. Quite different outcomes emerge from the basic assumptions as to the focus of the discipline, whether it be power, government, system, process, decision making, or policy formulation. The failure of the discipline to settle the question of identity has made for a creative and inventive exploration of many avenues of research, but the achieve-

Reactions against "scientism"

ment of a unified general theory of political behaviour on which there is common consent still lies ahead.

Even outside the United States, non-normative political science is not extensive, although it is not unknown. England has made many creative contributions to political theory and law throughout its history, as have other European countries, but it has also produced substantial works that can be classified as positive (*i.e.*, non-normative) political science. Among these are the studies of R.T. McKenzie and D.E. Butler in the field of political parties, voter behaviour, and pressure groups, S.E. Finer on interest groups, and U.W. Kitzinger on German elections. Notable work has been done in France in the field of political parties by Maurice Duverger and François Goguel. Work has also been done on the political socialization of children in French schools by Charles Roig and François Billon-Grand. In Denmark, beginnings were made in the systematic study of political science in 1959 with the founding of the Institute of Governmental Studies at the University of Århus. In Finland the first extensive studies on voting behaviour appeared in 1956.

In Japan there was a flowering of social sciences after World War II, but political science at first did not grow at the same pace as other social sciences because of lack of agreement about both subject matter and method, a difficulty felt from the beginning by the man recognized to be the founder of Japanese academic political science, Onozuka Kiheiji, who published *Principles of Political Science* in 1903. With the opening of the behaviour of public officials and institutions to scientific scrutiny, however, there has been much political inquiry that would qualify as political science, including, for example, systems analysis and works on political culture, political development, and process and behaviour.

Political science in Communist countries

Although in the past the objective study of political subjects by researchers in Communist regimes has been difficult, if not impossible, a somewhat more permissive policy in some countries has led to what may be the beginnings of scholarly political science. The most advanced political science is to be found in Poland and Yugoslavia, and Romania and Czechoslovakia have come to recognize political science as a discipline. The Soviet Union does not yet sanction political science, but scholars do conduct empirical research, which was endorsed in 1962 by the U.S.S.R. Academy of Sciences under the term "concrete sociological investigations." The greatest movement has been in the conduct of public-opinion polls using advanced Western techniques. Interest in the development of political science was evidenced in the publication in 1969 of *Politicheskaya nauka v SShA: Kritika burzhuaznykh Kontseptsy vlasti* ("Political Science in the U.S.A.: A Critique of Bourgeois Conceptions of Power"), by V.G. Kalensky. Although there is no officially sanctioned political science, there is a Soviet Association of Political Sciences, which has sent delegates to the meeting of the International Political Science Association. Despite its title, the Soviet Association of Political Sciences is most heavily oriented toward the state and the law, and its members have been critical of what they regard as the anti-Marxist bias of bourgeois political science. One of its members, however, F.M. Burlatsky, published a major article in *Pravda* in 1965 calling for the establishment of a genuine political science in which the findings would emerge from the data.

In Yugoslavia a political-science association was established in 1951, and in 1962 a faculty of the political sciences was established at the University of Zagreb. In Czechoslovakia a political-science association was formed in 1965 and became a member of the International Political Science Association; and in 1968 a political-science association was formed in Romania. Political science in Yugoslavia has tended to centre on traditional divisions of the discipline, such as political theory, comparative government, and international relations. In Poland, political science has centred on the study of political behaviour, on community power structures, on voter behaviour, and on public opinion. Techniques of considerable sophistication have been employed.

Political science is one of the means by which people seek to understand the human condition and man's fate. Or, as Aristotle believed, politics is the most important of human activities, and the sovereign science. For 24 centuries, at least, the greatest intellects and scholars have striven to state the universal elements of just order in human affairs. None has succeeded in this, so far, hopeless ambition, although most have contributed some special insight and added to the common wisdom. The effort to achieve magisterial comprehension will doubtless continue; and the search will change direction as experience requires, with the aid that new perceptions, concepts, and methods will provide. Progress toward establishing general laws, however, may never be as steady or as swift in political science as it has been in laboratory sciences like physics; for, as Albert Einstein once said, politics is more difficult than physics.

BIBLIOGRAPHY. Although works of classical political philosophy are both venerable and extensive, few of them qualify as modern political science because they are neither quantitative nor, in most respects, even empirical in tone and temper. ARISTOTLE'S *Politics* and MACHIAVELLI'S *The Prince* come closest to meeting empirical standards because the first is based in part on the study of numerous constitutions in the Attic world and the second has the reputation of having divorced politics and ethics, although it was in fact written to achieve normative ends. AUGUSTE COMTE, *Cours de philosophie positive*, 6 vol. (1830–42; Eng. trans., *The Positive Philosophy of Auguste Comte*, 2 vol., 1853), and *Système de politique positive*, 4 vol. (1851–54; Eng. trans., *System of Positive Polity*, 4 vol., 1875–77), are seminal statements in the 19th century on a science of society. LUDWIG GUMPLOWICZ, *Grundriss der Sociologie* (1885; Eng. trans., *The Outlines of Sociology*, 1899; 2nd ed., 1963); and GUSTAV RATZENHOFER in *Wesen und Zweck der Politik*, 3 vol. (1893), argue the case for the primacy of groups in studies of the state. A useful summary statement of the sociologies of the 19th century is NICHOLAS S. TIMASHEFF, *Sociological Theory: Its Nature and Growth*, 3rd ed. (1967). A good general work on the efforts of German jurists in the 19th century to cope with the facts of federalism is RUPERT EMERSON, *State and Sovereignty in Modern Germany* (1928). Among Americans who either studied in Germany in the 19th century or showed the influence of new European thought in their work were THEODORE D. WOOLSEY, *Political Science; or, The State Theoretically and Practically Considered*, 2 vol. (1877); JOHN W. BURGESS, *Political Science and Comparative Constitutional Law* (1891); WOODROW WILSON, *The State: Elements of Historical and Practical Politics* (1889), and *Congressional Government* (1885, reprinted 1961); and FRANK J. GOODNOW, *Politics and Administration: A Study in Government* (1900, reprinted 1967). An Englishman of the same period who brought a realistic inspiration to the study of American institutions was JAMES BRYCE, *The American Commonwealth*, 3 vol. (1888; abridged ed., 1959). An excellent statement of the content of political-science teaching in American colleges and universities, and spanning almost three centuries, is ANNA HADDOW, *Political Science in American Colleges and Universities, 1636–1900* (1939, reprinted 1969). The most notable precursor of the behavioural approach in the 20th century was ARTHUR F. BENTLEY, *The Process of Government: A Study of Social Pressures* (1908, reprinted 1949). Others were GRAHAM WALLAS, *Human Nature in Politics*, 4th ed. (1962); and WALTER LIPPMANN, *Public Opinion* (1922; paperback edition, 1965). Besides works of the Chicago School mentioned in the article, the following may be noted: CHARLES E. MERRIAM, *Chicago: A More Intimate View of Urban Politics* (1929, reprinted 1970), an early contribution to the realistic study of urban government; LEONARD D. WHITE, *The Prestige Value of Public Employment in Chicago* (1929), a study in popular attitudes; and HAROLD D. LASSWELL and DANIEL LERNER (eds.), *The Policy Sciences: Recent Developments in Scope and Method* (1951), an effort to bring scientific method to the study of choices in public policy. Support for the establishment of a value-free non-normative science of politics was also provided by other writers such as STUART A. RICE, *Quantitative Methods in Politics* (1928, reprinted 1969), which may be considered the first general work on the application of statistical methods to the study of politics; GEORGE E.G. CATLIN, *The Science and Method of Politics* (1927), who thought it possible to state laws of politics; and WILLIAM BENNETT MUNRO, *Invisible Government* (1928), who also thought it possible to state laws of politics and urged disassociation from the sociologists and imitation of the natural scientists. A useful summary survey of political science around the world after the end of World War II is *Contemporary Political Science*, published

in 1950 by the UNITED NATIONS EDUCATIONAL, SCIENTIFIC, AND CULTURAL ORGANIZATION. The revival of an interest in Bentley that occurred in the 1930s and 1950s is exemplified by the restatement of his views in DAVID B. TRUMAN, *The Governmental Process: Political Interests and Public Opinion* (1951); EARL LATHAM, *The Group Basis of Politics: A Study in Basing-Point Legislation* (1952); and JACK W. PELTASON, *Federal Courts in the Political Process* (1955). A community study using reputational methods that seemed to validate the existence of local elites was FLOYD HUNTER, *Community Power Structure: A Study of Decision Makers* (1953). Community studies using decisional techniques that seemed to validate the existence of pluralized powers in local decision-making were ROBERT A. DAHL, *Who Governs? Democracy and Power in an American City* (1961); and WALLACE S. SAYRE, and HERBERT KAUFMAN, *Governing New York City: Politics in the Metropolis* (1960). In systems theory and its applications, besides the works of Easton and Deutsch, there is GABRIEL A. ALMOND and JAMES S. COLEMAN (eds.), *The Politics of the Developing Areas* (1960), in which the research technique is structural-functional analysis. The study of voting behaviour was advanced by the refinement of survey techniques in PAUL F. LAZARSFELD, BERNARD BERELSON, and HAZEL GAUDET, *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign* (1944); and by ANGUS CAMPBELL, GERALD GURIN, and WARREN E. MILLER, *The Voter Decides* (1954). An exemplary work on the operation of political parties and factionalism in V.O. KEY and HEARD ALEXANDER, *Southern Politics in State and Nation* (1949). Trenchant criticisms of behavioralism are those of BERNARD CRICK, *The American Science of Politics* (1959); and HERBERT J. STORING (ed.), *Essays on the Scientific Study of Politics* (1962).

(Ea.L.)

Political Systems

The term political system may be used either narrowly or broadly. Narrowly defined, it is the set of formal legal institutions that comprise a "government" or a "state." This is the definition adopted by many studies of the legal or constitutional arrangements of advanced political orders. More broadly defined, it comprehends actual as well as prescribed forms of political behaviour, not only the legal organization of the state but also the realities of political life. Still more broadly defined, the political system is seen as a set of "processes of interaction" or as a subsystem of the social system interacting with other non-political subsystems, such as the economic system. This usage guides attention to the importance of informal sociopolitical processes and emphasizes the study of political development.

None of these approaches may be neglected. Traditional legal or constitutional analysis, using the first definition, has produced a huge body of literature on governmental structures, many of the specialized terms that are a part of the traditional vocabulary of political science, and several classifying schemes that are still instructive. Similarly, empirical analysis of political processes and the effort to identify the underlying realities of governmental forms have yielded a rich store of data and an important body of comparative theory. The third definition has inspired much recent scholarly work, particularly on non-Western states, that employs new kinds of data, new terms, and some new concepts and categories of analysis. The discussion that follows will draw upon all three approaches to the study of political systems.

The article is divided into the following sections:

I. Forms of government
  Supranational political systems
  National political systems
  Subnational political systems
  Types and models
  Development and change
II. The structure of government
  Contemporary forms of government
  The levels of government
  The divisions of government
III. The functions of government
  The indispensable functions of government
  The supervisory functions of government
  The regulatory functions of government
  The enterprising functions of government
  Public administration
IV. Development and change in political systems
  Stability and instability
  The problem of political change
V. The study of political systems
  Contemporary political science
  Future approaches to the study of political systems

I. Forms of government

The most important type of political system in the modern world is the nation-state. The world today is divided territorially into more than 100 states, in each of which a national government claims to exercise sovereignty—or the power of final authority—and seeks to compel obedience to its will by its citizens. This fact of the world's political organization suggests the distinction employed in the following section among supranational, national, and subnational political systems.

SUPRANATIONAL POLITICAL SYSTEMS

A principal result of the division of the world into a number of separate national entities, or states, that have contact with one another, share goals or needs, and face common threats is the formation of supranational relationships. In some cases, as in many alliances, these relationships are short-lived and fail to result in significant institutional development. In other cases, they lead to interstate organizations and supranational systems. The discussion below examines several types of supranational political systems, together with historical and contemporary examples of each.

Empires. All empires, since they are composed of peoples of different cultures and ethnic backgrounds, are ultimately held together by coercion and the threat of forcible reconquest. Imposing their rule on diverse political structures, they are characterized by the centralization of power and the absence of effective representation of their component parts. Although force is thus the primary instrument of imperial rule, it is also true that history records many cases of multi-ethnic empires that were governed peaceably for considerable periods and were often quite successful in maintaining order within their boundaries. The history of the ancient world is the history of great empires—Egypt, China, Persia, and imperial Rome—whose autocratic regimes provided relatively stable government for many subject peoples in immense territories over many centuries. Based on military force and religious belief, the ancient despotisms were legitimized also by their achievements in building great bureaucratic and legal structures, in developing vast irrigation and road systems, and in providing the conditions for the support of high civilizations. Enhancing and transcending all other political structures in their sphere, they could claim to function as effective schemes of universal order.

Ancient and modern empires

In contrast to the empires of the ancient world, the colonial empires of recent times fell far short of universal status. In part, these modern European empires were made up of "colonies" in the original Greek sense; peopled by immigrants from the mother country, the colonies usually established political structures similar to those of the metropolitan centre and were often able to exercise a substantial measure of self-government. In part, also, the European empires were composed of territories inhabited by native populations and administered by imperial bureaucracies. The government of these territories was generally more coercive than in the European colonies and more concerned with protection and supervision of the commercial, industrial, and other exploitative interests of the imperial power. The disintegration of these empires occurred with astonishing speed. The two world wars of the 20th century sapped the power of the metropolitan centres, while their own doctrines of democracy, equality, and self-determination undermined the principle of imperial rule. Powers such as Britain and France found it increasingly difficult to resist claims to independence couched in terms of the representative concepts on which their home governments were based, and they lacked the military and economic strength to continue their rule over restive native populations. In the two decades after 1945 nearly all the major colonial territo-

ries won their independence; the great colonial empires that had once ruled more than half the world were finally dismembered.

Leagues. One of the commonest forms of supranational organization in history is that of leagues, generally composed of states seeking to resist some common military or economic threat by combining their forces. This was the case with the early city leagues, such as the Achaean and Aetolian leagues in ancient Greece and the Hanseatic and the Swabian leagues in Europe; and to a great extent it was the case with the League of Nations. Other common features of leagues include the existence of some form of charter or agreement among the member states, an assembly of representatives of the constituent members, an executive organ for the implementation of the decisions of the assembly of representatives, and an arbitral or judicial body for adjudicating disputes.

The League of Nations

The League of Nations was one of the great experiments in supranational organization of the 20th century and the predecessor in several important respects of the United Nations. The Covenant of the League was drafted by a special commission of the Peace Conference after World War I, with Pres. Woodrow Wilson of the United States as its leading advocate, and approved by a plenary conference of the victorious powers in 1919. The initial membership of the League consisted of 20 states. The United States failed to take membership in the League, but by 1928 the organization had a total membership of 54. The machinery of the League consisted of an Assembly of all the member nations, acting through agents of their governments; a council on which the great powers were permanently represented and to which the other member powers were elected by the Assembly for three-year terms; a Secretariat to administer the internal affairs of the League; and a number of specialized agencies, such as the International Labour Organisation, that were responsible for implementing various economic and humanitarian programs on an international basis. The Covenant required that international disputes be submitted to peaceful settlement with a provision for adjudication or arbitration by the Permanent Court of International Justice or for intervention by the Council of the League. The Covenant also provided for the use of financial and economic penalties, such as embargoes, to enforce the decisions of the League and for joint military action against convicted aggressors. In practice, however, the League failed its most important tests and was unable to master the crises that led to World War II and its own collapse.

Confederations. Confederations are voluntary associations of independent states that, to secure some common purpose, agree to certain limitations on their freedom of action and establish some joint machinery of consultation or deliberation. The limitations on the freedom of action of the member states may be as trivial as an acknowledgement of their duty to consult with each other before taking some independent action or as significant as the obligation to be bound by majority decisions of the member states. Confederations usually fail to provide for an effective executive authority and lack viable central governments; their member states typically retain their separate military establishments and separate diplomatic representation; and members are generally accorded equal status with an acknowledged right of secession from the confederation. Historically, confederations have often proved to be a first or second step toward the establishment of a national state, usually as a federal union. Thus, the federal union of modern Switzerland was preceded by a confederation of the Swiss cantons; West Germany's modern federal arrangements may be traced to the German Confederation of the 19th century (the Deutsche Bund); and the federal constitution of the United States is the successor to the government of the Articles of Confederation. In some other cases confederations have replaced more centralized arrangements, as, for example, when empires disintegrate and are replaced by voluntary associations of their former colonies. The British Commonwealth, or Commonwealth of Nations, and the French Community are cases of this type.

The Articles of Confederation

An example of confederal arrangements that gave birth to a federal union is the Articles of Confederation (1781–89) that preceded the Constitution of the United States. The Articles established a Congress of the confederation as a unicameral assembly of ambassadors from the 13 states, each possessing a single vote. The Congress was authorized to appoint an executive committee of states

> to execute, in the recess of Congress, such of the powers of Congress as the United States, in Congress assembled, by the consent of nine States, shall from time to time think expedient to vest them with;

in turn, the committee of states could appoint a presiding officer or president for a term of one year. The Congress could also appoint such other committees and "civil officers as may be necessary for managing the general affairs of the United States," and was given the authority to serve as "the last resort or appeal in all disputes and differences, now subsisting or that hereafter may arise between two or more states." Although the Congress was given authority in important areas such as the regulation of foreign affairs, the establishment of coinage and weights and measures, the appointment of officers in the confederation's land and naval forces, and the issuance of bills of credit, all its powers were in fact dependent for their enforcement upon the states. The Congress lacked both an independent source of revenue and the executive machinery to enforce its will directly upon individuals. As the language of the Articles summarized the situation,

> each State retains its sovereignty, freedom and independence, and every power, jurisdiction and right, which is not by this Confederation expressly delegated to the United States in Congress assembled.

The Commonwealth of Nations

The Commonwealth of Nations is an example of a confederation born as the result of the decentralization and eventual disintegration of an empire. A number of former British colonies and dominions, together with the United Kingdom, composed the commonwealth in 1972: Australia, Barbados, Botswana, Canada, Cyprus, Fiji, Gambia, Ghana, Guyana, India, Jamaica, Kenya, Lesotho, Malawi, Malaysia, Malta, Mauritius, Nauru, New Zealand, Nigeria, Sierra Leone, Singapore, Sri Lanka (Ceylon), Swaziland, Tanzania, Tonga, Trinidad and Tobago, Uganda, Western Samoa, and Zambia. Ireland and South Africa had seceded from the commonwealth, and Rhodesia's status was unclear after a lengthy dispute with Britain and the other member nations. All Commonwealth states recognized the Queen as the head of the commonwealth. Barbados, Canada, Jamaica, Malta, and Trinidad and Tobago were monarchies in which the Queen was represented by a governor general appointed on the recommendation of the government of the country concerned and acting in accord with its own constitutional practice; Botswana, Cyprus, Ghana, India, Kenya, Nigeria, Sri Lanka, Tanzania, and Zambia were republics; Malaysia and Lesotho had their own sovereigns; and Uganda had a prime minister who was head of state. Commonwealth governments were represented in the capitals of other commonwealth countries by high commissioners equal in status to ambassadors. The nations of the commonwealth rarely acted in concert on the international scene, and, despite fairly regular meetings of the commonwealth prime ministers, there were at times severe strains in the relations among several of the member states. The fairly general use of the English language and of English common law, together with some common symbols and remaining cultural affinities, appeared to be the major ties binding together this loose assocation of former British imperial possessions.

Federations. The term federation is used to refer to groupings of states, often on a regional basis, that establish central executive machinery to implement policies or to supervise joint activities. In some cases such groupings are motivated primarily by political or economic concerns; in others, military objectives are paramount. Examples of the former include the association of Belgium, France, The Netherlands, Italy, Luxembourg, and West Germany in what is usually referred to as the Western European Community but is actually a combination of three main structures—the European Coal and Steel

The Western European Community

Community, established in 1952; the European Economic Community (Common Market), established in 1958; and the European Atomic Energy Community (Euratom), established in 1958. The Community quickly developed executive machinery exercising significant regulatory and directive authority over the governments and private business firms of the member countries. Although each of the member governments retains a substantial measure of sovereignty, and a systematic effort by one or more of the governments to resist the authority of the Community's agencies could endanger the whole fabric of cooperative effort, the Community has developed significant supranational features. These include the staffing of executive organs with persons other than governmental representatives, the making of binding decisions on important matters by majority vote, and the capability of the Community agencies to deal directly and authoritatively with individuals and companies within the member states. For example, the high authority of the Coal and Steel Community acts by majority vote of its members, without instruction from any of the governments, to "assure the achievement of the purposes stated in the Treaty"; and in pursuing this function it involves itself deeply in the economies of each of the member nations.

The North Atlantic Treaty Organization (NATO), established in April 1949, is an example of a modern military alliance endowed with complex and permanent executive machinery, employing multilateral procedures, and involving the continuous elaboration of plans for the conduct of joint military action by its member states (Belgium, Canada, Denmark, Greece, Iceland, Italy, Luxembourg, The Netherlands, Norway, Portugal, Turkey, the United Kingdom, the United States, and West Germany; France is associated with the NATO nations for some purposes). As stated in its treaty, the purpose of NATO is to maintain the security of the North Atlantic area by exercise of the right of collective security recognized in the Charter of the United Nations. An impressive array of institutional mechanisms was established, including a secretary general and a permanent staff, a council, a network of committees, a military command structure, various study groups, and liaison staffs; and an ongoing system of collaboration in planning and joint military exercises was brought into being. With the continued development of its organization, NATO gradually added a number of economic and cultural activities to its functions until it came to possess several of the features of multipurpose supranational organization.

The present world association

The United Nations organization. A supranational political system that does not fit precisely any of the conventional classifications of such systems is the United Nations, a voluntary association of most of the world's nation-states. Its membership had grown from an original 51 states to a total of 130 in 1971, when the government of the People's Republic of China was admitted in place of the government of the Republic of China (Taiwan). Founded at a conference in San Francisco in 1945 attended by representatives of all the nations that had declared war on Germany or Japan, the purposes of the organization are declared in its Charter to be the maintenance of international peace and security, the development of friendly relations among states, and international cooperation in solving the political, economic, social, cultural, and humanitarian problems of the world. Its organizational structure consists of a Security Council of five permanent members (China, France, the Soviet Union, the United Kingdom, and the United States) and ten nonpermanent members elected for two-year terms, a General Assembly, a Secretary General and a Secretariat, an Economic and Social Council, a Trusteeship Council, and the International Court of Justice. Attached to the United Nations are a number of specialized agencies, including the Food and Agriculture Organization, the International Atomic Energy Agency, the International Civil Aviation Organization, the International Labour Organisation, the International Monetary Fund, the International Telecommunications Union, the Universal Postal Union, the United Nations Educational, Scientific and Cultural Organization, the World Health Organization, and the International Bank for Reconstruction and Development (World Bank).

Aside from the rather generally stated and decidedly elusive aims of the Charter, the member states of the United Nations cannot be said to have any common goal; they lack even the unity that a common external threat to security brings to many supranational organizations. There is also difficulty in reaching and implementing decisions. Two different formulas are employed for voting in the two principal organs, the General Assembly and the Security Council. In the General Assembly a two-thirds majority decides on important matters; but, since the Assembly's decisions are not binding and are merely recommendations, this qualified majority principle must be viewed as of little significance. Although, on the other hand, the decisions of the Security Council may be binding, a unanimous vote of all five of the permanent members joined by the votes of at least four of the nonpermanent members is required; whenever important questions of peace and security are at stake, it is rarely possible to achieve agreement among the five great powers of the council. These difficulties, which might be fatal to the survival of many supranational organizations, are not in fact totally debilitating. The United Nations continues to serve as a very important forum for international debate and negotiation, and its specialized agencies play an important role in what is sometimes referred to as "the functional approach to peace."

NATIONAL POLITICAL SYSTEMS

The nation-state defined

The term nation-state is used so commonly but defined so variously that it will be well to indicate its usage in this article with some precision and to give historical and contemporary examples of nation-states. To begin with, there is no single basis upon which such systems are established. Many states were formed at a point in time when a people sharing a common history, culture, and language discovered a sense of identity. This was true in the cases of England and France, which were the first nation-states to emerge in the modern period, and of Italy and Germany, which were established as nation-states in the 19th century. In contrast, however, other states, such as India, the Soviet Union, and Switzerland, have come into existence without a common basis in race, culture, or language. It must also be emphasized that contemporary nation-states are creations of different historical periods and of very varied circumstances. Before the close of the 19th century the effective mobilization of governmental powers on a national basis had occurred only in Europe, the United States, and Japan. It was not until the 20th century and the collapse of the Ottoman, Habsburg, French, and British empires that the world could be fully organized on a national basis. In 1920 the League of Nations recognized seven nation-states as "Great Powers" (Britain, France, the United States, Germany, Italy, Japan, and Russia) and eventually admitted more than 40 other states to membership; in 1946 the United Nations admitted 51 states to membership, but the number had grown to 130 by 1971. The nation-states in today's world include the three "superpowers," the United States, the Soviet Union, and China; the established powers of second rank, Britain, France, Japan, and West Germany; emerging powers such as India and Indochina; and a host of old and new lesser states, such as Denmark, Gambia, Switzerland, Cyprus, Belgium, Malaysia, Norway, Chile, and others.

The characteristics that qualify these variously composed and historically differing entities as nation-states and distinguish them from other forms of social and political organization amount in sum to the independent power to compel obedience from the populations within their territories. The state is, in other words, a territorial association that may range in size from the Soviet Union to Singapore, in population from China to Luxembourg, and that claims supremacy over all other associations within its boundaries. As an association, the state is peculiar in several respects: membership is compulsory for its citizens; it claims a monopoly of the use of armed force within its borders; and its officers, who are the govern-

ment of the state, claim the right to act in the name of the land and its people.

A definition of the state in terms only of its powers over its members is not wholly satisfactory, however. Although all states make a claim to supremacy within their boundaries, they differ widely in their ability to make good their claims. States are, in fact, often challenged by competing associations within their boundaries; their supremacy is often more formal than real; and they are sometimes unable to maintain their existence. Moreover, a definition in terms of power alone ignores the fact that there are great differences among states in the structures they employ for the exercise of power, in the ways they use power, and in the ends to which they turn their power. Some of these differences are explored in the discussion that follows of two general categories of nation-states: the unitary state and the federal state. Partly from administrative necessity and partly because of the pressures of territorial interests, nearly all modern states provide for some distribution of governmental authority on a territorial basis. Systems in which power is delegated from the central government to subnational units and in which the grant of power may be rescinded at the will of the central government are termed unitary systems. Systems in which a balance is established between two autonomous sets of governments, one national and the other provincial, are termed federal. In federal systems, the provincial units are usually empowered to grant and take away the authority of their own subunits in the same manner as national governments in unitary systems. Thus, although the United States is federally organized at the national level, each of the 50 states is in a unitary relationship to the cities and local governments within its own territory.

Unitary nation-states. A majority of all the world's nation-states are unitary systems, including Belgium, Bulgaria, France, Great Britain, The Netherlands, Japan, Poland, Romania, the Scandinavian countries, Spain, and many of the Latin American and African countries. There are great differences among these unitary states, however, in the institutions and procedures through which their central governments interact with their territorial subunits.

Decentralization of power in unitary states

In one type of unitary system, decentralization of power among subnational governments goes so far that in practice, although not in constitutional principle, they resemble federal arrangements. In Britain, for example, there are important elements of regional autonomy in the relationship between Northern Ireland, Wales, and Scotland and the national government in London; and the complex system of elected local governments, although in constitutional theory subject to abrogation by Parliament, is in practice a fixed and fairly formidable part of the apparatus of British government. In other unitary systems of this type, decentralization on a territorial basis is actually provided for constitutionally, and the powers of locally elected officials are prescribed in detail. Thus, the Japanese constitution specifies certain autonomous functions to be performed by local administrative authorities.

A second type of unitary system makes substantially less provision for territorial decentralization of authority and employs rather strict procedures for the central supervision of locally elected governments. The classic example of this type is France. The French administrative system is built around 95 *départements*, each headed by a *préfet*, and 322 subdivisions of the *départements*, termed *arrondissements*, each headed by a *sous-préfet*. The *préfets* and *sous-préfets* are appointed by the government in Paris to serve as agents of the central government and also as the executives of the divisional governments or councils, which are composed of elected officials. The system thus combines central supervision of local affairs through appointed officers with territorial representation through locally elected governments.

A third type of unitary system provides for only token decentralization. In such cases, the officials responsible for managing the affairs of the territorial subdivisions are appointees of the central government, and the role of locally elected officers is either minimal or nonexistent. Examples of this kind of arrangement include Germany under Adolf Hitler and also several Communist countries. The Third Reich was divided into 42 *Gaue*, each headed by a gauleiter chosen for his personal loyalty to Hitler; in eastern Europe, the people's councils or people's committees are named by the centrally organized Communist parties; their appointment is confirmed by elections in which one slate of candidates is presented.

Federal systems. In federal systems, political authority is divided between two autonomous sets of governments, one national and the other subnational, both of which operate directly upon the people. Usually a constitutional division of power is established between the national government, which exercises authority over the whole national territory, and provincial governments that exercise independent authority within their own territories. Of the eight largest countries in the world, seven—Argentina, Australia, Brazil, Canada, India, the Soviet Union, and the United States—are organized on a federal basis. Federal states also include Austria, Cameroon, Czechoslovakia, Libya, Malaysia, Mexico, Nigeria, Switzerland, Tanzania, Venezuela, West Germany, and Yugoslavia.

The governmental structures and political processes found in these federal systems show great variety. One may distinguish, first, a number of systems in which federal arrangements reflect rather clear-cut cultural divisions. A classic case of this type is Switzerland, where the people speak four different languages—German, French, Italian, and Romansh—and the federal system unites 22 historically and culturally different entities, known as cantons. The Swiss constitution of 1848, as modified in 1874, converted a confederation originally formed in the 13th century by the three forest cantons of Uri, Schwyz, and Unterwalden into the modern federal state. The principal agencies of federal government are a bicameral legislature, composed of a National Council representing the people directly and a Council of States representing the constituent members as entities; an executive branch (Bundesrat) elected by both houses of the legislature in joint session; and a supreme court that renders decisions on matters affecting cantonal and federal relations. The Soviet Union's federal arrangements, although of a markedly different kind, also reflect underlying cultural and linguistic differences among the 15 union republics. The member republics of the Union of Soviet Socialist Republics are constitutionally endowed with territorial autonomy, a "right to secession," and equal representation in the Supreme Soviet. Ethnic subdivisions within the 15 union republics are also granted territorial autonomy and direct representation in the upper chamber of the Supreme Soviet. These arrangements, however, are rather more important in Soviet constitutional theory than in practice, since, in addition to the virtual monopoly of military, economic, and social controls vested in the central government, more than half the population and territory of the Soviet Union lies within one member republic, the Russian Socialist Federated Soviet Republic.

The Swiss and Soviet systems

In other systems, federal arrangements are found in conjunction with a large measure of cultural homogeneity. The Constitution of the United States delegates certain activities that concern the whole people, such as the conduct of foreign relations and war and the regulation of interstate commerce and foreign trade to the federal government; certain other functions are shared between the federal government and the states; and the remainder are reserved to the states. Although these arrangements require two separate bodies of political officers, two judicial systems, and two systems of taxation, they also allow extensive interaction between the federal government and the states. Thus, the election of Congress and the President, the process of amending the Constitution, the levying of taxes, and innumerable other functions necessitate cooperation between the two levels of government and bring them into a tightly interlocking relationship.

SUBNATIONAL POLITICAL SYSTEMS

Although national government is the dominant form of political organization in the modern world, an extraordi-

Basic political communities

nary range of political forms exists below the national level—tribal communities, the intimate political associations of villages and towns, the governments of regions and provinces, the complex array of urban and suburban governments, and the great political and administrative systems of the cities and the metropolises. These subnational entities are, in a sense, the basic political communities—the foundation on which all national political systems are built.

Tribal communities. The typical organization of mankind in its early history was the tribe. Today, in many parts of the world, the tribal community is still a major form of human political organization. Even in advanced systems, traces can still be found of its influence. Some of the *Länder* of modern Germany, such as Bavaria, Saxony, or Westphalia, have maintained their identity since the days of the Germanic tribal settlements. In England, too, many county boundaries can be explained only by reference to the territorial divisions in the period after the end of the Roman occupation.

In many of the new states of Africa, the tribe is still the effective community and a principal vehicle of political consciousness. Most of these states are the successors to the administrative units established by colonial regimes and owe their present boundaries to the often arbitrary decisions of imperial bureaucracies or to the territorial accommodations of rival colonial powers. The result was often the splintering of the tribal communities or their aggregation in largely artificial entities. That the force of tribalism is far from exhausted is suggested by the experience of Nigeria and other new states where a resurgence of tribal rivalries has presented great obstacles to the process of nation building.

Importance of kinship in tribal societies

In its primary sense, the tribe is a community organized in terms of kinship, and its subdivisions are the intimate kindred groupings of moieties, gentes, and totem groups. Its territorial basis is rarely defined with any precision, and its institutions are typically the undifferentiated and intermittent structures of an omnifunctional social system. The leadership of the tribe is provided by the group of adult males, the lineage elders acting as tribal chiefs, the village headmen, or the shamans, or tribal magicians. These groups and individuals are the guardians of the tribal customs and of an oral tradition of law. Law is thus not made but rather invoked; its repository is the collective memory of the tribal council or chief men. This kind of customary law, sanctioned and hallowed by religious belief, nevertheless changes and develops, for each time it is declared something may be added or omitted to meet the needs of the occasion.

Rural communities. The village has traditionally been contrasted with the city: the village is the home of rural occupations and tied to the cycles of agricultural life, while the inhabitants of the city practice many trades, and its economy is founded on commerce and industry; the village is an intimate association of families, while the city is the locus of a mass population; the culture of the village is simple and traditional, while the city is the centre of the arts and sciences and of a complex cultural development. The village and the city offer even sharper contrasts as political communities. Historically, the village has been ruled by the primitive democracy of face-to-face discussion in the village council or by a headman whose decisions are supported by village elders or by other cooperative modes of government; urban government has never been such a simple matter, and monarchical, tyrannical, aristocratic, and oligarchic forms of rule have all flourished in the city. In the village, the boundaries among political, economic, religious, and other forms of action have not been as clearly drawn as in cities.

The origins and development of the apparatus of government can be seen most clearly in the simple political society of the rural community. The transformation of kin-bound societies with their informal, folk-sustained systems of sociopolitical organization into differentiated, hierarchical societies with complex political structures began with the enlargement of the rural community —an increase in its population, the diversification of its economy, or its interaction with other communities. The rudimentary organs of communal government were then elaborated, the communal functions received more specialized direction, and leadership roles were institutionalized. This was sometimes a process that led by gradual stages to the growth of cities. Elsewhere, however, as in the case of ancient Attica, the city was established as the result of a process of *synoikismos,* or the uniting of a number of tribal or village communities. This was undoubtedly the origin of Athens, and, according to its legendary history, Rome also was established as a result of the forcible unification of the tribes that dwelt on the hills surrounding the Palatine Hill.

Even in the nation-states of today's world, the contrasts between the village or the town and the city as centres of human activity are readily apparent. In the country, life is more intimate, the human contacts more informal, the structure of society more stable. In the city, the individual becomes anonymous, the contacts between people are mainly formal, and the standing of the individual or the family in society is subject to rapid change. In many contemporary systems, however, the differences in the forms of government of rural and urban communities appear to be growing less pronounced. In the United States, for example, rural institutions have been seriously weakened by the movement of large numbers of people to the city. The township meeting of New England and other forms of direct citizen participation in the affairs of the community have declined in importance and have often been displaced by more formal structures and the growth of local governmental bureaucracies.

Cities. Cities first emerged as complex forms of social and political organization in the valleys of the Euphrates and the Tigris, the Nile, the Yellow River, and the Yangtze. These early cities broke dramatically with the patterns of primitive life and the rural societies from which they sprang. Kinship as the basis of society was replaced with status determined by class and occupation; the primitive magical leaders of the tribe were displaced by temple priesthoods presiding over highly developed religious institutions and functioning as important agencies of social control; earlier systems of rule by the tribal chieftains and the simple forms of communal leadership gave way to kingships endowed with magical powers and important religious functions; and specialized functionaries in the royal courts became responsible for supervising new kinds of governmental activity. Many other developments contributed to the growing centralization of power in these city civilizations. Barter was replaced by more effective systems of exchange, and the wealth generated in commerce and the specialized city trades became both an object of taxation and an instrument of power. Class distinctions emerged as the result of a division of labour and advances in technical development. A military order and a professional soldiery were created and trained in new techniques of warfare, and a slave class provided the work force for large-scale projects of irrigation, fortification, and royal architecture. As these developments proceeded, the city was able to project its power even further into the surrounding countryside, to establish its rule over villages and other cities in its sphere, and finally to become the centre of such early empires as those of Sumeria, Egypt, China, Babylonia, Assyria, and Persia.

Organization of ancient cities

A very different form of city life emerged among the Greeks. The Greek polis also broke with the folkways of primitive society, but its political development was in striking contrast to the despotism of the Oriental city empires and their massive concentrations of power in the hands of king and priest. As the polis transcended its origins in village life, the powers of the tribal chief dwindled and passed into the hands of aristocratic families. The kingship of Homeric tradition vanished, the "kings" who remained became mere dignitaries in the religious and ceremonial life of the city, and new magistracies and other civic offices were founded. These offices became the focus of factional struggle among the aristocratic families and, later, with the weakening of aristocratic rule, the chief prizes in a contest of power between the nobility

The Greek polis

and the common citizens. Eventually, these developments issued in the characteristic form of Greek city government. A citizen body, always a much narrower group than the total population but often as numerous as the population of freeborn males, acquired power in the direction of the city government through the election of its officers and direct participation in the city councils. Although often interrupted by episodes of oligarchic or tyrannical rule and by periods of civic dissension and class rivalry, the main theme of governmental development in the Greek city was the elaboration of structures that permitted the control of political affairs by its citizens.

The medieval city

Autonomous cities also sprang up in Europe in the later Middle Ages. Medieval city life, although it differed from that of the polis and was coloured by the forms of feudal society, also emphasized the principle of cooperative association. Indeed, for the first time in the history of city civilization, the majority of the inhabitants of the city were free. The development of trades, the growth of commerce, and the mobilization of wealth emancipated the city from its feudal environment, and the merchant and craft guilds became the matrices of a new kind of city democracy. In time, the guilds were transformed into closed corporations and became a basis for oligarchic control; and the city's independence was threatened by the rise of the new nation-states. Tempting targets for the ambition of kings, Venice, Genoa, Florence, Milan, Cologne, Amsterdam, Hamburg, and other free cities of Europe eventually succumbed to monarchical control. Theirs was an important legacy, however, for the political order of the medieval city was a powerful influence in the development of the constitutional structures of the modern democratic state.

Modern cities

Although cities are no longer independent, the almost universal increase in urban population has made them more important than ever before as centres of human activity. The political organization of modern cities differs from country to country. Even within the same nation-state, there are often important contrasts in the structures of city government. In the United States, for example, three principal types of city government are usually distinguished: the council–manager form, the mayor–council form, and the commission form.

More than half of all American cities with populations of between 25,000 and 250,000 operate under council–manager governments. In council–manager systems the council is generally small, elected at large on a nonpartisan ballot for overlapping four-year terms; no other offices are directly elected, and the mayor, who presides at council meetings and performs mainly ceremonial functions, is chosen by the council from among its members. The manager, a professional city administrator, is selected by the council, serves at the council's pleasure, and is responsible for supervising the city departments and municipal programs, preparing the budget, and controlling expenditures.

Mayor–council governments are found in two basic forms, the "weak" mayor and the "strong" mayor. The former was typical of the 19th-century municipal organization and is now mainly confined to cities of less than 25,000 population; the latter is a common arrangement in cities with populations of more than 5,000,000. In weak-mayor–council governments, a number of officials, elected or appointed for lengthy terms, wield important administrative powers; the council, typically elected by divisions of the city called wards, is responsible for the direction of the major city agencies; and the mayor's powers of appointment and removal and his control over the city budget are severely limited. In many cases, strong-mayor–council governments evolved from weak-mayor–council systems as an independently elected mayor won the power of veto over council ordinances, strengthened his control over appointment and removal, and established himself as the city's chief budgetary officer; at the same time, also, the elective administrative officers and the semi-autonomous appointive boards and commissions were often eliminated and the number of councilmen reduced.

The commission plan, which is now found in fewer than 200 cities of more than 10,000 population, concentrates legislative and executive powers in the hands of a small group of commissioners. The commissioners serve individually as the heads of administrative departments and choose one of their number to act as a ceremonial mayor without executive authority.

The variety in the governmental structures of American cities is paralleled in many other countries, for everywhere in the modern world the government of the city continues to challenge man's political invention. Although no longer sovereign, cities are the centres of modern civilization and, both in terms of the services demanded of them and the range and importance of the functions they exercise, the most important of contemporary subnational political systems. Moreover, it is in the cities that most of the problems of modern industrial society seem to have their focus. These problems are not only governmental but also technological, cultural, and economic. They are found in their most acute form in the great metropolitan centres and in that vast urban agglomeration known as the megalopolis. In political terms, the issue that is posed appears to be whether these huge centres of population can continue as effective communities with democratically manageable governments.

Regions. In many contemporary national political systems the forces of history and administrative necessity have joined to produce regional communities at an intermediate level between the local and the national community. In some cases—the Swiss canton, the English county, the German *Land*, and the American state—these regional communities possess their own political institutions and exercise governmental functions. In other cases, however, the territorial community is a product of ethnic, cultural, linguistic, physiographic, or economic factors and maintains its identity without the support of political structures.

Regional communities

As subnational political systems, regional communities are sometimes based in tradition, even tracing their origin to a period prior to the founding of the nation; in other cases, they are modern administrative units created by national governments for their own purposes. Examples of both types may be found in the history of regionalism in France and its complex pattern of internal territorial divisions. Before the French Revolution, France was divided into ancient provinces—Burgundy, Gascony, Brittany, Normandy, Provence, Anjou, Poitou, and others. After the Revolution, in what seems to have been an effort to discourage regional patriotism and threats of separatism, the Napoleonic government superimposed a new regional structure of *départements* on the old provincial map. More than a century and a half later, in the era of rapid communications and national economic planning, the French national government announced a regrouping of the Napoleonic *départements* into much larger Gaullist *régions*. Recognizing, perhaps, the continuing strength of the provincial attachments of Gascon, Breton, Normand, and Provençal and the survival of old regional folk cultures with their distinctive patterns of speech, the new *régions* were given boundaries similar in many cases to the traditional provincial boundaries of pre-republican France.

The history of the French regional communities is not a special case, for political, administrative, economic, and technical forces have led many other national governments to replace traditional territorial divisions with new regional units. In England, for example, the traditional structure of county governments (52 in number) was replaced by a system of administrative counties (61 plus London); and, in 1969, a royal commission proposed a further reform that would abolish 39 counties. Attempts have also been made to use older regional communities as the infrastructure for new systems of regional government. Thus, the Italian constitution provides for semi-autonomous government in five special regions—Valle d'Aosta, Sardinia, Sicily, Trentino-Alto Adige, and Friuli-Venezia Giulia—which, in different ways, are historically distinct from the rest of Italy. In yet other cases the fear of competition from regional governments or of sep-

aratist movements has led national governments to make various efforts to resist the development of regional political structures. Again, Italy provides a convenient example, for its constitution requires the establishment of 14 other autonomous regions, but Italian governments refused to implement this provision of the 1947 constitution until recently. It should be noted that the Italian republic of 1870–1922 and its Fascist successor state also made similar efforts to combat regional political development, the former by the creation of a large number of administrative provinces and the latter by establishing *corporazione* to represent occupations regardless of geographic location.

In several modern states the growth of vast conurbations and the rise of the megalopolis have prompted the development of new kinds of regional governmental structures. The Port of London Authority, the Port of New York Authority, and the San Francisco Bay Area Transit Authority are examples of regional systems designed to serve the needs of urban communities that have outgrown the boundaries of existing city governments. New regional structures have also resulted from the increased responsibility of national governments for the administration of comprehensive social and economic programs. The Tennessee Valley Authority, for example, is both a national agency and a regional government whose decisions affect the lives of the inhabitants of all the states and cities in its sphere. Other examples of new regional administrative structures include the zonal councils established in India for social and economic planning purposes, the districts of the Interstate Commerce Commission in the United States, and the governmental and economic units established in Britain to deal with the problems of industrially depressed areas.

TYPES AND MODELS

The almost infinite range of political systems has been barely suggested in this brief review. Confronted by the vast array of political forms, political scientists have attempted to classify and categorize, to develop typologies and models, or in some other way to bring analytic order to the bewildering variety of data. Many different schemes have been developed. There is, for example, the classical distinction between governments in terms of the number of rulers—government by one man (monarchy or tyranny), government by the few (aristocracy or oligarchy), and government by the many (democracy). There are schemes classifying governments in terms of their key institutions (for example, parliamentarism, cabinet government, presidentialism). There are classifications that group systems according to basic principles of political authority or the forms of legitimacy (charismatic, traditional, rational–legal, and others). Other schemes distinguish between different kinds of economic organization in the system (the laissez-faire state, the corporate state, and Socialist and Communist forms of state economic organization) or between the rule of different economic classes (feudal, bourgeois, and capitalist). And there are modern efforts to compare the functions of political systems (capabilities, conversion functions, and system maintenance and adaptation functions) and to classify them in terms of structure, function, and political culture.

Approaches to the analysis of political systems

Although none is comprehensive, each of these principles of analysis has some validity, and the classifying schemes that are based on them, although in some cases no longer relevant to modern forms of political organization, have often been a major influence on the course of political development. The most influential of such classifying schemes is undoubtedly the attempt of Plato and Aristotle to define the basic forms of government in terms of the number of power holders and their use or abuse of power. Plato held that there was a natural succession of the forms of government: an aristocracy (the ideal form of government by the few) that abuses its power develops into a timocracy (in which the rule of the best men, who value wisdom as the highest political good, is succeeded by the rule of men who are primarily concerned with honour and martial virtue), which through greed develops into an oligarchy (the perverted form of government by the few), which in turn is succeeded by a democracy (rule by the many); through excess, the democracy becomes an anarchy (a lawless government), to which a tyrant is inevitably the successor. Abuse of power in the Platonic typology is defined by the rulers' neglect or rejection of the prevailing law or custom (*nomos*); the ideal forms are thus *nomos* observing (*ennomon*), and the perverted forms are *nomos* neglecting (*paranomon*). Although disputing the character of this implacable succession of the forms of government, Aristotle also based his classification on the number of rulers and distinguished between good and bad forms of government. In his typology it was the rulers' concern for the common good that distinguished the ideal from perverted forms of government. The ideal forms in the Aristotelian scheme are monarchy, aristocracy, and polity (a term conveying some of the meaning of the modern concept of "constitutional democracy"); when perverted by the selfish abuse of power, they are transformed respectively into tyranny, oligarchy, and ochlocracy (or the mob rule of lawless democracy). The concept of the polity, a "mixed" or blended constitutional order, fascinated political theorists for another millennium. To achieve its advantages, innumerable writers from Polybius to St. Thomas Aquinas experimented with the construction of models giving to each social class the control of appropriate institutions of government.

Another very influential classifying scheme was the distinction between monarchies and republics. In the writings of Machiavelli and others, the tripartism of classical typologies was replaced by the dichotomy of princely and republican rule. Sovereignty in the monarchy or the principality is in the hands of a single ruler; in republics, sovereignty is vested in a plurality or collectivity of power holders. Reducing aristocracy and democracy to the single category of republican rule, Machiavelli also laid the basis in his analysis of the exercise of princely power for a further distinction between despotic and nondespotic forms of government. In the work of Montesquieu, for example, despotism, or the lawless exercise of power by the single ruler, is contrasted with the constitutional forms of government of the monarchy and the republic. As a result of the decline of monarchies and the rise of new totalitarian states terming themselves republics, this traditional classification is now, of course, of little more than historical interest.

The usefulness of all the traditional classifications has been undermined by the momentous changes in the political organization of the modern world. Typologies based on the number of power holders or the formal structures of the state are rendered almost meaningless by the standardization of "democratic" forms, the deceptive similarities in the constitutional claims and governmental institutions of regimes that actually differ markedly in their political practices, and the rise of new political orders in the non-Western world. A number of modern writers have attempted to overcome this difficulty by constructing classifying schemes that give primary importance to social, cultural, economic, or psychological factors. The most influential of such schemes is the Marxist typology, which classifies types of rule on the basis of economic class divisions and defines the ruling class as that which controls the means of production in the state. A monistic typology that also emphasized the importance of a ruling class was developed by an Italian theorist of the early 20th century, Gaetano Mosca. In Mosca's writings all forms of government appear as mere facades for oligarchy or the rule of a political "elite" that centres power in its own hands. Another classification, which distinguishes between "legitimate" and "revolutionary" governments, was suggested by Mosca's contemporary Guglielmo Ferrero. Using a sociopsychological approach to the relations between rulers and ruled, Ferrero held that a legitimate government is one whose citizens voluntarily accept its rule and freely give it their loyalty; in revolutionary systems, the government fears the people and is feared by them. Legitimacy and leadership are also the basis of a typology developed by the German sociologist Max We-

Modern classifying systems

ber. In Weber's scheme there are three basic types of rule: charismatic, in which the authority or legitimacy of the ruler rests upon some genuine sense of calling and in which the followers submit because of their faith or conviction in the ruler's exemplary character; traditional, in which, as in hereditary monarchy, leadership authority is historically or traditionally accepted; and rational–legal, in which leadership authority is the outgrowth of a legal order that has been effectively rationalized and where there is a prevailing belief in the legality of normative rules or commands. The Weberian typology has been elaborated by a number of recent writers who have found it particularly useful for comparing and classifying the emergent political orders of the non-Western world.

A serviceable classification of political systems must penetrate beneath formal appearances to underlying realities; these realities, however, do not consist only of the facts of social and economic organization. Important differences often exist between political systems having very similar socioeconomic structures. That is why some recent sociological classifications and schemes of analysis fail as tools of political inquiry: they cannot effectively distinguish between certain societies whose political orders are full of contrasts. The political system itself must be the primary focus of inquiry and the phenomena of politics the principal facts of investigation. Such an approach may involve many different kinds of analysis, but it must begin with an examination of the ways in which power is acquired and transferred, exercised, and controlled. This is important for comparing advanced political orders and also for drawing important distinctions between regimes in the underdeveloped areas of the world.

The acquisition and transfer of power. A key problem of all political orders is that of succession. "The king is dead; long live the king" was the answer, not always uncontested, of European hereditary monarchy to the question of who should rule after the death of the king. A second, closely related problem is in what manner and by whom a present ruler may be replaced or deprived of power. To this second question hereditary monarchy gave no definite answer, although the concept of *diffidatio,* or the severance of the bond of allegiance between king and feudal lord, was invoked more than once in the medieval period. Political systems, even those of primitive tribal societies, have approached both problems in a variety of ways. Anthropological records show that tribal chiefs or kings were sometimes selected as a result of ritual tests or the display of magical signs and proofs of divine origin, usually as determined by the tribal elders or magical leaders; in other cases, a principle of heredity, often diluted by a choice among heirs in terms of physique or warrior ability, was applied; in still other cases, the chief was elected, often from among the adult males of a select group of families. Techniques for the removal of tribal rulers were equally varied. Sometimes the ruler would be killed after a specified period or when his magical powers weakened or when his physical prowess or health failed; in other cases the chief was exposed to periodic tests of his magical powers or required to accept challenges to combat from other qualified candidates for rule; and in some cases the elders could remove him from office.

Techniques for assuring the succession are also varied in the modern world. Succession procedures range from the complex hieratic process of identifying a reincarnated Dalai Lama, which was practiced until quite recently in Tibet, to the subtle, informal procedures by which parliamentary majorities choose a successor to the office of prime minister in Britain. In fact, however, the succession practices of modern political systems appear to be of four main types: (1) heredity, (2) constitutional prescription, (3) election, and (4) force.

Hereditary succession

Although dictators still occasionally seek to establish their sons as their heirs, they usually rely on force rather than the claims of heredity to achieve their object. Apart from a few states, mostly in the Arab world, where the dynastic ruler is the effective head of the government, the hereditary principle of succession is now almost exclusively confined to the constitutional monarchies of western Europe. There is some irony in the fact that the line of succession is more securely established in these monarchies now than at any point in their earlier history: intradynastic struggle, it appears, is much less likely when kingship is mainly ceremonial. Heredity may be reinforced or modified by constitutional prescription: this was the case, for example, of the famous Act of Settlement that secured the Hanoverian succession in Britain.

Succession by constitutional prescription. A leading example of succession by constitutional prescription is the United States. Article II, Section 1, of the Constitution of the U.S. provides:

> In case of the removal of the President from office, or of his death, resignation, or inability to discharge the powers and duties of the said office, the same shall devolve on the Vice President, and the Congress may by law provide for the case of removal, death, resignation, or inability, both of the President and Vice President, declaring what officer shall then act as President, and such officer shall act accordingly, until the disability be removed, or a President shall be elected.

A constitutional amendment was ratified in 1967 elaborating these procedures to include further arrangements for dealing with the problem of presidential disability. The original language of the Constitution has been the basis for the peaceful succession of Vice Presidents John Tyler, Millard Fillmore, Andrew Johnson, Chester Arthur, Theodore Roosevelt, Calvin Coolidge, Harry S. Truman, and Lyndon Johnson. Constitutionally prescribed arrangements for assuring the succession are not always so successful, and many states whose constitutions contain very similar provisions have experienced succession crises that were resolved only by violence.

Succession by election. Election is a principle of succession also frequently combined with force. In cases of closely contested elections or where there is doubt as to the validity or proper form of the election, the result is often a disputed succession. The Great Schism in the papacy in the 14th century and the disputed succession to the elective kingship of Hungary in the 16th century are instructive examples of the failure of elective systems to assure an orderly succession. Force is also the effective basis of succession in several contemporary states in which *pro forma* confirmation by electoral majorities is given to the ruler who seizes power.

The problem of succession imposes great strains on any political order: the continuity of rule is broken, established patterns of action are interrupted, and the future suddenly becomes uncertain. This political crisis tests the character of regimes in ways that are of some importance for comparative political analysis. A number of interesting comparisons may be drawn from the study of succession practices, but perhaps the most important is the distinction between those systems in which the problem is resolved primarily by force and those systems in which heredity, constitutional prescription, or election assure a peaceful and orderly succession.

The transfer of power

Political orders are subjected to another kind of strain when the rule of their present power holders is challenged and the question arises of depriving them of authority. This is the problem of the transfer of power: whether, in what way, and by whom a present ruler may be displaced. Like succession, it is a recurrent problem in all political systems, and, as in the case of succession practices, the ways in which political systems respond to the strains involved offer important clues to their character. It is, in a sense, the fundamental political crisis, for all systems are in some way shaken, often violently, sometimes to the point of destruction, by the struggles between established rulers and their rivals.

Revolutions, which are the result of the crisis in its most extreme form, involve the overthrow not merely of the government but of the political order itself. Typically, a revolution is preceded by a series of strains within the system: challenges to the authority of the government mount, and its legitimacy is increasingly questioned; the exercise of power becomes coercive, and the challenge to rule assumes ever more violent forms; eventually, the

struggle comes to a dramatic climax in the destruction of the old order. The coup d'etat is another form of violent response to the crisis of rule, but it is distinguished from the revolution in that it involves the overthrow only of the government: the political order is not immediately affected, for the coup is managed by an individual or group within the government or within the ruling class. In some cases, however, the coup d'etat is merely a preliminary stage to revolution. Sometimes this happens when the new ruler leads a governmentally imposed revolution: this was the role played by Napoleon I, Napoleon III, Mussolini, and Hitler. At other times, coups are actually prompted by fear of revolution but succeed only in further weakening the claims to legitimacy of the existing order: this has recently been the case in some countries in the non-Western world where conservative-led coups were quickly overthrown by revolutionary movements. In addition to revolutions and coups d'etat, the crisis of rule may prompt other forms of violent political reaction, including civil war and secession, resistance movements and rebellions, guerrilla warfare and terrorism, class warfare, and peasant revolts.

Causes of the overthrow of governments

The causes of internal conflict leading to the forcible overthrow of governments are extremely varied. They include tensions created by rapid social and economic development; the rise of new social classes and the refusal of established elites to share their power; problems of the distribution of wealth and the grievances of different economic groups and interests; the rise of corrosive social and political philosophies and the estrangement of intellectuals; conflict of opinions over the ends of government; factional struggles among power holders or within the ruling class; the rise of a charismatic leader; oppressive rule that alienates powerful groups; weak rule that tolerates antigovernmental or revolutionary movements; and many different combinations of these and other social, economic, and political factors. All political systems experience some of these conditions with some frequency. Yet there are a number of modern states that have avoided internal wars and the forcible overthrow of their governments for considerable periods.

It appears that rulers in the contemporary world are generally safe from violent challenges if they possess an effective monopoly of military, economic, and political power, linked with certain important social controls; or, alternatively, if they are obliged to exercise limited powers for specified periods and are required to yield office to rivals who meet certain qualifications. The first is the definition of a modern totalitarian regime, fully and efficiently organized; the second describes the governments of several contemporary constitutional democracies. In the first case, the government secures itself by force combined with social and psychological means of preventing the formation of opposition. In the second case, alternatives to internal war are provided by the opportunities for oppositions to influence the exercise of power and ultimately to replace the government. The great achievement of constitutional democracy has been to give reasonable security to governments from forcible overthrow by compelling them to accept limitations on their power, by requiring them to forgo the use of force against rivals who agree to accept the same limitations, and by establishing well-known legal procedures through which these rivals may themselves constitute the government.

The exercise and control of power. The foregoing discussion has suggested a distinction among political systems in terms of the role played by force in the acquisition and transfer of power. The role of force is vital, also, in distinguishing among political systems in terms of the exercise and control of power. Here the contrast is essentially between "autocratic" and "nonautocratic" governments, for totalitarianism is only a recent, although very successful, species of autocracy, to which constitutionalism is the principal contemporary antithesis. Autocracy is characterized by the concentration of power in a single centre, be it an individual dictator or a group of power holders such as a committee or a party leadership. This centre relies on force to suppress opposition and to limit social developments that might eventuate in opposition. The power of the centre is not subject to effective controls or limited by genuine sanctions: it is absolute power. In contrast, nonautocratic government is characterized by the existence of several centres, each of which shares in the exercise of power. Nonautocratic rule allows the development of social forces that generate a variety of interests and opinions. It also subjects the power holders to reciprocal controls and to effective sanctions of law.

Autocracy

In appearance, autocracy may sometimes be difficult to distinguish from nonautocratic rule. Often, autocracies attempt to borrow legitimacy by adopting the language of the constitutions of nonautocratic regimes or by establishing similar institutions. It is a common practice, for example, in many modern totalitarian states to establish institutions—parliaments or assemblies, elections and parties, courts and legal codes—that differ little in appearance from the institutional structures of constitutional democracies. Similarly, the language of totalitarian constitutions is often couched in terms of the doctrines of popular rule or democracy. The difference is that in totalitarian regimes neither the institutions nor the constitutional provisions act as effective checks on the power of the single centre: they are essentially facades for the exercise of power through hierarchical procedures that subject all the officials of the state to the commands of the ruling individual or group. The underlying realities of autocratic rule are always the concentration of power in a single centre and the mobilization of force to prevent the emergence of opposition.

Totalitarianism, as already noted, is the most successful of contemporary forms of autocratic rule; it is distinguished from previous forms in its use of state power to impose an official ideology on its citizens. Nonconformity of opinion is treated as the equivalent of resistance or opposition to the government, and a formidable apparatus of compulsion, including various kinds of state police or secret police, is kept in being to enforce the orthodoxy of the proclaimed doctrines of the state. A single party, centrally directed and composed exclusively of loyal supporters of the regime, is the other typical feature of totalitarianism. The party is at once an instrument of social control, a vehicle for ideological indoctrination, and the body from which the ruling group recruits its members.

Constitutional democracy

In the modern world, constitutional democracy is the chief type of nonautocratic government. The minimal definition in institutional terms of a constitutional democracy is that it should provide for a regularized system of periodic elections with a free choice of candidates, the opportunity to organize competing political parties, adult suffrage, decisions by majority vote with protection of minority rights, an independent judiciary, constitutional safeguards for basic civil liberties, and the opportunity to change any aspect of the governmental system through agreed procedures. Two features of constitutional democracy require emphasis in contrasting it with modern totalitarian government: the constitution, or basic law, and the political party. A constitution, as the example of British constitutional democracy suggests, need not be a single written instrument; indeed, the essence of a constitution is that it formalizes a set of fundamental norms governing the political community and determining the relations between the rulers and the people and the interaction among the centres of power. In most modern constitutional democracies, however, there is a constitutional document providing for fixed limitations on the exercise of power. These provisions usually include three major elements: an assignment of certain specified state functions to different state organs or offices, the delimitation of the powers of each organ or office, and the establishment of arrangements for their cooperative interaction; a list of individual rights or liberties that are protected against the exercise of state power; and a statement of the methods by which the constitution may be amended. With these provisions a concentration of power in the hands of a single ruler is prevented, certain areas of political and social life are made immune to governmental intervention, and peaceable change in the political

order is made possible. The political party is the other chief instrument of constitutional democracy, for it is the agency through which the electorate is involved in both the exercise and transfer of power. In contrast with the centralized, autocratic direction of the totalitarian single-party organization, with its emphasis on ideological conformity and restricted membership, the political parties of constitutional democracy are decentralized, concerned with the integration of many interests and beliefs, and open to public participation. In constitutional democracies there is usually some measure of competition among two or more parties, each of which, if it cannot hope to form a future government, has some ability to influence the course of state action. The party in a constitutional democracy is at once a means of representing a mass electorate in the exercise of power and also a device for allowing the peaceful replacement of one set of power holders with another.

The distinction between autocratic and nonautocratic rule should not conceal the existence of a number of intermediate types of government that combine elements of both. In these cases, also, the best procedure for comparative purposes is to investigate the power configurations underlying the formal structures and to examine the extent to which power is concentrated in a single centre or the role that is played by force in the maintenance of the regime. It is a type of analysis that, by guiding attention to the relative weight of coercive and consensual power and the scope of individual freedom in the political order, allows comparisons between systems in terms of their most important attributes.

DEVELOPMENT AND CHANGE

The difficulties of analyzing political change

Political life is shaped by a wide variety of factors, including social and cultural conditions, economic organization, intellectual and philosophical influences, geography or climate, and historical circumstance. Recurrent attempts have been made to reduce this range of variables to analytically manageable dimensions. This is partly the motive, for example, of Marxist and other efforts to relate specific types of political systems to stages of economic development or particular kinds of socioeconomic organization. Although interesting interrelations between political and economic development have been discovered, such monistic, or single-factor, approaches are inadequate to the task of explaining political change. The problem is not only that there are many factors that should be examined but also that they are found in different combinations from one society to another. All political orders are unique as products of history and creations of the peculiar forces and conditions of their environment. A second problem that confronts comparative analysis is the difficulty of devising measures of political development. The definition of what is modern or what constitutes an advanced or developed political system has troubled many recent writers. Clearly, the older notions of development toward the goals of constitutionalism or democracy must now be seriously questioned, and to judge the maturity of a political system in terms of the extent to which it adopts any particular set of institutions or techniques of rule is an equally doubtful procedure. Another difficulty is that political change is not simply a reaction to "objective" factors such as economic forces but also the product of conscious manipulation. In explaining the growth and development of political systems it is impossible to ignore the fact that men, having considered the advantages and disadvantages of different forms of government, often decide to adopt one form rather than another. A similar problem arises from the fact that the nature of the interaction between political systems and their environment is extremely complex. For example, to treat the political system as merely the outgrowth of particular patterns of social or economic organization is to ignore the fact that changes in social and economic structures are often the product, sometimes the intended product, of governmental action.

These difficulties of analysis have prevented the emergence of any satisfactory theory to explain the processes of political change or growth. In the absence of such a theory, however, several writers have recently attempted to identify certain basic phases in the development of national political systems. For example, five major steps in the emergence of the advanced nation-states of the modern world are often distinguished: (1) unification and independence or autonomy; (2) development and differentiation of political institutions and political roles; (3) transfer of power from traditional elites; (4) further institutional and political role differentiation accompanied by the development of a number of organized social interests and growth in governmental functions; and (5) use of state power in attempts to guide or control social and economic activity, extensive exploitation of resources as the result of technological development, and full participation in the international political system.

Steps in the emergence of advanced nation-states

Other writers distinguish among "traditional," "transitional," and "modern" societies in an effort to identify differences and regularities in social, economic, cultural, and political development. The social structure of the traditional society is described as hierarchical, class bound, based on kinship, and divided into relatively few effectively organized social groupings; its economic basis is primarily agricultural, and industry and commerce are relatively undeveloped; its political institutions are those of sacred monarchy, rule by a nobility, and various forms of particularism. The social system of the transitional society is typified by the formation of new classes, especially a middle class and a proletariat, and conflict among ethnic, religious, and cultural groupings; its economic system experiences major tensions as the result of technological development, the growth of industry, urbanization, and the use of rapid communications; its political institutions are typically authoritarian, although constitutional forms also make their appearance. Modernity is seen as the age of high social mobility, equality, universal education, mass communications, increasing secularism, and sociocultural integration; in its economic system, the modern society experiences a further technological revolution, massive urbanization, and the development of a fully diversified economy; its political institutions are those of democracy and modified totalitarianism, and, in either case, a specialized bureaucracy is used to carry on the expanding functions of government.

These efforts to identify stages of "modernization" are poor substitutes for a general theory of political change, but they serve to emphasize the increasing complexity of all the structures—social, economic, and political—of the modern state. The elaboration of the institutions and procedures of modern government appears to be partly a reflection of the social and economic forces at work in the contemporary world and partly the result of efforts to control these forces through governmental action. The complex structures of advanced political orders are treated in the discussion that follows.

II. The structure of government

The study of governmental structures must be approached with great caution, for political systems having the same kind of legal arrangements and using the same type of governmental machinery often function very differently. A parliament, for example, may be an important and effective part of a political system; or it may be no more than an institutional facade of little practical significance. A constitution may provide the framework within which the political life of a state is conducted; or it may be no more than a piece of paper, its provisions bearing almost no relationship to the facts of political life. Political systems must never be classified in terms of their legal structures alone: the fact that two states have similar constitutions with similar institutional provisions and legal requirements should never, by itself, lead to the conclusion that they represent the same type of political system.

To be useful, the study of governmental structures must always proceed hand in hand with an investigation of the actual facts of the political process: the analyst must exercise the greatest care in distinguishing between form and reality and between prescription and practice. Approached in this way, an examination of the organiza-

tional arrangements that governments use for making decisions and exercising power can be a valuable tool of political inquiry.

CONTEMPORARY FORMS OF GOVERNMENT

The rapidity of change

Few states in the modern world have constitutional arrangements that are more than a century old. Indeed, the vast majority of all the world's states have constitutions written in the 20th century. This is true of states, such as Germany, Italy, and Japan, that were defeated in World War II and of other states, such as the Soviet Union, Spain, and China that have experienced civil war and revolutions in the course of the century. Great Britain and the United States are almost alone among major contemporary nation-states in possessing constitutional arrangements that predate the 20th century.

Even in Britain and the United States, the 20th century has seen much change in the governmental system. In the United States, for example, the relationship of legislature and executive at both the national and the state levels has been significantly altered by the growth of bureaucracies and the enlargement of the executive's budgetary powers. In Britain, even more far-reaching changes have occurred in the relationship between the prime minister and Parliament and in Parliament's role in supervising the executive establishment. In both countries, the appearance of the welfare state, the impact of modern technology on the economy, and international crises have resulted in major alterations in the ways in which the institutions of government function and interact.

The modern student of constitutional forms and institutional arrangements confronts an endlessly changing world. In many parts of the world, in countries as different as France, Pakistan, Argentina, Tanzania, and South Vietnam, there are continuing experiments with new constitutions. All systems, moreover, even without formal constitutional change, undergo a continual process of adjustment and mutation as their institutional arrangements respond to and reflect changes in the social order and the balance of political forces.

Decline of traditional monarchy

Contemporary monarchy. The ancient distinction among monarchies, tyrannies, oligarchies, and constitutional governments, like other traditional classifications of political systems, is no longer very descriptive of political life. A king may be a ceremonial dignitary in one of the parliamentary democracies of western Europe, or he may be an absolute ruler in one of the emerging states of North Africa, the Middle East, or Asia. In the first case his duties may be little different from those of an elected president in many republican parliamentary regimes; in the second his role may be much the same as that of countless dictators and strong men in autocratic regimes throughout the underdeveloped areas of the world.

It may be said of the reigning dynasties of modern Europe that they have survived only because they failed to retain or to acquire effective powers of government. Royal lines have been preserved only in those countries of Europe in which royal rule was severely limited prior to the 20th century or in which royal absolutism had never firmly established itself. More successful dynasties, such as the Hohenzollerns in Germany, the Habsburgs in Austria-Hungary, and the Romanovs in Russia, which continued to rule as well as to reign at the opening of the present century, have paid with the loss of their thrones. Today, in countries such as Great Britain or The Netherlands or Denmark, the monarch is the ceremonial head of state, an indispensable figure in all great official occasions and a symbol of national unity and of the authority of the state; but he is almost entirely lacking in power. Monarchy in the parliamentary democracies of modern Europe has been reduced to the status of a dignified institutional facade behind which the functioning mechanisms of government—cabinet, parliament, ministries, and parties—go about the tasks of ruling.

The 20th century has also seen the demise of most of the hereditary monarchies of the non-Western world. Thrones have toppled in Turkey, in China, in most of the Arab countries, in the principates of India, in the tribal kingdoms of Africa, and in several countries of Southeast Asia. The kings who maintain their position do so less by the claim of legitimate blood descent than by their appeal as popular leaders responsible for well-publicized programs of national economic and social reform or as national military chieftains. In a sense, these kings are less monarchs than monocrats, and their regimes are little different from several other forms of one-man rule found in the modern world.

Contemporary dictatorship. While royal rule, as legitimized by blood descent, has almost vanished as an effective principle of government in the modern world, monocracy—a term that comprehends the rule of the remaining non-Western royal absolutists, of the generals and strong men of Latin America and Asia, of the messianic leaders of postcolonial Africa, and of the totalitarian heads of Communist and Fascist states—still flourishes. Indeed, the 20th century, which has witnessed the careers of Kemal Atatürk, Benito Mussolini, Adolf Hitler, Joseph Stalin, Francisco Franco, Mao Tse-tung, Juan Perón, Tito, Gamal Abdel Nasser, Sukarno, Kwame Nkrumah, and Charles de Gaulle, could appear in history as the age of plebiscitary dictatorship.

Monocratic rule

The contemporary world provides examples of several types of monocratic or dictatorial rule. In many of the new states of Africa and Asia, for example, dictators quickly established themselves on the ruins of constitutional arrangements inherited from Western colonial powers. In some of these countries, presidents and prime ministers captured personal power by banning opposition parties and building primitive replicas of the authoritarian one-party systems of the Communist world. In other new countries the armies seized power, and military dictatorships were established. Whether as presidential dictatorships or as military dictatorships, the regimes that came into being appear to have had common roots in the social and economic problems of the new state. The constitutional systems inherited from the colonial powers proved unworkable in the absence of a strong middle class; local traditions of autocratic rule retained a powerful influence; the army, one of the few organized forces in society, was also often the only force capable of maintaining order; and a tiny intellectual class was impatient for economic progress, frustrated by the lack of opportunity, and deeply influenced by the example of authoritarianism in other countries. The dictatorships that resulted proved highly unstable, and few of the individual dictators were able to satisfy for long the demands of the different groups that supported their bids for power.

Although similar in some respects to the dictatorships of the new nations, the caudillos of 19th- and 20th-century Latin America represent a very different type of monocratic rule. In its 19th-century form, *caudillismo* was the result of the breakdown of central authority. After a brief period of constitutional rule, each of the former Spanish colonies in the Americas experienced a collapse of effective national government. A self-proclaimed leader, usually an army officer, heading a private army typically formed from the peasantry with the support of provincial landowners, established his control over one or more provinces, and then marched upon the national capital. The famous 19th-century caudillos—Antonio López de Santa Anna of Mexico or Juan Manuel de Rosas of Argentina, for example—were thus essentially provincial leaders who seized control of the national government to maintain the social and economic power of provincial groups. The 20th-century dictatorships in Latin-American countries have had different aims. The modern caudillo is less a provincial than a national leader. The Perón regime, for example, was established by nationalistic army officers committed to a program of national reform and various ideological goals. Often, too, the dictators of contemporary Latin America have allied themselves with a particular social class, attempting either to maintain the interests of established economic groupings or to press far-reaching social reforms.

Dictatorship in the technologically advanced, totalitarian regimes of modern Communism or Fascism is distinctively different from the authoritarian regimes of either Latin America or the new states of Africa and Asia. Nazi

Germany under Hitler and the Soviet Union under Stalin are the leading examples of modern totalitarian dictatorships. The crucial elements of both were the identification of the state with the single mass party and of the party with its charismatic leader, the use of an official ideology to legitimize and maintain the regime, the employment of a terroristic police force and a controlled press, and the application of all the means of modern science and technology to control the economy and individual behaviour. The two systems of dictatorship, however, may be distinguished in several ways. Fascism, in its National Socialist form, was primarily a counterrevolutionary movement that mobilized middle and lower middle class groups to pursue nationalistic and militaristic goals and whose sole principle of organization was that of obedience to the Führer. By contrast, Soviet Communism grew out of a revolutionary theory of society, pursued the goal of revolutionary overthrow of capitalist systems internationally, and employed the complex bureaucratic structures of the Communist Party as mechanisms of governmental organization.

Western constitutional democracies have provided examples of another type of contemporary dictatorship. At various points in the 20th century, during periods of domestic or foreign crisis, most constitutional regimes have conferred emergency powers on the executive, suspending constitutional guarantees of individual rights or liberties, or declaring some form of martial law. Indeed, the constitutions of some Western democracies explicitly provide for the grant of emergency powers to the executive in a time of crisis in order to protect the constitutional order. In many cases, of course, such provisions have been the instruments by which dictators have overthrown the regime. Thus, the proclamation of emergency rule was the beginning of the dictatorships of Mussolini in Italy, of Kemal Atatürk in Turkey, of Józef Piłsudski in Poland, of António de Olveira Salazar in Portugal, of Franz von Papen and Hitler in Germany, and of Engelbert Dollfuss and Kurt von Schuschnigg in Austria. In other democracies, however, constitutional arrangements have survived quite lengthy periods of crisis government. After World War II, for example, in both the United States and Britain, the use of extraordinary powers by the executive came to a halt with the end of the wartime emergency. Similarly, although the 1958 constitution of the Fifth Republic of France contained far-reaching emergency powers conferred on the president —"when the institutions of the Republic, the independence of the nation, the integrity of its territory or the fulfillment of its international obligations are threatened with immediate and grave danger, and when the regular functioning of the constitutional authority is interrupted" —their implicit threat to the constitutional order has not been realized.

Many forces at work in the 20th century appear to lend impetus to the rise of monocratic forms of rule. In nearly all political systems, the powers of chief executives have increased in response to the demanding social, economic, and military crises of the age. The complex decisions required of governments in a technological era, the perfectionist impulses of the great bureaucratic structures that have developed in all industrialized societies, and the imperatives of national survival in a nuclear world continue to add to the process of executive aggrandizement. The question for many constitutional regimes is whether the limitation and balance of power that are at the heart of constitutional government can survive the growing enlargement of executive power.

Contemporary oligarchy. In the Aristotelian classification of government there were two forms of rule by the few: aristocracy and its debased form, oligarchy. Although the term oligarchy is rarely used to refer to contemporary political systems, the phenomenon of irresponsible rule by small groups has not vanished from the world.

Small leadership groups

Many of the classical conditions of oligarchic rule were found until recently in those parts of Asia in which governing elites were recruited exclusively from a ruling caste—a hereditary social grouping set apart from the rest of society by religion, kinship, economic status, prestige, and even language. In the contemporary world, in some countries that have not experienced the full impact of industrialization, governing elites are still often recruited from a ruling class—a stratum of society that monopolizes the chief social and economic functions in the system. Such elites exercise power in the interests of their class.

The simple forms of oligarchic rule associated with pre-industrial societies are, of course rapidly disappearing. Industrialization produces new, differentiated elites that replace the small leadership groupings that once controlled social, economic, and political power in the society. The demands of industrialization compel recruitment on the basis of skill, merit, and achievement, rather than on the basis of inherited social position and wealth. New forms of oligarchic rule have also made their appearance in many advanced industrial societies. Although governing elites in these societies are no longer recruited from a single class, they are often not subjected to effective restraints on the exercise of their power. Indeed, in some circumstances, the new elites may use their power to convert themselves into a governing class whose interests are protected by every agency of the state. This is the criticism that is made, for example, of the Communist Party bureaucracy in the Soviet Union and in some eastern European states.

Oligarchic tendencies of a lesser degree have been detected in all the great bureaucratic structures of advanced political systems. The growing complexity of modern society and its government thrusts ever greater power into the hands of administrators and committees of experts. Even in constitutional regimes, no fully satisfactory answer has been found to the question of how these bureaucratic decision makers can be held accountable and their powers effectively restrained without, at the same time, jeopardizing the efficiency and rationality of the policy-making process.

Contemporary constitutional government. Constitutional government is defined by the existence of a constitution—which may be a legal instrument or merely a set of fixed norms or principles generally accepted as the fundamental law of the polity—that effectively controls the exercise of political power. The essence of constitutionalism is the control of power by its distribution among several state organs or offices in such a way that they are each subjected to reciprocal controls and forced to cooperate in formulating the will of the state. Although constitutional government in this sense flourished in England and in some other historical systems for a considerable period, it is only recently that it has been associated with forms of mass participation in politics. In England, for example, constitutional government was not harnessed to political democracy until after the Reform Act of 1832 and subsequent 19th-century extensions of the suffrage. In the contemporary world, however, constitutional governments are also generally democracies, and they are usually referred to as constitutional democracies or constitutional–democratic systems.

The contemporary political systems that combine constitutionalism and democracy share a common basis in the primacy they accord to the will of the majority of the people as expressed in free elections. In all such systems, political parties are key institutions, for they are the agencies by which majority opinion in a modern mass electorate is mobilized and expressed. Indeed, the history of the political party in its modern form is coincidental with the development of contemporary constitutional–democratic systems. In each case, the transition from the older forms of constitutionalism to modern constitutional democracy was accompanied by the institutionalization of parties and the development of techniques of party competition. The essential functions of political parties in a constitutional democracy are the integration of a multitude of interests, beliefs, and values into one or more programs or proposals for change and the nomination of party members for elective office in the government. In both functions, the party serves as a link between the rulers and the ruled: in the first case by allowing the electorate

to register an opinion on policy and in the second by giving the people a chance to choose their rulers. Of course, the centralized, autocratically directed, and ideologically orthodox one-party systems of contemporary totalitarian regimes perform neither of these functions.

The British and U.S. systems compared

The two major types of constitutional democracy in the modern world are exemplified by the United States and Great Britain. The United States is the leading example of the presidential system of constitutional democracy; Britain, although its system is sometimes referred to as a cabinet system, in recognition of the role of the Cabinet in the government, is the classic example of the parliamentary system. The U.S. presidential system is based on the doctrine of separation of powers and distinguishes sharply between the personnel, although not between the functions, of the legislature and the executive; the British parliamentary system provides for the integration or fusion of legislature and executive. In the U.S. system the separation of legislature and executive is reinforced by their separate election and by the doctrine of checks and balances that provides constitutional support for routine disagreements between the branches; in the British system the integration of legislature and executive is reinforced by the necessity for their constant agreement, or for a condition of "confidence" between the two, if the normal processes of government are to continue. In the U.S. system, reciprocal controls are provided by such devices as the presidential veto of legislation (which may be overridden by a two-thirds majority in Congress), the Senate's role in ratifying treaties and confirming executive nominations, congressional appropriation of funds, and judicial review of legislation; in the British system, the major control device is the vote of "no confidence" or the rejection of legislation that is considered vital.

The prestige of constitutional democracy was once so great that many thought all the countries of the world would eventually accede to the examples of the United States or Britain and establish similar arrangements. Recent history, however, has confounded this expectation. The collapse of the Weimar Constitution in Germany in the 1930s and the recurrent political crises of the Fourth Republic in France after World War II suggested that constitutional democracy, even in advanced systems, carries no guarantee of stability. The failure of both presidential and parliamentary systems to work as expected in less advanced countries that modelled their constitutions on those of the United States and Britain resulted in a further diminution in the prestige of both systems. Successful examples of established constitutional democracies are now limited to a dozen or so northwest European and North American countries, together with New Zealand and Australia in the Pacific. Systems in which experiments are being made in the use of constitutional-democratic arrangements with some degree of success include Mexico, Uruguay, Chile, the Philippines, India, and Malaya. In all these countries, constitutionalism and democracy have combined to produce political change under conditions of relative stability.

THE LEVELS OF GOVERNMENT

Most national societies have passed through a stage in their social and political development, usually referred to as feudalism, in which a weak and ineffectively organized national government competes for territorial jurisdiction with local power holders. In medieval England and France, for example, the crown was perennially threatened by the power of the feudal nobles, and a protracted struggle was necessary before the national domain was subjected to full royal control. Elsewhere, innumerable societies continued to experience this kind of feudal conflict between local magnates and the central government well into the modern era. The warlords of 19th- and 20th-century China, for example, were just as much the products of feudal society as the warring barons of 13th-century England and presented the same kind of challenge to the central government's claim to exercise sovereign jurisdiction over the national territory. By the 1970s, feudalism was almost extinct. The social patterns that had formerly supported the power of local landowners were rapidly disappearing, and central governments had generally acquired a near monopoly of communications and military technology, enabling them to project their power into areas once controlled by local rulers.

Feudalism

In nearly all national political systems, central governments are better equipped than ever before to exercise effective jurisdiction over their territories. In much of the underdeveloped world, nationalist political movements and a variety of modern economic forces have swept away the traditional structures of local government, and the quasi-autonomous governments of village and tribe and province have been replaced by centrally directed systems of subnational administration. Even in the heavily industrialized states of the modern world, there has been an accelerating tendency toward greater centralization of power at the national level. In the United States, for example, the structure of relationships among the governments at the national, state, and local levels has changed in a number of ways to add to the power of the federal government in Washington. Even though the system of national grants-in-aid appears to have been designed as a means of decentralizing administration, the effect has been decidedly centralist, for the conditional character of the grants has allowed the federal government to exercise influence on state policies in fields that were once invulnerable to national intervention.

The nation-state

Contemporary national government. The nation-state is the dominant type of political system in the contemporary world, and nationalism, or the creed that centres the supreme loyalty of the people upon the nation-state, is the dominating force in international politics. The national ideal triumphed as a result of the wars of the 19th and 20th centuries. The Napoleonic Wars, which spread the doctrines of the French Revolution, unleashed nationalism as a force in Europe and led to the Risorgimento in Italy and the emergence of Bismarck's Germany. The two world wars of the 20th century carried the principles of national self-determination and liberal democracy around the world and gave birth to the independence movements that resulted in the foundation of new states in eastern Europe in 1919 and the emergence from colonial status of countries in Asia and Africa after 1945.

All the major forces of world politics—war and the threat of war, the development of national economies, the demand for social services, and many others—have reinforced the national state as the primary focus of men's loyalties. Wars have played the major part in strengthening national governments and weakening political regionalism and localism. The attachments that men have to subnational political communities are loosened when they must depend for their security on the national power. Even in the new age of total war—which few nations are capable of waging and even fewer of surviving—men look for their security to national governments rather than to international organizations. In nearly all contemporary states, the national budget is dominated by expenditures for defense, the military employs the largest fraction of the work force, and questions of national security pervade the discussion of politics.

The lesson of the last three decades is that national sovereignty remains the most important obstacle to the emergence of new forms of supranational government and to effective international cooperation. Almost everywhere, attempts to achieve federation and other forms of multinational communication have foundered on the rocks of nationalism. The collapse of the Federation of Rhodesia and Nyasaland and the Federation of Malaya are paralleled by the failure of the Organization of American States and the Arab League: in all cases, nationalism has been the divisive force. National sovereignty has been successfully limited in only two major areas of the contemporary world: in eastern Europe, where, as a result of Soviet occupation, a number of formerly independent states were subjected to a form of centralized control from Moscow; and in western Europe, where several nation-states joined in a loose confederation (the European Economic Community) to achieve certain economic goals. Even in these cases, nationalism has been a major threat to continued unity. Several of the countries of

eastern Europe have recently declared their independence of Moscow, and France under President de Gaulle made clear its hostility to any form of supranational union of the countries belonging to the European Economic Community.

At the international level, anarchy is the principal form of contemporary rule, for the nation-state's freedom of action is limited only by its power. Some forces are at work, however, that may eventually lead to a lessening of the role of the nation-state and to the emergence of effective supranational governments. The first of these is industrial and commercial development. The development of national industries in the 19th and early 20th centuries played a major part in strengthening national as against regional and local political entities, but the scale of economic activity has now outgrown national markets. Industrial combines and commercial groupings have emerged that cross national frontiers and require international markets. This integrative influence of economic development may be assisted by the threat of nuclear war. Many national frontiers can no longer be adequately defended in the era of the intercontinental missile, and many states, even some that were until recently great powers, are incompetent to preserve themselves. Together these forces may assist the progress of efforts to develop new kinds of cooperative union among nation-states.

Although the failure of efforts to achieve world government and to develop an effective system of international law may be regretted, it should perhaps be remembered that the nation-state continues to function as an extremely effective system for maintaining order within its boundaries. In some cases, this is achieved with remarkably little coercion and in such a way that the progress of civilization is encouraged. In the conditions of our age, world government might well involve much higher levels of coercion and much less civilization.

Decline of regional and state government

Contemporary regional and state government. The 18th-century political philosopher Montesquieu wrote that governments are likely to be tyrannical if they are responsible for administering large territories, for they must develop the organizational capacity characteristic of despotic states. It was partly this fear that led the American founding fathers to provide for a federal system and to divide governmental functions between the government in Washington and the state governments. Modern technology and mass communication are often said to have deprived Montesquieu's axiom of its force. Yet the technology that makes it possible for large areas to be governed democratically also holds out the spectre of an even greater tyranny than Montesquieu foresaw.

In all political systems the relationships between national and regional or state governments have been affected by technology and new means of communication. In the 18th century Thomas Jefferson, in arguing that local government, or the government closest to the people, was best, could claim that citizens knew most about their local governments, somewhat less about their state governments, and least about the national government. In the present-day United States, however, the concentration of the mass media on the issues and personalities of national government has made nonsense of this proposition. As several recent studies have demonstrated, people know much less about local government than national government and turn out to vote in much larger numbers in national elections. The necessity for employing systems for the devolution of political power is reduced when a central government can communicate directly with citizens in all parts of the national territory, and the vitality of subnational levels of government is sapped when public attention is focussed on national problems.

Another general development that has lessened the importance of regional or state government is the rise of efficient national bureaucracies. In nearly all political systems, there has been some tendency toward bureaucratic centralization, and in some cases national bureaucracies have almost completely replaced older systems of regional and provincial administration. In the United States, for example, complex programs of social security, veterans' benefits, income taxes, agricultural subsidies, and many others that bear directly on individuals are centrally administered; and great agencies of government, such as the Department of Health, Education, and Welfare, maintain regional and local offices for the supervision of other national activities.

Even in systems in which a division of functions between national and subnational governments is constitutionally prescribed, the prevailing trend in intergovernmental relations is toward increasing involvement of the national government in areas once dominated by regional or state governments. Thus, the original constitutional arrangements prescribed by the Allied powers for the West German republic in 1949 won general acclaim at the time because they provided for greater decentralization than had the Weimar Constitution; but, as soon as Germany was free to amend its own constitution, several state functions were reassigned to the national government. In the United States, also, the collapse of the doctrine of "dual federalism," according to which the powers of the national government were restricted by the powers reserved to the states, signalled the end of an era in which the states could claim exclusive jurisdiction over a wide range of functions. Today, forms of cooperative federalism involving joint action by national and state governments are increasingly common. Such cooperative relationships in the United States include programs of public assistance, the interstate-highway system, agricultural extension programs, and aid to education. In some areas, such as school desegregation, the national government has used broad powers to compel states to conform to national standards.

Efforts made to halt the trend toward centralization and to re-invigorate regional or state governments have met with little success. In the U.S. a Commission on Intergovernmental Relations established by President Eisenhower in 1953 concluded that it could recommend no major reversion of functions to state governments. Similarly, efforts in France and Italy to decentralize parts of the national administrations have had few practical results. Political regionalism appears to be in steep decline almost everywhere, whether in China or in the American South. The attachments that bind men to localities and allow the growth of genuine subnational political communities have weakened under the impact of technology and the growth of national economies. Only where political regionalism has always been a cloak for movements of national independence—for example, in Scotland, Wales, Northern Ireland, Quebec, and Brittany—are there popular attempts to reverse the trend toward national centralization.

The urban metropolis

Contemporary city and local government. Political scientists since Aristotle have recognized that the nature of political communities changes when their populations grow larger. One of the central problems of contemporary government is the vast increase in urban population and the progression from "polis to metro-polis to mega-polis." The catalog of ills that have resulted from urban growth includes political and administrative problems of extraordinary complexity.

The rise of great urban centres has had a major impact on the politics of all advanced industrial societies. In the United States, for example, the issues of metropolitan politics have become key political issues at the national level. Programs relating to urban renewal, mass-transit systems, school desegregation, and air and water pollution have become of pressing national importance, along with other social problems of the metropolis, such as urban crime, juvenile delinquency, the disorganization of the family, ghetto conditions, and unemployment. The growing involvement of the national government with such issues is in part a consequence of the impoverishment of the cities and the collapse of effective municipal government. The tax base of U.S. cities has dwindled with the flight of the middle classes to the suburbs and the relocation of industry; at the same time, although "bossism" and machine politics have almost vanished, newer forms of city governments have failed to establish effective systems of administration or to win support from

stable majorities of city voters. In part, too, the national government gives priority to such issues because of the political power of the urban areas: cities have become the focuses of national politics, for they are rich in the votes that decide national campaigns. Aside from such fiscal and political pressures, the national government is inevitably concerned with the threat posed by racial conflict, ghetto violence, and other kinds of social chaos in the city.

The metropolis suffers from several acute governmental and administrative failures. Responsibility for the issues that transcend the boundaries of local governments has not been defined, for representative institutions have failed to develop at the metropolitan level. In most cases, there are no effective governmental structures for administering area-wide services or for dealing comprehensively with the common problems of the metropolitan community. The result has been the appearance of a new class of problems created by government itself, including uneven levels of service for metropolitan residents, inequities in financing government services and functions, and variations in the democratic responsiveness of the governments scattered through the metropolitan area. The tangled pattern of local governments, each operating in some independent sphere, does not allow the comprehensive planning necessary to deal with the escalating problems of urban life.

Efforts to create new governing structures for metropolitan communities are among the most interesting developments in contemporary government. In the United States these efforts include the creation of special districts to handle specific functions, area-wide planning agencies, interstate compacts, consolidated school and library systems, and various informal intergovernmental arrangements. Although annexation of outlying areas by the central city and city–county consolidations have been attempted in many cases, the reluctance of urban areas to surrender their political independence or to pay for central-city services has been an obstacle. The Los Angeles plan, by which the county has assumed responsibility for many area-wide functions, leaving the local communities with substantial political autonomy, may represent a partial solution to the problem of urban–suburban tensions. In other cases, "metropolitan federation" has been attempted. One of the earliest and most influential examples of a federated system of metropolitan government is the Greater London Area in Britain, which encompasses 32 London boroughs and places effective governing powers in the hands of an elected city council. In Canada the city of Toronto and its suburbs adopted a metropolitan "constitution" in 1953 under which mass transit, highways, planning, and several other functions are controlled by a council composed of elected officials from the central city and surrounding governments. In 1957 Miami and Dade counties in Florida chose "metro" government: Miami and 27 suburban cities retained control of local functions, while area-wide functions such as fire and police protection and transportation were allocated to the new federal structure. In 1962 Nashville, Tennessee, combined with Davidson County to form a single metropolitan government; and in 1968 Jacksonville, Florida, joined with Duval County in a similar arrangement. Other examples of various degrees of area-wide consolidation in the United States include Baton Rouge, Louisiana; Seattle, Washington; Portland, Oregon; and Indianapolis, Indiana.

Most of the major problems of contemporary politics seem to have found their focus in the metropolis, and there is almost universal agreement that new governing systems must be devised for the metropolitan community if the problems are ever to be resolved.

THE DIVISIONS OF GOVERNMENT

Legislative, executive, and judicial functions

In his *Politics*, Aristotle differentiated three categories of state activity—deliberations concerning common affairs; decisions of executive magistrates, and judicial rulings—and indicated that the most significant differences among constitutions concerned the arrangements made for these activities. This threefold classification is not precisely the same as the modern distinction among legislature, executive, and judiciary. Aristotle intended to make only a theoretical distinction among certain state functions and stopped short of recommending that they be assigned as powers to separate organs of government. Indeed, since Aristotle held that all power should be wielded by one man, pre-eminent in virtue, he never considered the concept of separated powers. In the 17th century the English political philosopher John Locke also distinguished the legislative from the executive function but, like Aristotle, failed to assign these to separate organs or institutions. Montesquieu was the first to make the modern division among legislative, executive, and judiciary. Arguing that the purpose of political association is liberty, not virtue, and that the very definition of liberty's great antagonist, tyranny, is the accumulation of all power in the same hands, he urged the division of the three functions of government among three separate institutions. After Montesquieu, the concept of separation of powers became one of the principal doctrines of modern constitutionalism. Nearly all modern constitutions, from the document written at Philadelphia in 1787 through the French Declaration of the Rights of Man and of the Citizen of August 1789 up to the constitutions of the new states of Africa and Asia, provide for the separate establishment of legislative, executive, and judiciary. The functional division among the branches of government is never precise. In the American system, for example, the doctrine of checks and balances justifies several departures from the strict assignment of functions among the branches. Parliamentary forms of government depart even further from the concept of separation and integrate both the personnel and the functions of the legislature and the executive. Indeed, the principle of shared rather than separated powers is the true essence of constitutionalism. In the constitutional state, power is controlled because it is shared or distributed among the divisions of government in such a way that they are each subjected to reciprocal checks and forced to cooperate in the exercise of political power. In the nonconstitutional systems of totalitarianism or autocracy, although there may be separate institutions such as legislatures, executives, and judiciaries, power is not shared but rather concentrated in a single organ. Because this organ is not subjected to the checks of shared power, the exercise of political power is uncontrolled or absolute.

The legislature. The characteristic function of all legislatures is the making of law. In most systems, however, legislatures also have other tasks, such as selection and criticism of the government, supervision of administration, appropriation of funds, ratification of treaties, impeachment of executive and judicial officials, acceptance or refusal of executive nominations, determination of election procedures, and public hearings on petitions. Legislatures, then, are not simply lawmaking bodies. Neither do they monopolize the function of making law. In most systems the executive has a power of veto over legislation, and, even where this is lacking, the executive may exercise original or delegated powers of legislation. Judges, also, often share in the lawmaking process, through the interpretation and application of statutes or, as in the U.S. system, by means of judicial review of legislation. Similarly, administrative officials exercise quasi-legislative powers in making rules and deciding cases that come before administrative tribunals.

Varieties of legislatures

Legislatures differ strikingly in their size, the procedures they employ, the role of political parties in legislative action, and in their vitality as representative bodies. In size, the British House of Commons, with 630 members, is among the largest; the Icelandic lower house, with 40, the New Zealand House of Representatives, with 80, and the Senate of Nevada, with 17 members, are among the smallest. Most legislatures are bicameral, although New Zealand, Denmark, the state of Queensland, in Australia, and Nebraska, in the United States, have all abolished their second chambers. The procedures of the United States House of Representatives, which derive from a manual of procedure written by Thomas Jefferson, are among the most elaborate of parliamentary rules, requir-

ing study and careful observation over a considerable period before members become proficient in their manipulation. Voting procedures range from the formal procession of the division or teller vote in the British House of Commons to the electric voting methods employed in the California legislature and in some other American states. Another point of difference among legislatures concerns their presiding officers. These are sometimes officials who stand above party and, like the speaker of the British House of Commons, exercise a neutral function as parliamentary umpires; sometimes they are the leaders of the majority party and, like the speaker of the United States House of Representatives, major political figures; and sometimes they are officials who, like the vice president of the United States in his role as presiding officer of the Senate, exercise a vote to break ties and otherwise perform mainly ceremonial functions.

Legislative parties are of various types and play a number of roles or functions. In the United States House of Representatives, for example, the party is responsible for assigning members to all standing committees; the party leadership fills the major parliamentary offices, and the party membership on committees reflects the proportion of seats held by the party in the House as a whole. The congressional party, however, is not disciplined to the degree found in British and some other European legislative parties, and there are relatively few "party line" votes in which all the members of one party vote against all the members of the other party. In the House of Commons, party-line voting is general; indeed, it is very unusual to find members voting against their party leadership, and, when they do, they must reckon with the possibility of penalties such as the "withdrawal of the whip" or the loss of their official status as party members.

It is often said that the 20th century has dealt harshly with legislatures and that this is an age of executive aggrandizement. Certainly, executives in most countries have assumed an increasingly large role in the making of law, through the initiation of the legislation that comes before parliaments, assemblies, and congresses, through the exercise of various rule-making functions, and as a result of the growth of different types of delegated legislation. It is also true that executives have come to predominate in the sphere of foreign affairs and, by such devices as executive agreements, which are frequently used in place of treaties, have freed themselves from dependence upon legislative approval of important foreign-policy initiatives. Moreover, devices such as the executive budget and the rise of specialized budgetary agencies in the executive division have threatened the traditional fiscal controls of legislatures. This decline in legislative power, however, is not universal. The United States Congress, for example, has preserved a substantial measure of its power. Indeed, congressional oversight of the bureaucracy is an area in which it has added to its power and has developed new techniques for controlling the executive. The difficulties of recent presidents with legislative programs of foreign aid and the insistent congressional criticism of executive policies in Southeast Asia also suggest that Congress continues to play a vital role in the governing process.

The executive. Political executives are government officials who participate in the determination and direction of government policy. They include heads of state and government leaders—presidents, prime ministers, premiers, chancellors, and other chief executives—and many secondary figures, such as cabinet members and ministers, councillors, and agency heads. By this definition, there are several thousand political executives in the U.S. national government, including the president, dozens of political appointees in the Cabinet departments, in the agencies, in the commissions, and in the White House staff, and hundreds of senior civil servants. The same is true of most advanced political systems, for the making and implementation of government policy require very large executive and administrative establishments.

The crucial element in the organization of a national executive is the role assigned to the chief executive. In presidential systems, such as in the United States, the president is both the political head of the government and also the ceremonial head of state. In parliamentary systems, such as in Great Britain, the prime minister is the national political leader, but another figure, a monarch or elected president, serves as the head of state. In mixed presidential–parliamentary systems, such as that established in France under the constitution of 1958, the president serves as head of state but also wields important political powers, including the appointment of a prime minister and Cabinet to serve as the government.

Varieties of executives

The manner in which the chief executive is elected or selected is often decisive in shaping his role in the political system. Thus, although he receives his seals of office from the monarch, the effective election of a British prime minister usually occurs in a private conclave of the leading members of his party in Parliament. Elected to Parliament from only one of 630 constituencies, he is tied to the fortunes of the legislative majority that he leads. By contrast, the American president is elected by a nationwide electorate, and, although he leads his party's ticket, his fortunes are independent of his party. Even when the opposition party controls the Congress, his fixed term and his independent base of power allow him considerable freedom of manoeuvre. These contrasts explain many of the differences in the roles of the two chief executives. The British prime minister is invariably a man who has served for many years in Parliament and has developed skills in debate and in political negotiation. His major political tasks are the designation of the other members of the Cabinet, the direction of parliamentary strategy, and the retention of the loyalty of a substantial majority of his legislative party. The presidential chief executive, on the other hand, often lacks prior legislative and even national-governmental experience, and his main concern is with the cultivation of a majority in the electorate through the leadership of public opinion. Of course, since the president must have a legislative program and often cannot depend on the support of a congressional majority, he may also need the skills of a legislative strategist and negotiator.

Another important area of contrast between different national executives concerns their role in executing and administering the law. In the U.S. presidential system, the personnel of the executive branch are constitutionally separated from the personnel of Congress: no executive officeholder may seek election to either house of Congress, and no member of Congress may hold executive office. In parliamentary systems the political management of government ministries is placed in the hands of the party leadership in parliament. In the U.S. system the president often appoints men to Cabinet positions who have had little prior experience in politics, and he may even appoint members of the opposition party. In the British system, Cabinet appointments are made to consolidate the prime minister's personal ascendancy within the parliamentary party or to placate its different factions. These differences extend even further into the character of the two systems of administration and the role played by civil servants. In the U.S. system a change in administration is accompanied by the exodus of a very large number of top government executives—the political appointees who play the vital part in shaping day-to-day policy in all the departments and agencies of the national government. In Britain, when political control of the House of Commons changes, only the ministers, their parliamentary secretaries, and one or two other top political aids are replaced. For all practical purposes, the ministries remain intact and continue under the supervision of permanent civil servants.

In nearly all political systems, even in constitutional democracies where executive responsibility is enforced through free elections, the 20th century has seen an alarming increase in the powers of chief executives. The U.S. presidency, like the office of prime minister in Britain, has greatly enlarged the scope of its authority. One of the future tasks of representative government is to develop more constitutional restraints on the abuse of executive powers while retaining their advantages for effective rule.

The judiciary. Like legislators and executives, judges are major participants in the policy-making process; and courts, like legislatures and administrative agencies, promulgate rules of behaviour having the nature of law. The process of judicial decision making, or adjudication, is distinctive, however, for it is concerned with specific cases in which an individual has come into conflict with society by violating its norms or in which individuals have come into conflict with one another, and it employs formal procedures that contrast with those of parliamentary or administrative bodies.

Varieties of court systems

Established court systems are found in all advanced political systems. Usually there are two judicial hierarchies, one dealing with civil and the other with criminal cases, each with a large number of local courts, a lesser number at the level of the province or the region, and one or more courts at the national level. This is the pattern of judicial organization in Britain, for example. In some countries—for example, in France—although there is a double hierarchy, the distinction is not between courts dealing with criminal cases and other courts dealing with civil cases but rather between those that handle all civil and criminal cases and those that deal with administrative cases or challenges to the administrative authority of the state. Reflecting the federal organization of its government, the United States possesses two court systems: one set of national courts and 50 sets of state courts. By contrast, West Germany and the Soviet Union, both of which are federal in governmental organization, each possess only one, integrated court system.

Local courts are found in all systems and are usually of two types. The first type deals with petty offenses and may include a traffic court, a municipal court, a small-claims court, and a court presided over by a justice of the peace or a local magistrate. The second type, sometimes called trial courts, are courts of first instance in which most cases of major importance are begun. These are the state superior courts in the United States, the county courts and quarter sessions in Britain, the *tribunal de grande instance* in France, and the district courts, or *Landgerichte*, in Germany. In some systems there is a level above the local court, usually referred to as assize courts, in which exceptionally serious crimes, such as homicide, are tried. Courts of appeal review the procedures and the law in the lower court and either reverse or affirm its judgment or, in some cases, return the case for a new trial. In all systems there are national supreme courts that hear appeals and exercise original jurisdiction in cases of the greatest importance, such as those involving conflict between a state and a national government. Outside the regular court systems, there are sometimes found specialized judicial tribunals, such as administrative courts, or courts of claims that deal with special categories of cases.

III. The functions of government

In all modern states, governmental functions have greatly expanded with the emergence of government as an active force in guiding social and economic development. In Communist countries, government has a vast range of detailed controls over all types of social and economic behaviour. Even in the United States, where there remains a much greater attachment than in most societies to the idea that government should be only an umpire adjudicating the rules by which other forces in society compete, such governmental activities as the Tennessee Valley Authority or the use of credit controls to prevent economic fluctuations are now accepted with relatively little question. Government has thus become the major or even the dominant organizing power in all contemporary societies.

The historical stages by which governments have come to exercise their contemporary functions make an interesting study in themselves. The scope of government in the ancient polis involved the comprehensive regulation of the ends of human existence. As Aristotle expressed it, what was not commanded by government was forbidden. The extent of the functions of government in the ancient world was challenged by Christianity and its existence on a division of those things that belong separately to Caesar and to God. When the feudal world succeeded the Roman Empire, however, the enforcement of the sanctions of religion became one of the first objects of political authority. The tendencies that began in the 18th century separated church from state and state from society, and the modern concept of government came into being. The American Colonies' Declaration of Independence expresses the classic modern understanding of those ends that governmental functions exist to secure. The first aim of government is to secure the right to life; this comprehends the safety of fellow citizens as regards one another and the self-preservation of the nation as regards foreign powers. Life exists for the exercise of liberty, in terms of both natural and civil rights, and these, along with other specific functions of government, provide those conditions upon which men may pursue happiness, an end that is finally entirely private and beyond the competence of government.

With the advent of the Marxist conception of the state, the ends of human existence once again became the objects of comprehensive government regulation. Marxism sees the state as a product of class warfare that will pass out of existence in the future age of perfect freedom. Aristotle believed human perfection to be possible only within political society; Marx believed that the perfection of man would follow upon the abolition of political society. Before the final disposal of the state, however, many Marxists believe that forceful use of governmental power is justified in order to hasten mankind's progress toward the last stage of history.

THE INDISPENSABLE FUNCTIONS OF GOVERNMENT

The maintenance of authority

The first right of men and nations is self-preservation. The task of maintaining the nation, however, is more complex than the individual's duty of self-preservation, for the nation must seek to command the attachment of a community of citizens as well as to preserve itself from external violence. As Thomas Hobbes insisted, civil war constitutes the greatest threat to governments, for it represents the dissolution of the "sovereign power." In modern terms, civil war signifies that the government has lost one of the basic attributes of political authority: its monopoly of force and its control over the use of violence. In a fundamental sense, political authority may be preserved from the threat of civil war only when there exists in the political community an agreement on the basic principles of the regime. Such a consensus is the result, among other things, of a shared "ideology" that gives fellow citizens a sense of communal belonging and recognizes interlocking values, interests, and beliefs. Ideology, in this sense, may be the product of many different forces. Sometimes it is associated with ancient customs, sometimes with religion, sometimes with severe dislocations or the sort of common need that has led to the formation of many nation states, and sometimes with the fear of a common enemy. The ideological commitment that men call patriotism is typically the product of several of these forces.

Governments neglect at their peril the task of strengthening the ideological attachment of their citizens to the regime. In this sense, civic education should be counted among the essential functions of the state, for it is primarily through systems of education that citizens learn their duties. Indeed, as a number of recent sociological studies have shown, the process of political socialization that transforms men into citizens begins in kindergarten and grade school. Even more than this, education is the instrument by which governments further the cohesion of their societies and build the fundamental kinds of consensus that support their authority. It is not surprising, therefore, that national systems of education are often linked to central elements of the regimes. The educational system of the Soviet Union emphasizes political indoctrination; in France, public education is traditionally mixed with the teachings of the Roman Catholic Church; in Great Britain, until very recently, a private system of education supported the class divisions of society; and in the United States a primarily secular form of public edu-

cation has traditionally used constitutional documents as the starting point of children's training in patriotism.

The preservation of the authority of the state also requires a governmental organization capable of imposing its jurisdiction on every part of the national territory. This involves the maintenance of means of communication, the use of administrative systems, and the employment of police forces capable of controlling domestic violence. The police function, like education, is often a key to the character of a regime. In Nazi Germany, Hitler's Brownshirts took over the operation of local and regional police systems and often supervised the administration of law in the streets. In the Soviet Union, the security police acts to check any deviation from the policy of the party or state. In the United States the police powers are left in the hands of the 50 states and the local agencies of government. They serve to check civil and criminal offenses; with the exception of certain offenses created by the McCarran Act and some parallel statutes, political crimes as such are unknown. In addition, there are state militias that act, under the control of the governors of the various states, in moments of local emergency, such as riots or natural catastrophes. The Federal Bureau of Investigation (FBI), the only equivalent of a national police force, is an agency established to carry out specific assignments dealing with a limited but important class of crimes. Since there is no comprehensive federal criminal code, there is not, strictly speaking, a federal police.

Governments must preserve themselves against external as well as domestic threats. For this purpose they maintain armed forces and carry on intelligence activities. They also try to prevent the entry of aliens who may be spies or saboteurs, imprison or expel the agents of foreign powers, and embargo the export of materials that may aid a potential enemy. The ultimate means of preserving the state against external threats, of course, is war. In war, governments usually enlarge the scope of their domestic authority; they may raise conscript forces, imprison conscientious objectors, subject aliens to internment, sentence traitors to death, impose extraordinary controls on the economy, censor the press, compel settlement of labour disputes, impose internal-travel limitations, withhold passports, and provide for summary forms of arrest.

Many forces generate clashes between nations, including economic rivalry and disputes over trade, the desire to dominate strategic land or sea areas, religious or ideological conflict, and imperialistic ambition. All national governments develop organizations and policies to meet these and other situations. They have foreign ministries for the conduct of diplomatic relations with other states, for representing them in the United Nations and other international organizations, and for negotiating treaties. Some governments conduct programs such as foreign aid, cultural exchange, and other activities designed to win goodwill abroad.

In the 20th century, relationships among governments have been affected by a developing awareness that world peace depends upon multinational and international cooperation. The League of Nations and the United Nations, together with their associated agencies, represent major efforts to establish substitutes for traditional forms of diplomacy. Regional alliances and joint efforts, such as the Organization of American States, the North Atlantic Treaty Organization, the European Economic Community, and the Organization of African Unity, represent another type of cooperation among nations.

THE SUPERVISORY FUNCTIONS OF GOVERNMENT

Mediation of conflicts

The conflict of private interest is the leading characteristic of the political process in constitutional democracies, and the supervision, mediation, arbitration, and adjudication of such conflicts are among the key functions of their governments. Representative institutions are themselves a device for the resolution of conflict. Elections in constitutional democracies provide opportunities for mass participation in a process of open debate and public decision; assemblies, congresses, and other parliamentary institutions provide for public hearings on major issues of policy and require formal deliberative procedures at different stages of the legislative process; and political parties integrate a variety of interests and effect compromises on policy that win acceptance from many different groups.

If the interests that compete in the political process are too narrow or restricted, efforts may be made to control or change the rules of competition. Thus, laws have been enacted that seek to prevent discrimination from locking racial and other minorities out of the democratic process; the franchise has been extended so as to permit all groups, including minorities such as women, blacks, and 18-year-olds, to have a voice in elections; and government bodies such as courts and administrative agencies enforce legislation against groups considered to be too large or monopolistic.

Judicial processes offer a means by which some disputes in society are settled according to rule and legal authority, rather than by political struggle. In all advanced societies, law is elaborated in complex codes governing rights and duties and procedural methods; and court systems are employed that adjudicate disputes in terms of the law. In constitutional systems such as the United States, the judiciary is deeply involved in the process of public decision making; the courts actually produce much of the substantive law that bears on private individuals and economic groups in society.

THE REGULATORY FUNCTIONS OF GOVERNMENT

Regulation of the economy

Government regulation of economic life is not a new development. The national mercantilist systems of the 18th century provided for regulation of the production, distribution, and export of goods by government ministries; even during the 19th century, governments continued to intervene in the economy. The government of the United States, for example, from its inception in 1789, allotted funds or subsidies for the support of agriculture, maintained a system of tariffs for its own revenue and the support of domestic manufacturers, patronized the arts and sciences, and engaged in various kinds of public works to advance commerce and promote the general welfare. In France, even more elaborate governmental schemes of economic regulation were practiced throughout the 19th century, including a variety of Socialist experiments such as the Public Workshops that Louis Blanc established in Paris in 1848. In Britain, the various factory acts of the 19th century represented an effort by government to control working conditions in industry.

After World War II the ability of a government to regulate or control the economy became one of the chief tests of its success, and regulatory agencies multiplied to the point at which they are now often referred to as "the fourth branch of government." The extent of the controls imposed on the economy is one of the principal distinctions among capitalist, Socialist, and Communist systems. In Communist countries it is a matter of doctrine that the means of production should be owned and therefore controlled by the state; in the 1960s and 1970s, however, some Communist countries experimented with various ways of decentralizing economic decision making in an effort to increase flexibility and reduce the weight of the central bureaucracy. In Britain the Labour governments nationalized some major industries, including coal, steel, and the railroads, prompted partly by Socialist doctrine and partly by the failure of British industry to remain competitive in international markets. In the United States the government has involved itself in the economy primarily through its regulatory powers. In France the government has gone further and has engaged in national economic planning in cooperation with private business organizations.

The regulation of industrial conditions and of labour–management relations has been a major concern of most Western governments. In the United States the first regulatory efforts in this field were made during the Progressive era at the turn of the 20th century, when the wages, hours, and working conditions of women and children in industry became a matter of public scandal. A little later, the conditions, hours, food, and wages of merchant seamen were brought under government regulation; an eight-hour day was set for railway crews; and workmen's

compensation laws were instituted. With the Depression of the 1930s, minimum wages were introduced for workers in many industries, hours of work were set, and the right to collective bargaining was given legal sanction.

Regulation of transportation has been another major activity in most Western political systems, beginning with the railroads. In the United States their monopolistic practices attracted the criticism of agricultural interests and led eventually to the Interstate Commerce Act of 1887, which regulated railroad rates; subsequent legislation covered the hours, conditions, and wages of railroad employees, among other things. Other modes of land and air transportation have since been brought under regulatory controls implemented by government agencies.

In most European countries, facilities of communication—the telegraph, telephone, radio, and television—are owned and operated by the government. In the United States most of these facilities have remained in private ownership, although they are regulated by the Federal Communications Commission. The regulation by government of important instruments of public opinion such as radio, television, and newspapers has important implications for the freedoms of speech and press and other individual rights. In the United States and Great Britain, government censorship of the press and other media has been restricted to matters of national security. This is also generally true of other Western constitutional democracies, although the celebrated *Der Spiegel* affair in Germany in 1962 and some extraordinary controls imposed on the media in France have been widely criticized. In Communist countries and in many of the new states of Africa and Asia, very extensive controls are imposed on the press, and government-owned newspapers are often the principal sources of political news.

Other forms of government regulation of the economy involve the use of taxes and tariffs, the regulation of weights and measures, and the issuance of money.

Protection of political and social rights

To some extent, all modern governments assume responsibility for protecting the political and social rights of their citizens. The protection of individual rights has taken two principal forms: first, the protection of liberty in the face of governmental oppression; second, the protection of individual rights against hostile majorities and minorities. Even Communist governments recognize certain rights of individuals and minorities in the ordinary spheres of life, if not in matters touching the interests of the state. In the 1960s and 1970s there appeared to be a broadening of the sphere of private rights in the Soviet Union. Authors, artists, and scientists were able to engage in criticism of official policies, not with impunity but with much greater freedom than in the Stalin era; and there were some signs that the Soviet state had begun to recognize the rights of the Jewish minority and the claims of some other national groups. In the United States the Supreme Court expanded the rights of the criminally accused; and, after 1954, the national government acted to bar legal discrimination against ethnic minorities. Indeed, almost all the freedoms detailed in the first ten amendments to the Constitution have been extended since World War II.

Another type of government regulation bearing on the individual concerns the law of immigration and emigration. The great mass migrations of the 19th and early 20th centuries came to an abrupt halt after 1914 with the proliferation of government controls on the freedom of movement across national boundaries.

THE ENTERPRISING FUNCTIONS OF GOVERNMENT

Provision of goods and services

All modern governments participate directly in the economy, purchasing goods, operating industries, providing services, and promoting various economic activities. One of the indispensable functions of government—national defense—has made governments the most important consumers of goods, and they have not hesitated to use their resulting pricing, purchasing, and contracting powers to achieve various economic aims. In efforts to avoid dependence on private sources of strategic goods and defense materials, some governments have taken a further step and established their own military production plants. In wartime, governments have assumed control over entire industries and have subjected the work force to military direction.

In nearly all political systems, certain functions are recognized as public, or belonging to the government. In addition to national defense, these include the maintenance of domestic peace, public education, fire protection, traffic control, aid to the indigent, conservation of natural resources, flood control, and postal services. But governments have assumed responsibility for many other commercial operations, even in non-Socialist countries. The sale of electric power, for example, is one of the established enterprising functions of national, state, and local governments in the United States and Canada. Municipally owned power utilities exist in about 2,000 cities in North America; 14 states of the U.S. have public power districts; and the U.S. government markets power through the Tennessee Valley Authority, the Bureau of Reclamation, the Bonneville Power Administration, and the Southwestern and Southeastern Power administrations. Another range of functions is performed by other national agencies, including the Rural Electrification Administration, which loans money to rural cooperatives to finance local power projects; the Export–Import Bank of Washington, which makes loans to finance export and import trade; the Maritime Administration, which holds mortgages on ships; and the Veterans Administration, which makes loans for farm or home purchases to military veterans. The United States government, acting through agencies such as the Social Security Board, the Federal Deposit Insurance Corporation, and the Federal Housing Administration, is also the largest insurer in the nation. Through other agencies, such as the Housing and Home Finance Agency, the Urban Renewal Administration, and the Public Works Administration, the national government has also developed a major role in the construction and rental of residential housing for low-income persons.

Other miscellaneous enterprises in which governments are involved include the provision of health care, the operation of public transport facilities, the development of public works, airport and port maintenance, and water-supply systems. In Great Britain the government operates hospitals and provides medical care under the National Health Service. In the United States, many state and local governments operate hospitals on a commercial basis, although providing some charity care. At the local level in the United States, the Port of New York Authority constructs and operates bridges, terminals, and airports. The states in the Delaware Basin recently joined in a compact to establish an agency to control the use of water from the basin, institute programs to prevent pollution, provide recreation facilities, transmit and sell hydroelectric power, and provide watershed management. Cities in the United States and Canada operate more than 70 urban transit systems, 600 municipal gas utilities, and more than 4,000 water-supply systems. Cities are also generally responsible for garbage disposal and sometimes operate commercial slaughterhouses, coal yards, laundries, ice plants, and golf courses. Finally packaged liquor sales are often made by the state governments, at either wholesale or retail or both.

PUBLIC ADMINISTRATION

Civil service

While the functional objectives of government administration vary from system to system, all countries that are technologically developed have evolved systems of public administration. A number of common features may be detected in all such systems. The first is the hierarchical, or pyramidal, character of the organization by which a single chief executive oversees a few subordinates, who in turn oversee their chief subordinates, who are in turn responsible for overseeing other subordinates, and so on, until a great structure of personnel is integrated and focussed on the components of a particular program. A second common feature is the division of labour or specialization within the organization. Each individual in the hierarchy has specialized responsibilities and tasks. A third feature is the maintenance of detailed official re-

cords and the existence of precise paper procedures through which the personnel of the system communicate with each other and with the public. Finally, tenure of office is also characteristic of all public bureaucracies.

The various national civil services, despite their similarities, also show important differences, particularly in the way in which individuals are recruited and in the status accorded them in the political system. The British civil service, for example, has traditionally been composed of three classes, or grades—clerical, executive, and administrative. Administrative civil servants, the highest grade, are recruited by examination from among recent university graduates. The top managers of the different government ministries are drawn from this elite group. They remain in office despite changes in government and are accorded immense prestige in the British political system. The U.S. civil service is organized into 18 grades. Although promotion from the lower grades is the typical means by which positions in the top grades are filled, there is also a flow of individuals into senior positions from private business and the professions. The U.S. equivalent of the administrative civil servant in Britain is usually a political appointee recruited by each new administration from private life or from a position in politics.

IV. Development and change in political systems

The student of political systems grapples with a subject matter that is today in constant flux. He must deal not only with the major processes of growth, decay, and breakdown but also with a ceaseless ferment of adaptation and adjustment. The magnitude and variety of the changes that occurred in the world's political systems between the second and eighth decades of the 20th century suggest the dimensions of the problem. Great empires disintegrated; nation-states emerged, flourished briefly, and then vanished; world wars twice transformed the international system; new ideologies swept the world and shook established groups from power; all but a few nations experienced at least one revolution and many nations two or more; domestic politics in every system were contorted by social strife and economic crisis; and everywhere the nature of political life was changed by novel forms of political activity, new means of mass communication, the enlargement of popular participation in politics, the rise of new political issues, the extension of the scope of governmental activity, the threat of nuclear war, and innumerable other social, economic, and technical developments.

STABILITY AND INSTABILITY

Although it is possible to identify a number of factors that obviously have a great deal to do with contemporary development and change in the world's political systems —industrialization, population growth, the "revolution of rising expectations" in the less developed countries, and international tensions—there is no agreed theory to explain the causes of political change. Some social scientists have followed Aristotle's view that political instability is generally the result of a situation in which the distribution of wealth fails to correspond with the distribution of political power and have echoed his conclusion that the most stable type of political system is one based on a large middle class. Others have adopted Marxist theories of economic determinism that view all political change as the result of changes in the mode of production. Still others have focussed on governing elites and their composition and have seen in the alienation of the elite from the mass the prime cause of revolutions and other forms of violent political change.

In the discussion that follows, a distinction is drawn between unstable and stable political systems, and an attempt is made to suggest ways of understanding the processes of political development and change.

Crisis situations

Unstable political systems. In modern times the great majority of the world's political systems have experienced one form or another of internal warfare leading to violent collapse of the governments in power. Certain crisis situations seem to increase the likelihood of this kind of breakdown. Wars and, more particularly, national military defeats have been decisive in prompting many revolutions. The Paris Commune of 1871, the Russian revolutions of 1905 and 1917, Hitler's overthrow of the Weimar Constitution in Germany, and the revolutions in China all occurred in the aftermath of national military disasters. Many factors in such a situation, including the cheapening of human life, the dislocation of population, the ready availability of arms, the disintegration of authority, the discrediting of the national leadership, material scarcities, and a sense of wounded national pride, contribute to the creation of an atmosphere in which radical political change and violent mass action are acceptable to large numbers of people. Economic crises are another common stimulus to revolutionary outbreaks, for they produce not only the obvious pressures of material scarcity and deprivation but also a threat to the individual's social position, a sense of insecurity and uncertainty as to the future, and an aggravation of the relationships among social classes. A severe national economic crisis works, in much the same way as a military disaster, to discredit the existent leadership and the present regime. Another triggering factor is the outbreak of revolutions in other political systems. Revolutions have a tendency to spread: the Spanish Revolution of 1820 had repercussions in Naples, Portugal, and Piedmont; the French Revolution of July 1830 provoked similar outbreaks in Poland and Belgium; the Russian Revolution of 1917 was followed by a dozen other revolutions; and the colonial liberation movements after World War II appear to have involved a similar chain reaction.

Crisis situations test the stability of political systems in extremely revealing ways, for they place extraordinary demands on the political leadership and the structure and processes of the system. Since the quality of the political leadership is often decisive, those systems that provide methods of selecting able leaders and replacing them possess important advantages. Although leadership ability is not guaranteed by any method of selection, it is more likely to be found where there is free competition for leadership positions. The availability of established methods of replacing leaders is equally, if not more, important, for the result of crises is often to disgrace the leaders in power, and, if they cannot be replaced easily, their continued incumbency may discredit the whole regime. The stamina and resolve of the ruling elite are also important. It is often said that a united elite, firmly believing in the justice of its own cause and determined to employ every measure to maintain its power, will not be overthrown. Most revolutions have gotten under way not when the oppression was greatest but only after the government had lost confidence in its own cause.

Other conditions of the survival of political systems relate to the effectiveness of the structures and processes of government in meeting the demands placed on them. Political systems suffer violent breakdown when channels of communication fail to function effectively, when institutional structures and processes fail to resolve conflicts among demands and to implement acceptable policies, and when the system ceases to be viewed as responsive by the individual and groups making demands on it. Usually, a system has failed over a period of some time to satisfy persistent and widespread demands; then, exposed to the additional strains of a crisis situation, it is unable to maintain itself. Revolutions and other forms of violent collapse are thus rarely sudden catastrophes but rather the result of a process of considerable duration that comes to its climax when the system is most vulnerable.

Causes of the breakdown of political systems

Unstable political systems are those that prove vulnerable to crisis pressures and that break down into various forms of internal warfare. The fundamental causes of such failures appear to be the lack of a widespread sense of the legitimacy of state authority and the absence of some general agreement on appropriate forms of political action. Governments suffer their gravest handicap when they must govern without consent or when the legitimacy of the regime is widely questioned. This is often the case in systems that have experienced prolonged civil war, that are torn by tensions among different national or ethnic

groups, or in which there are divisions along sharply drawn ideological or class lines. The problem is often most acute where there is a pretender to the throne, a government in exile, a neighbouring state sympathetic to a rebel cause, or some other focus for the loyalty of dissidents. To some degree, also, the problem of legitimacy confronts all newly established regimes. Many of the new states of Africa and Asia, for example, have found it a source of great difficulty. Often they have emulated the form of Western institutions but failed to achieve their spirit: borrowing eclectically from Western political philosophies and systems of law, they have created constitutional frameworks and institutional structures that lack meaning to their citizens and that fail to generate loyalty or a sense that government exercises rightful powers.

Closely related to the problem of legitimacy as a cause of the breakdown of political systems is the absence of a fundamental consensus on what is appropriate political behaviour. A regime is fortunate if there are well-established, open channels of political action and settled procedures for resolving grievances. Although the importance of such "rules of the game" is that they allow change to occur in mainly peaceful ways, stable political systems often show surprising tolerance for potentially violent forms of political behaviour, such as strikes, boycotts, and mass demonstrations. Such forms of political behaviour are not permitted in systems where there are no agreed limits to the role of violence and where there is a high risk that violence may escalate to the point of actual warfare. If the government cannot count upon widespread support for peaceful political procedures, it must restrict many kinds of political action. Such restriction, of course, inhibits still more the development of open methods of citizen participation in politics and adds to tension between the government and the people.

Managed and evolutionary systems

Stable political systems. The simplest definition of a stable political system is one that survives through crises without internal warfare. Several types of political systems have done so, including despotic monarchies, militarist regimes, and other authoritarian and totalitarian systems. After 1868, in the period of the restoration regime under the Meiji emperor, Japan succeeded, without major political breakdowns, in building an industrial state and developing commercial structures that transformed traditional Japanese society. This achievement was based on the development of centralized patterns of political control and the growth of a type of authoritarianism involving the rule of a military elite. Similarly, some of the totalitarian regimes of the contemporary world have demonstrated an impressive capability for survival. The key to their success is their ability to control social development, to manage and prevent change, and to bring under governmental direction all the forces that may result in innovations that are threatening to the system.

In some systems, survival does not depend on the detailed management of the society or close governmental control over social processes but is the result of sensitive political response to the forces of change, of flexible adjustment of the structures of the system to meet the pressures of innovation, and of open political processes that allow gradual and orderly development. Much of the Western democratic world has achieved peaceful progress in this way, despite new political philosophies, population increases, industrial and technological innovations, and many other social and economic stresses.

Such evolutionary change is possible when representative institutions provide effective channels for the communication of demands and criticisms to governments that rely upon majority support. The election of legislators and executive officials, competition between political parties, constitutional guarantees of freedom of speech and press, the right of petition, and many other structures and procedures perform this function in contemporary constitutional democracies. In such systems, social and economic problems are quickly transformed into issues in the open arenas of politics; governments are obliged to shape policies that reflect a variety of pressures and effect compromises among many conflicting demands.

The representative mechanisms that have produced evolutionary change in Western constitutional democracies are themselves subject to a continuous process of adjustment and mutation. Indeed, representative institutions must develop in ways that reflect social and economic developments in the society or they will lose their legitimacy in the minds of the people. In political systems such as the United States, for example, subtle shifts in the function and relative power of different institutions are continuously being made and, over time, produce entirely new structures and very different patterns of institutional behaviour. It is as a result of this process that the presidency has accumulated a range of new powers that have given it primacy among the branches of American government. This process also explains the growth of administrative agencies that perform both legislative and judicial functions. This process of dynamic adjustment is crucial, for institutions that remain static in a changing society are unable to serve as agencies of evolutionary change.

THE PROBLEM OF POLITICAL CHANGE

Four types of political change

The study of political change is difficult, for change occurs in many different ways and at many different points in the political system. One may distinguish several major types of change. First are changes of the most fundamental type—transformations not only of the structure of government but of the whole polity. Such change is not limited to political life but transforms also the social order, the moral basis, and the values of the whole society. Drastic change of this kind occurred in the four great revolutions of the modern era—the English Revolution of the 17th century, the American Revolution, the French Revolution, and the Russian Revolution. These movements had the most profound effect on social and political life, permanently altering the beliefs by which men live. Their consequences were felt not only in the societies in which they occurred but also in many other political systems, in which, as a result of their example, revolutions of an equally fundamental character occurred.

Each of these major revolutions was something of a world revolution, for it resulted in a basic change in the ways in which men in all political systems viewed the nature of politics and the purpose of political life. The independence movements in the colonial empires after World War II, for example, were fuelled by those principles of individual liberty and representative government that were once the slogans of 18th-century American and French revolutionaries. Marxist revolutionary concepts emphasizing economic progress and radical social change have shaped the development of many of the new nations. The continuing impact of such ideas is an example of another way in which fundamental political change occurs. The nature of a political system may be transformed not suddenly or violently in the course of revolution but by the gradual, corrosive influence of ideas and by the accumulating impact of different political philosophies.

A second type of change involves alterations to the structure of the political system. Such change is not fundamental, in the sense of a basic transformation of the nature of the regime, but it may produce great shifts in policy and other political outcomes. Because the structure of a political system—that is, its formal and informal institutional arrangements—is a major determinant of policy outcomes, it is frequently the target of political action of various kinds. The political activist, the reformer, and the revolutionary share the recognition that the policies of a government may be effectively changed by adjusting the institutional forms through which the government acts. In some systems, structural change has been accomplished by legal means. In the United States, for example, such major institutional reforms as the direct election of the Senate and the limitation on presidential terms were made by constitutional amendment; and in Britain the various reforms of Parliament were accomplished by statute. In other systems, structural changes are often achieved by revolution and other violence.

A third type of political change involves the replacement of leaders. Again, the recognition that to change the personnel of a government may be an effective way of

changing government policy prompts many kinds of political action, ranging from election contests to political assassination and various forms of coup d'etat. In some systems the existence of established means of changing political leaderships works to prevent violent types of political action. In the United States the quadrennial contests for the presidency afford a constitutional opportunity to throw the whole executive leadership out of office. At the other extreme, the coup d'etat leads to the abrupt, often violent replacement of national executives. Although it is a type of revolution, the coup d'etat usually does not involve prolonged struggle or popular participation; after seizing office, the principal aim of the leaders of the coup is usually the restoration of public order. The coup d'etat occasionally develops into much more than the replacement of one set of governmental leaders by another and may prove to be the initial stage of a truly revolutionary process; *e.g.*, the coups d'etat that initiated Communist rule in Czechoslovakia in 1948 and ended King Farouk I's regime in Egypt in 1952.

Government policy itself may be an important agency of political change. The social and economic policies of Franklin D. Roosevelt's New Deal and the Socialist programs implemented by the British Labour Party after 1945 are examples. In both cases, government policies resulted in far-reaching modifications to the functioning of the political system: a vast expansion in the role of government in the economy, the use of taxes to redistribute wealth, an increase in the political influence of organized labour, and the implementation of national programs of social welfare. Major policy change of this type, of course, is often a response to widespread pressures and demands that, if not satisfied by the system, may intensify and lead to various forms of violent political action. At other times, however, policy changes are imposed by a government to achieve the political, social, and economic goals of a single class, of an elite, or of the political leadership itself.

Many important questions remain as to the reasons for change, the ways in which change occurs, and the effects of change. Political scientists are still not completely certain, for example, why some systems have managed to avoid violent political change for considerable periods, while in other systems change is typically accomplished through coups d'etat, revolutions, and other forms of internal warfare. As suggested above, the explanation may have much to do with the existence in countries such as the United States and Great Britain of well-established political institutions that permit peaceful change, the presence in the population of widely shared attitudes toward the government, and the existence of basic agreement on the legitimacy of state authority. Clearly, however, other factors are also involved. Perhaps one of the chief goals of the study of political systems should be to determine as exactly as possible the conditions and prerequisites of those forms of change that permit the peaceful and evolutionary development of human society.

V. The study of political systems

CONTEMPORARY POLITICAL SCIENCE

The academic discipline of political science underwent a profound transformation in the four decades after 1930. The study of politics, particularly in the United States, gradually developed into a behavioral science. First, the social sciences generally experienced a technological revolution—a radical change in the equipment, techniques, and procedures used in research. Second, there was also a revolution in the philosophical posture and theoretical orientation of students of politics: interest gradually shifted from the traditional concerns of speculative philosophy to the theoretical problems of empirical analysis; increasingly, political scientists became concerned with "systematic theory," turning for guidance to the work of logicians and philosophers of science rather than to the classics of political theory. A third development was the rise of prestigious new social-science disciplines. Political scientists displayed mounting enthusiasm for interdisciplinary cooperation and borrowed increasingly from the theories, methods, conceptual schemes, and research results of other behavioral sciences. The effect of these three developments was to bring into question many of the approaches traditionally used in political inquiry.

New techniques and methods of analysis. The revolution in social-science technology has placed many novel research tools at the disposal of political scientists. Today many scholars use the interview techniques of the mass survey and the "panel" method; metric techniques, such as the scalogram, scale analysis, and factor analysis are frequently employed; and increasing interest has been shown in the use of specialized computer routines for purposes such as "pattern searches," "multiple regression analysis," and "simulation." Some of these technical innovations have already had a dramatic impact in more than one field of political science. The probability sample survey, for example, although it began with voting and public-opinion studies, has been extended to the investigation of specialized elites and institutionalized groups such as legislators and bureaucratic organizations. Similarly, metric techniques such as scalogram and factor analysis have been applied not only to voting and public-opinion research but also to the treatment of legislative roll-call votes and judicial decisions. Observational and sociometric techniques have been increasingly used in several areas for the study of small and middle-sized groups and for the analysis of group properties such as authority, communication, morale, cohesion, and consensus. Other techniques have been shown to have potential usefulness. For example, "experimental" or "prototype" studies and techniques—planned and deliberate "interferences" in real-life situations—may eventually be brought to bear upon the study of decision-making processes. Similarly, computer simulation—already a fruitful device when applied to problems of strategy in conjunction with game-theoretic analysis—may prove to be useful. The effect has been a revolutionary expansion in the data resources available to political scientists.

The empirical approach

New theoretical approaches. Increasingly, political scientists are committed to the belief that theory and research are mutually interrelated activities. Theoretical questions, it is said, must be stated in operational terms —in the form of testable hypotheses—for purposes of empirical research. And, in turn, empirical findings must be brought to bear upon theoretical formulations. Some students seek to develop general theoretical frameworks designed merely to orient the observer and to serve as "research guides" or "heuristic devices." At another extreme, there are attempts to build logically consistent, deductive models of the political process and to utilize theory as a tool of explanation and prediction.

The ultimate goal of much modern political inquiry is explanation, not description. A majority of contemporary political scientists are no longer content to describe the universe of politics. They seek, instead, to explain political behaviour—why people behave politically as they do and why, as a result, political processes and systems function as they do. With this emphasis, there is a new interest in the logic of scientific explanation. Much self-conscious attention is now paid to the problems of logical inference and to the use of different kinds of proof or techniques of validation in empirical analysis.

The interdisciplinary orientation. A leading development in modern political science is the willingness of students to borrow from the concepts, theories, and data of other social sciences. For example, the concept of role has been taken from psychology and sociology to construct a "role-theory model of the legislative actor." Similarly, structural–functional analysis has been borrowed from anthropology and sociology by students of both comparative politics and public administration. Systems analysis, which has many disciples in other behavioral sciences, has been utilized in fields of political inquiry.

Borrowings from other fields

Recent developments in the study of political leadership have owed a great deal to the pioneering efforts of sociologists and psychologists. In their studies of elites and social mobility, sociologists developed techniques that political scientists could use, with little modification, to determine political mobility and the representativeness of political elites. The results of primary- and middle-size-

group research in psychology and sociology had an obvious interest to students of cabinets, legislative and executive committees, administrative bodies, and executive staff organizations. The conception of the relational nature of leadership was a valuable corrective for political scientists who had vainly sought to isolate stable characteristics and personality traits in political elites. The application of sociometric techniques that could elicit interindividual choices among group members was important in identifying informal leadership structures in middle-sized political groups. And the distinction between formal and effective authority, a result of early sociological studies, found rapid application by students of politics.

Criticisms of traditional approaches. The older political science was characterized by the use of historical, legal, and prescriptive approaches that have increasingly come to be questioned.

Criticisms of the historical approach. Many modern political scientists do not consider history to be a science. The logic of historical explanation, the use of theory in historical research, and the procedures, sources, and goals of historical analysis have all been criticized. To begin with, historical explanations of events are viewed dubiously. The mere statement of a historical sequence, it is said, does not amount to causal explanation. The historian may describe the genesis of a particular event, but his premises do not suffice to establish the event deductively. Historians' attitudes to theory are also suspect. On the one hand, philosophers have united to discredit historicism—the construction of grand, overarching theories of the first causes and final ends of history viewed as a deterministic process. On the other hand, social scientists have condemned the specialized, "scientific" history in which concern with detailed facts is at the expense of hypotheses, generalizations, and theory of all kinds. On a less abstract level, the historian's procedures have also been criticized. Historical analysis, it is claimed, consists merely of an impressionistic linking of variables and an intuitive treatment of complex phenomena. Historical sources are said to be inadequate, since historians rely upon materials that come to them by the haphazard process of documentary survival. And, finally, the historian's chief goals of realistic description and accurate reconstruction of historical events are also questioned. Mere description, it is said, is not the aim of science.

The modern tendency among social scientists, therefore, is to deny that history has an autonomous status as a social science and to insist that it is a mere auxiliary, a reservoir of data, for other disciplines. Among political scientists, modern behavioralist writers have most sharply expressed their discontent with historical-descriptive approaches to political subjects. They believe that, where historians are principally concerned with the uniqueness of the phenomena investigated, the political scientist should be interested in the search for hypotheses, generalizations, and even universal laws. In nearly all fields of political science, moreover, the goal of realistic description has been replaced by the effort to evolve analytical schemes that will relate one event to another causally rather than chronologically and systematically rather than impressionistically.

Criticisms of legal approaches. Legal approaches to political subjects have also met with considerable criticism. Even among early political scientists there were some who contrasted "real" with "legal" forms and functions of government and who preferred realistic to legal methods of analysis. But it was the development of new areas of inquiry—political parties, pressure groups, and public opinion—that finally brought legalism in political science into disrepute. Students had now to deal with a new and broader universe of politics, in which extralegal groups and processes were a primary focus of attention. Legal sources—constitutional clauses, statutory provisions, judicial dicta, and the like—had little relevance in these new areas. Increasingly, political scientists contrasted formal, legal materials with the data thrown up by the new methods of empirical analysis. A belief gradually developed that the perspectives of constitutional law were false and artificial and that only the distinctively political approaches that had been developed in the new areas of inquiry could capture the realities of the political process. Today, a diminishing minority of political scientists continue to employ legal approaches.

Even the traditional subjects of the public law syllabus have been affected by the new, nonlegal approaches. The court system, for example, is now represented as an arena for the play of political pressures; and the traditional simplistic model of a neutral judiciary whose only function was to interpret or apply laws made by the representative body has been replaced with a view of the courts as an integral part of the political system and of judges as policy makers. More recently, legal studies have been further revolutionized by the application of methods of modern social science to the analysis of judicial opinions.

Criticisms of prescriptive approaches. Political scientists' attitudes to the use of prescriptive and reform-oriented approaches also have changed. One position, modelled on the procedures and aims of the physical sciences, holds that the student of political subjects should not evaluate, criticize, or advocate the reform of the phenomena with which he deals. A less rigorous position concludes that, although prescriptive or meliorative activities and the attempt to advise policy makers should have their place in the discipline, the chief purpose of inquiry should not be policy advice but the pursuit of knowledge. A third position insists that, though political scientists can never choose the goals of politics, their skills can be used in problem solving, and political science as a policy science may have a legitimate function in clarifying (without evaluating) policy alternatives and estimating (without judging) their consequences. All these positions are a considerable distance from the crusading, curative, and reformist attitudes of earlier political scientists and from the ethical inquiries into the nature of the good state once typical of political theory.

Even among those political scientists who continue to hold that it is proper and laudable to search for means to improve the working of the political system, there is evident a new caution and hesitancy. Reform is understood to be a very difficult and complex undertaking, while proposals for wholesale change, involving major constitutional revision, are considered unrealistic. Adjustments to governmental mechanisms, it is agreed, must match existent developments. Governmental institutions, moreover, are no longer regarded as separate and isolable from the remainder of the government and the society; they are viewed as components of a system, affecting and affected by needs and changes in the society and the government as a whole. It is recognized that the features of a governmental system that one may wish to change are often the product of deeper social and political forces that are irreversible and beyond human control. It is likewise recognized that the attempt to modify some apparently minor but awkward piece of institutional machinery may lead to major, unforeseen, and unwanted changes in other parts of the system. The effort to transfer bits and pieces of one political system to another is seen to be impossible for the very reason that it is right to use the word system. The stress on formal and structural changes to accomplish reorganization or institutional reform has been largely replaced with a recognition of the importance of the fluid, dynamic, and behavioral forces that affect institutional decision making. And, finally, the development of the modern, more complex view of institutions as patterns of anticipated behaviour has been accompanied by the growth of an understanding that institutional change does not come easily and that there is great toughness and resilience in the established forms of behaviour. All these insights have worked to sharpen an awareness of the limitations of political analysis and of the difficulties and dangers of governmental reform.

Weaknesses of contemporary political science

Although these innovations have been accompanied by a vast expansion in the number of professional political scientists and in the financial resources available for research, there are many troubling features of the contemporary discipline. In the view of many critics, the promise of a more realistic political science is still unfulfilled. In many cases, the new theoretical approaches have yielded

not the promised additions to our knowledge but sterile controversies among a new generation of schoolmen. In many cases, also, the new techniques and methods of analysis have led to investigation of trivial questions or to an obsessive concern with methodological problems for their own sake. The interdisciplinary orientation, too, has often resulted in a dependence on the conceptual apparatus of sociology, psychology, and other modern social science disciplines that is at least as dangerous as the older emphasis on legal and historical approaches. Finally, the reaction to prescription has made the discipline vulnerable to charges that it is incapable of confronting the important problems of the modern world.

FUTURE APPROACHES TO THE STUDY OF POLITICAL SYSTEMS

The study of advanced systems

Although it is hazardous to speculate on the future development of political science as a discipline, it seems clear that one must not expect any general reaction to the major trends of recent decades. All the main features of the behavioral revolution are now deeply entrenched in the discipline, and they will certainly continue to shape most of the writing and research on political systems for a long time to come. In a few areas, however, one may hope for adjustments that will correct some of the obvious weaknesses of recent scholarship. Much of the literature on comparative politics of the last four decades has dealt with non-Western political systems. This emphasis yielded much information on the process of political development and change, and many of the new concepts that are now employed to explain the role of social forces in shaping political life are the product of work on the new nations of Africa and Asia.

This progress in the understanding of non-Western systems has no parallel in the study of the political institutions of advanced systems. Scholarly work on the politics of western European countries is still largely dominated by legal and historical approaches. Research on the nonlegal political institutions and processes—parties, interest groups, public opinion, leadership recruitment, and the like—of many European constitutional democracies is still in its infancy. The advanced political systems of the Communist world have also been given much less attention than they merit.

One of the trends of the future may be a renewal of interest in the politics of advanced political systems. Many of the tools of inquiry and the theoretical concepts that are needed for such an enterprise have been made available in studies of the politics of the United States and Britain. The developing role of the European Common Market, the re-emergence of Germany and Japan as great powers, and the changes that are occurring in the relationships among the Communist states of eastern Europe may provide a new focus for research.

The limits of analysis

The contemporary behavioral political scientist is no longer able to confine his research to formal political institutions. He must treat the whole social environment of the political system. Indeed, every social relationship that involves the exercise of authority or the allocation of social values has been brought within his professional purview. Thus, the field of comparative politics has become unmanageably broad and almost hopelessly complex.

One way of lightening the burden of the political scientist might be to break down the field of comparative politics into more manageable segments of specialization. Another, perhaps more fruitful development might be to redefine the limits of analysis. This would involve not returning to the narrow legalism of earlier political science but placing more emphasis on the political system itself and on the structures and processes of politics.

A recent tendency among political scientists, particularly those who have sought to escape from the normative emphasis of traditional political science, has been to avoid dealing with issues of public policy. They study the processes by which policies are made rather than the policies themselves. Research on other political systems is even less concerned with policy.

Widespread controversy in the United States over national policy issues has made the subject matter of much "scientific" research on politics appear trivial and irrelevant. Some American political scientists, reacting to what they perceive as the dehumanizing tendency of behavioral research, have begun to insist that policy evaluation should be a principal concern of the discipline. If this movement gathers impetus, it may have important consequences for research on other political systems. Techniques that have proved useful in studying the processes of public decision making in the United States might be applied with great profit to other political systems. Eventually, there may develop a school of comparative politics that focusses on policy processes and that is concerned with comparing and evaluating the policy outputs of different political systems.

Perhaps the future will ultimately bring many more fundamental changes in the study of political systems. More realistic ways of comparing political systems may be discovered or rediscovered, and a discipline may emerge that is appropriate to modern conditions. Such a discipline would be sensitive to differences in the character of regimes, attuned to revealing contrasts in policy, and concerned with the protection of freedom.

BIBLIOGRAPHY

General works: Perhaps the best single introduction to the subjects covered in this article is CARL J. FRIEDRICH, *Man and His Government: An Empirical Theory of Politics* (1963). Other excellent general works include KARL LOEWENSTEIN, *Political Power and the Governmental Process*, 2nd ed. (1965); and R.M. MACIVER, *The Modern State* (1926) and *The Web of Government* (1947). A good introductory text is JOHN C. WAHLKE and ALEX N. DRAGNICH (eds.), *Government and Politics* (1966); and some carefully selected readings are available in HARRY ECKSTEIN and DAVID E. APTER (eds.), *Comparative Politics* (1963). For many subjects, *The International Encyclopedia of the Social Sciences*, 17 vol. (1968), is an excellent reference.

Supranational political systems: The history of the empires of the ancient world may be studied in a number of monumental works, including ARNOLD J. TOYNBEE, *A Study of History*, 1st–2nd ed., 12 vol. (1948–61); OSWALD SPENGLER, *Der Untergang des Abendländes*, 2 vol. (1918–23; Eng. trans., *The Decline of the West*, 2 vol., 1926–28); and EDUARD MEYER, *Geschichte des Altertums*, 5 vol. (1884–1902, reprinted 1910–39). Among other general works, S.N. EISENSTADT, *The Political Systems of Empires* (1963), compares the ancient empires in terms of bureaucracy; and KARL WITTFOGEL, *Oriental Despotism: A Comparative Study of Total Power* (1957), traces the origins of their political orders. Treatments of the major empires of the ancient world include JAMES H. BREASTED, *Development of Religion and Thought in Ancient Egypt* (1912); A.T. OLMSTEAD, *History of Assyria* (1923, reprinted 1960); JOHN A. WILSON, *The Culture of Ancient Egypt* (1956); WILLIAM C. HAYES, *The Scepter of Egypt* (1953); R.E.M. WHEELER, *The Indus Civilization* (1953); OTTO FRANKE, *Geschichte des chinesischen Reiches*, 5 vol. (1930–52); WOLFRAM EBERHARD, *Chinas Geschichte* (1948; Eng. trans., *A History of China*, 3rd ed. rev., 1969); ETIENNE BALAZS, *Chinese Civilization and Bureaucracy* (1964); MIKHAIL ROSTOVTSEV, *The Social and Economic History of the Roman Empire* (1963); A.E.R. BOAK, *History of Rome to 565 A.D.* (1921); GEORGE OSTROGORSKI, *Geschichte des byzantinischen Staates* (1965; Eng. trans., *History of the Byzantine State*, 2nd ed., 1968). Histories of modern colonial empires are found in J. HOLLAND ROSE *et al.* (eds.), *The Cambridge History of the British Empire*, 9 vol. (1929–59; 2nd ed., 1963–); RICHARD KOEBNER and HELMUT D. SCHMIDT, *Imperialism: The Story and Significance of a Political World, 1840–1960* (1964); CHARLES DE LANNOY and HERMAN V. LINDEN, *Histoire de l'expansion coloniale des peuples européens*, 3 vol. (1907–21); STEPHEN H. ROBERTS, *The History of French Colonial Policy, 1870–1925* (1963); ROBERT L. SCHUYLER, *The Fall of the Old Colonial System* (1945); MARY E. TOWNSEND, *The Rise and Fall of Germany's Colonial Empire, 1884–1918* (1930); CLARENCE H. HARING, *The Spanish Empire in America* (1947). A discussion of the post-imperial era in Asia and Africa is provided by RUPERT EMERSON in *From Empire to Nation* (1960). An excellent discussion of modern supranational political institutions is available in INIS L. CLAUDE, *Swords into Ploughshares*, 4th ed. (1971); the same author discusses the conditions for the development of different types of international organization in his *Power and International Relations* (1962). GERARD J. MANGONE, *A Short History of International Organization* (1954), is a convenient reference for the development of most contemporary supranational political institutions. Standard treatments of the League of Nations and the United Nations

are available in FRANCIS P. WALTERS, *A History of the League of Nations* (1960); ALFRED E. ZIMMERN, *The League of Nations and the Rule of Law, 1918–1935* (1938); and HERBERT G. NICHOLAS, *The United Nations As a Political Institution*, 2nd ed. (1962). More specialized works that may be consulted include THOMAS HOVET, *Bloc Politics in the United Nations* (1958); STEPHEN M. SCHWEBEL, *The Secretary-General of the United Nations* (1952); and ALEXANDER DALLIN, *The Soviet Union at the United Nations* (1962). The development of supranational institutions in Europe is discussed in ERNST B. HAAS, *The Uniting of Europe* (1958); MICHAEL CURTIS, *Western European Integration* (1965); WALTER HALLSTEIN, *United Europe: Challenge and Opportunity* (1962); and UWE W. KITZINGER, *The Politics and Economics of European Integration* (1963). Good discussions of the political and constitutional organization of the British Commonwealth are found in WILLIAM S. LIVINGSTON (ed.), *Federalism in the Commonwealth* (1963); J.D.B. MILLER, *The Commonwealth in the World* (1958); and K.C. WHEARE, *The Constitutional Structure of the Commonwealth* (1960). An extremely useful commentary on various forms of supranational political organization is provided in IVO D. DUCHACEK, *Comparative Federalism: The Territorial Dimension of Politics* (1970). A journal published since 1947 by the World Peace Foundation, *International Organization* (quarterly), contains informative articles on current developments in international organization and also provides selective bibliographies.

National political systems: The phenomenon of nationalism is treated in HANS KOHN, *The Idea of Nationalism* (1961); and EUGEN LEMBERG, *Nationalismus*, 2 vol. (1964); and a historical sketch of the emergence of the idea and its reality is provided in CARLTON J.H. HAYES, *The Historical Evolution of Modern Nationalism* (1931, reprinted 1968). Among the many works on federal and unitary systems that the student should consult are ARTHUR W. MACMAHON (ed.), *Federalism, Mature and Emergent* (1955); WILLIAM S. LIVINGSTON, *Federalism and Constitutional Change* (1956); VALERIE EARLE (ed.), *Federalism* (1968); CARL J. FRIEDRICH, *Trends of Federalism in Theory and Practice* (1968); PETER HAY, *Federalism, and Supranational Organizations* (1966); K.C. WHEARE, *Federal Government*, 4th ed. (1963); and GEORGE C.S. BENSON *et al.*, *Essays in Federalism* (1961). IVO D. DUCHACEK (*op. cit.*) should also be consulted. Works on individual nation-states are so numerous that the student may wish to consult some standard texts that include selected bibliographies: HAROLD ZINK, *Modern Governments*, 2nd ed. (1962); HERMAN FINER, *The Major Governments of Modern Europe* (1960); and GWENDOLEN M. CARTER and JOHN H. HERZ, *Major Foreign Powers*, 5th ed. (1967). Comparative studies of different aspects of contemporary national government are also numerous. The bureaucracies and administrative procedures of different modern governments are treated in BRIAN CHAPMAN, *The Profession of Government: The Public Service in Europe* (1959); JOSEPH G. LA PALOMBARA (ed.), *Bureaucracy and Political Development* (1963); FRITZ MORSTEIN MARX, *The Administrative State* (1957); WILLIAM J. SIFFIN (ed.), *Toward the Comparative Study of Public Administration* (1957); ERICH STRAUSS, *The Ruling Servants: Bureaucracy in Russia, France and—Britain?* (1960); and J.D. THOMPSON *et al.* (eds.), *Comparative Studies in Administration* (1959). Constitutions, courts, and other judicial institutions are comparatively treated in LESLIE WOLF-PHILLIPS (ed.), *Constitutions of Modern States* (1968); K.C. WHEARE, *Modern Constitutions*, 2nd ed. (1966); WILLIAM GEORGE ANDREWS (ed.), *Constitutions and Constitutionalism*, 3rd ed. (1968); HENRY J. ABRAHAM, *Courts and Judges* (1959) and *The Judicial Process: An Introductory Analysis of the Courts of the United States, England, and France*, 2nd ed. rev. (1968); FREDA CASTBERG, *Freedom of Speech in the West: A Comparative Study of Public Law in France, the United States, and Germany* (1960); RUSSELL F. MOORE (ed.), *Modern Constitutions* (1957); DOROTHY PEASLEE XYDIS, *Constitutions of Nations*, rev. 3rd ed., 4 vol. (1965–70); and CHARLES F. STRONG, *A History of Modern Political Constitutions* (1964). National legislatures are discussed in LORD CAMPION and D.W.S. LIDDERDALE, *European Parliamentary Procedure* (1953); PAUL EINZIG, *The Control of the Purse* (1959); JOHN E. KERSELL, *Parliamentary Supervision of Delegated Legislation* (1960); and K.C. WHEARE, *Legislatures* (1963). Political parties, elections, and interest groups are comparatively treated in ROBERT R. ALFORD, *Party and Society: The Anglo-American Democracies* (1963); DAVID E. BUTLER (ed.), *Elections Abroad* (1959); ROBERT A. DAHL (ed.), *Political Oppositions in Western Democracies* (1966); MAURICE DUVERGER, *Les Partis politiques*, 5th ed. (1964; Eng. trans., *Political Parties*, 2nd ed. rev., 1959); HENRY W. EHRMANN (ed.), *Interest Groups on Four Continents* (1958); JOSEPH G. LA PALOMBARA and MYRON WEINER (eds.), *Political Parties and Political Development* (1966); W.J.M. MACKENZIE, *Free Elections* (1958); SIGMUND NEUMANN (ed.), *Modern Political Parties* (1956); and T.E. SMITH, *Elections in Developing Countries* (1960). General comparative studies of national political systems include GABRIEL ALMOND and SIDNEY VERBA, *The Civic Culture: Political Attitudes and Democracy in Five Nations* (1965); GABRIEL ALMOND and JAMES S. COLEMAN (eds.), *The Politics of the Developing Areas* (1960); GABRIEL ALMOND and G. BINGHAM POWELL, JR., *Comparative Politics: A Developmental Approach* (1966); ZBIGNIEW BRZEZINSKI and SAMUEL P. HUNTINGTON, *Political Power: USA/USSR* (1964); JAMES S. COLEMAN (ed.), *Education and Political Development* (1965); ROBERT C. FRIED, *Comparative Political Institutions* (1966); LUCIAN W. PYE (ed.), *Communications and Political Development* (1963); and LUCIAN W. PYE and SIDNEY VERBA (eds.), *Political Culture and Political Development* (1965).

Subnational political systems: Excellent general treatments of local and city government are found in WILLIAM A. ROBSON, *The Development of Local Government*, 3rd rev. ed. (1954); and (ed.), *Great Cities of the World: Their Government, Politics, and Planning*, 2nd ed. rev. (1957); and in LEWIS MUMFORD, *The City in History* (1961). On early forms of city government, the student should consult HENRI PIRENNE, *Les Villes du moyen âge* (1927; Eng. trans., *Medieval Cities*, 1956); and FUSTEL DE COULANGES, *La Cité antique* (1864; Eng. trans., *The Ancient City*, 1873, many reprints). A useful comparative study of local government is found in G.M. HARRIS, *Comparative Local Government* (1949). City and local governments in contemporary Western systems are treated in D.N. CHESTER, *Central and Local Government* (1951); HERMAN FINER, *English Local Government*, 4th ed. rev. (1950); and BRIAN CHAPMAN, *Introduction to French Local Government* (1953). Political and sociological analyses of the local community include ROBERT A. DAHL, *Who Governs?* (1961); DONALD J. BOGUE, *The Structure of the Metropolitan Community* (1949); FLOYD HUNTER, *Community Power Structure* (1953); ROBERT S. and HELEN M. LYND, *Middletown: A Study in Contemporary American Culture* (1929); and H.D. MALAVIYA, *Village Panchayats in India* (1966). For tribal forms of government, see J.A. NOON, *Law and Government of the Grand River Iroquois* (1949); and ISAAC SCHAPERA, *The Khoisan Peoples of South Africa* (1930) and *Government and Politics in Tribal Societies* (1956).

Types and models: The student of classical typologies of political systems will first refer to Book IV of Aristotle's *Politics* and then perhaps consult Book VIII of Plato's *Republic* together with a number of passages in *The Laws*. Leading modern efforts to develop types and models of political systems may be found in GAETANO MOSCA, *Elementi di scienza politica*, 2nd ed. (1923; Eng. trans., *The Ruling Class*, rev. ed., 1939); GUGLIELMO FERRERO, *Pouvoir* (1942; Eng. trans., *The Principles of Power*, 1942); and MAX WEBER, *Wirtschaft und Gesellschaft* (1922; Eng. trans. of pt. 1, *The Theory of Social and Economic Organization*, 1947). The Marxist typology, of course, is set forth in the writings of KARL MARX, FRIEDRICH ENGELS, LENIN, and their disciples. Contemporary efforts to develop typologies may be studied in DAVID E. APTER, "A Comparative Method for the Study of Politics," *American Journal of Sociology*, 64:221–237 (1958); DAVID EASTON, *A Systems Analysis of Political Life* (1965); GABRIEL ALMOND and G. BINGHAM POWELL, JR. (*op. cit.*); and ALEXANDER J. GROTH, *Comparative Politics: A Distributive Approach* (1971).

Development and change: There is a large contemporary literature centring on the question of modernization and political development. In addition to works by Gabriel Almond noted above, the student should consult PETER H. MERKL, *Modern Comparative Politics* (1970), which includes an excellent bibliography; and DAVID E. APTER, *The Politics of Modernization* (1965). Other works are CYRIL E. BLACK, *The Dynamics of Modernization* (1966) and (ed.), *The Transformation of Russian Society: Aspects of Social Change Since 1861* (1960); S.N. EISENSTADT, *Modernization: Protest and Change* (1966); JASON FINKLE and RICHARD W. GABLE (eds.), *Political Development and Social Change* (1966); EVERETT E. HAGEN, *On the Theory of Social Change* (1962); ROBERT T. HOLT and JOHN E. TURNER, *The Political Basis of Economic Development* (1966); SAMUEL P. HUNTINGTON, *Political Order in Changing Societies* (1968); JOHN H. KAUTSKY (ed.), *Political Change in Underdeveloped Countries* (1962); DANIEL LERNER, *The Passing of Traditional Society* (1958); MAX F. MILLIKAN and DONALD L.M. BLACKMER (eds.), *The Emerging Nations* (1961); DANKWART RUSTOW, *A World of Nations: Problems of Political Modernization* (1967); EDWARD A. SHILS, *Political Development in the New States* (1962); I. ROBERT SINAI, *The Challenge of Modernisation* (1964); FRED VON DER

MEHDEN, *Politics of the Developing Nations* (1964); IMMANUEL WALLERSTEIN, *Social Change: The Colonial Situation* (1966); BARBARA WARD, *The Rich Nations and the Poor Nations* (1961); and ARISTIDE R. ZOLBERG, *Creating Political Order: The Party-States of West Africa* (1966).

Forms, levels, and divisions of government: One of the best introductions to the study of different forms of government and an excellent review of the differences between constitutionalism and autocracy, democracy and totalitarianism, is KARL LOEWENSTEIN, *Political Power and the Governmental Process*, 2nd ed. (1965). Also excellent are CARL J. FRIEDRICH, *Constitutional Government and Politics* (1937); and with ZBIGNIEW BRZEZINSKI, *Totalitarian Dictatorship and Autocracy*, 2nd rev. ed. (1965). On totalitarianism, other books that should be consulted are ZBIGNIEW BRZEZINSKI, *The Soviet Bloc*, rev. ed. (1967), and *Ideology and Power in Soviet Politics*, rev. ed. (1967); MILOVAN DJILAS, *The New Class* (1957); MERLE FAINSOD, *How Russia Is Ruled*, rev. ed. (1963); ALEX INKELES and R.A. BAUER, *The Soviet Citizen: Daily Life in a Totalitarian Society* (1959); ERNST NOLTE, *Der Faschismus in seiner Epoche* (1963; Eng. trans., *The Three Faces of Fascism*, 1966); J.L. TALMON, *The Rise of Totalitarian Democracy* (1952); ALEXANDER DALLIN and ALAN F. WESTIN (eds.), *Politics in the Soviet Union* (1966); and HERBERT MCCLOSKY and JOHN E. TURNER, *The Soviet Dictatorship* (1960). On constitutional forms, the student should refer to WALTER BAGEHOT, *The English Constitution*, rev. ed. (1911, reprinted 1966); JAMES BRYCE, *The American Commonwealth*, rev. ed., 2 vol. (1931–33)); D.L. KEIR, *The Constitutional History of Modern Britain, 1485–1951*, 9th ed. (1969); IVOR JENNINGS, *Cabinet Government*, 3rd ed. (1969), and *Parliament*, 2nd ed. (1957); C.W. CASSINELLI, *The Politics of Freedom* (1961); V.O. KEY, *Politics, Parties, and Pressure Groups*, 5th ed. (1964), and *Public Opinion and American Democracy* (1961); A.D. LINDSAY, *The Modern Democratic State* (1943); C.H. MCILWAIN, *Constitutionalism, Ancient and Modern*, rev. ed. (1947); HERBERT MORRISON, *Government and Parliament*, 3rd ed. (1964); CLINTON ROSSITER, *Constitutional Dictatorship* (1948); and HERBERT J. SPIRO, *Government by Constitution: The Political Systems of Democracy* (1959). A good general introduction to the major European forms of constitutional and totalitarian governments is available in SAMUEL H. BEER and ADAM B. ULAM (eds.), *Patterns of Government: The Major Political Systems of Europe*, 2nd ed. rev. (1962). CARL J. FRIEDRICH offers an excellent introduction to the problem of the relationship between different levels of government in *Man and His Government* (1963)—see especially pt. 5, "Ranges and Levels of Government." Other books to consult on this subject are CLAUDE AKE, *A Theory of Political Integration* (1967); HAROLD ALDERFER, *Local Government in Developing Countries* (1964); REINHARD BENDIX, *Nation-Building and Citizenship* (1964); BRIAN CHAPMAN, *Introduction to French Local Government* (1953); L.G. COWAN, *Local Government in West Africa* (1958); RICHARD COX, *Pan-Africanism in Practice* (1964); KARL W. DEUTSCH, *Nationalism and Social Communication* (1953) and *The Nerves of Government* (1963; with new introd., 1966); RUPERT EMERSON, *From Empire to Nation* (1960); RICHARD R. FAGEN, *Politics and Communication* (1966); JAMES FESLER, *Area and Administration* (1949); HERMAN FINER, *English Local Government* (1933); J.A.G. GRIFFITH, *Central Departments and Local Authorities* (1966); LUCIAN W. PYE (ed.), *Communications and Political Development* (1963); WILLIAM A. ROBSON, *Local Government in Crisis* (1966); and MARC J. SWARTZ (ed.), *Local-Level Politics* (1968). Short introductions to the divisions of government, suitable for the beginning student, are available in ROBERT C. FRIED, *Comparative Political Institutions* (1966); and JOHN C. WAHLKE and ALEX N. DRAGNICH (*op. cit.*). On the legislative division of government, in addition to the books listed in an earlier section of this bibliography, the student should consult BERTRAM GROSS, *The Legislative Struggle* (1953); WILLIAM J. KEEFE and MORRIS S. OGUL, *The American Legislative Process*, 2nd ed. (1968); D.W.S. LIDDERDALE, *The Parliament of France* (1951); STEPHEN KING-HALL and R.K. ULLMANN, *German Parliaments* (1954); JOHN C. WAHLKE and HEINZ EULAU (eds.), *Legislative Behaviour* (1959); and JOHN C. WAHLKE *et al.*, *The Legislative System* (1962). Additional reading on executives should include C.G. BROWNE and T.S. COHN (eds.), *The Study of Leadership* (1958); EDWARD S. CORWIN, *The President: Office and Powers, 1787–1957*, 4th rev. ed. (1957); ALVIN W. GOULDNER (ed.), *Studies in Leadership* (1950); HOOVER INSTITUTE STUDIES, *Elites*, 5 vol. (1951–55); RICHARD E. NEUSTADT, *Presidential Power* (1960); and ERIC STRAUSS (*op. cit.*). Further reading on courts and systems of law is available in HAROLD J. BERMAN, *Justice in the U.S.S.R.* (1963); FELIX S. COHEN, *Ethical Systems and Legal Ideals* (1933); RENE DAVID and HENRY DE VRIES, *The French Legal System* (1957); JEROME FRANK, *Courts on Trial* (1949); RICHARD M. JACKSON, *The Machinery of Justice in England*, 3rd ed. (1960); HANS KELSEN, *The Communist Theory of Law* (1955); W.F. MURPHY and C.H. PRITCHETT, *Courts, Judges, and Politics* (1961); J.W. PELTASON, *Federal Courts in the Political Process* (1955); J.R. SCHMIDHAUSER, *The Supreme Court: Its Politics, Personalities, and Procedures* (1960); GLENDON SCHUBERT (ed.), *Judicial Decision-Making* (1963); and BERNARD SCHWARTZ, *French Administrative Law and the Common-Law World* (1954).

The functions of government: The functions of government may be studied in many works dealing with each of the world's advanced political systems. For the United States, the student should consult JAMES E. ANDERSON, *The Emergence of the Modern Regulatory State* (1962); MELVIN ANSHEN and FRANCIS D. WORMUTH, *Private Enterprise and Public Policy* (1954); MORROE BERGER, *Equality by Statute* (1952); RICHARD E. CAVES, *Air Transport and Its Regulators* (1962); ROBERT A. DAHL and CHARLES E. LINDBLOM, *Politics, Economics, and Welfare* (1953); KENNETH CULP DAVIS, *Administrative Law* (1951); ERNST FREUND, *Administrative Powers over Persons and Property* (1928); D.E. HATHAWAY, *Government and Agriculture: Public Policy in a Democratic Society* (1963); CARL KAYSEN and DONALD F. TURNER, *Antitrust Policy* (1959); JAMES M. LANDIS, *The Administrative Process* (1938); AVERY LEISERSON, *Administrative Regulation* (1942); JOHN H. LEEK, *Government and Labor in the United States* (1952); MICHAEL D. REAGAN, *The Managed Economy* (1963); EMMETTE S. REDFORD, *American Government and the Economy* (1965) and *Administration of National Economic Control* (1952); GEORGE A. STEINER, *Government's Role in Economic Life* (1953); GEORGE W. TAYLOR, *Government Regulation of Industrial Relations* (1948); and DONALD S. WATSON, *Economic Policy: Business and Government* (1960). For Great Britain, the student should consult PATRICK ABERCROMBIE, *Town and Country Planning*, 3rd ed. (1959); W.P. ALEXANDER, *Education in England: The National System—How It Works*, 2nd ed. (1964); MAX BELOFF, *New Dimensions in Foreign Policy* (1961); A.M. CARTTER, *The Redistribution of Income in Postwar Britain* (1955); HARRY ECKSTEIN, *The English Health Service* (1958); PETER GOLDMAN, *The Welfare State* (1964); F.A. JOHNSON, *Defense by Committee* (1960); T.H. MARSHALL, *Social Policy* (1965); and HAROLD E. RAYNES, *Social Security in Britain*, 2nd ed. (1960). On France there are WARREN C. BAUM, *The French Economy and the State* (1958); JOHN S. HARLOW, *French Economic Planning* (1966); MARIO EINAUDI, MAURICE BYE, and ERNESTO ROSSI, *Nationalization in France and Italy* (1955); W.R. FRASER, *Education and Society in Modern France* (1963); JOHN and ANNE-MARIE HACKETT, *Economic Planning in France* (1963); WALLACE C. PETERSON, *The Welfare State in France* (1960); and JOHN B. SHEAHAN, *Promotion and Control of Industry in Postwar France* (1963).

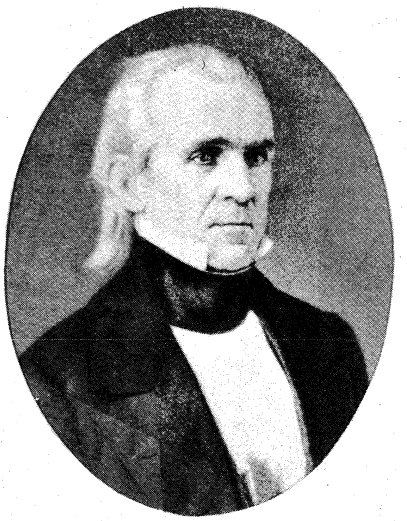
The changing of government: Further reading on the politics of change should include DAVID E. APTER (ed.), *Ideology and Discontent* (1964); ROBERT N. BELLAH, *Religion and Progress in Modern Asia* (1965); CRANE BRINTON, *The Anatomy of Revolution*, rev. ed. (1952); HARRY ECKSTEIN (ed.), *Internal War* (1964); SAMUEL E. FINER, *The Man on Horseback: The Role of the Military in Politics* (1962); CLIFFORD GEERTZ (ed.), *Old Societies and New States* (1963); MANFRED HALPERN, *The Politics of Social Change in the Middle East and North Africa* (1963); CHALMERS JOHNSON, *Revolutionary Change* (1966); JOSEPH LOPREATO, *Peasants No More* (1967); PAUL E. SIGMUND (ed.), *The Ideologies of the Developing Nations* (1962); KALMAN H. SILVERT (ed.), *Churches and States: The Religious Institution and Modernization* (1967); and ROBERT E. WARD and DANKWART RUSTOW (eds.), *Political Modernization in Japan and Turkey* (1964).

The analysis of political systems: An excellent brief review of contemporary trends in the discipline of political science is available in W.J. MACKENZIE, *Politics and Social Science* (1967). Other works that should be consulted include ROBERT R. BROWN, *Explanation in Social Science* (1963); MAURICE DUVERGER, *Méthodes des sciences sociales* (1959); Eng. trans., *An Introduction to the Social Sciences*, 1965); DAVID EASTON, *A Systems Analysis of Political Life* (1965); HEINZ EULAU, *The Behavioral Persuasion in Politics* (1963); HARRY V. JAFFA, *Equality and Liberty* (1965); ABRAHAM KAPLAN, *The Conduct of Inquiry* (1964); EUGENE J. MEEHAN, *The Theory and Method of Political Analysis* (1965); W.G. RUNCIMAN, *Social Science and Political Theory* (1963); EDWARD A. SHILS and HENRY A. FINCH (eds.), *Max Weber on the Methodology of the Social Sciences* (1949); FRANCIS J. SORAUF, *Political Science: An Informal Overview* (1965); and T.D. WELDON, *The Vocabulary of Politics* (1953).

(D.A.He.)

Polk, James K.

A friend of Pres. Andrew Jackson and a loyal proponent of his political principles, James Knox Polk, the 11th president of the United States, distinguished himself by the skill, equalled by few presidents, with which he won legislative approval for the large majority of his goals in foreign and domestic policy. The most important legacy of his presidency was the acquisition of the territories that today comprise the states of Texas, New Mexico, California, and Oregon.



Polk, daguerreotype by Mathew Brady, 1849.
By courtesy of the Library of Congress, Washington, D.C.

Polk was born in Mecklenburg County, North Carolina, on November 2, 1795. At the age of 11 he accompanied his family to Tennessee, where his father operated a prosperous farm in Maury County. Though ill health made formal schooling impossible during his childhood, at the age of 20 he successfully passed the entrance requirements for the second-year class of the University of North Carolina. He was "correct, punctual, and industrious," and as a graduating senior in 1818 he was the Latin salutatorian of his class—the pre-eminent scholar in both the classics and mathematics.

Early career

After graduation he returned to Tennessee and began to practice law in Nashville. His interest in politics, which had fascinated him even as a young boy, was encouraged by his association with leading public figures in the state. In 1820 he was admitted to the bar. Because he was a confirmed Democrat and an unfailing supporter of Andrew Jackson and because his style of political oratory became so popular that he was characterized as the "Napoleon of the stump," his political career was assured. His rapid rise to political power was furthered by his wife, Sara Childress Polk (1803–91), whom he married January 1, 1824, while serving in the state house of representatives (1823–25). The social prominence of her family and her personal charm were distinct assets for a politically ambitious lawyer. As an official hostess she won the admiration and esteem of the leading figures of the day, and for 25 years she was her husband's close companion in state and national politics.

James K. Polk was by nature a student of government, by experience a legislator, and by force of circumstance an administrator. He was not an easy man to know or to like. Even close companions did not relish his austerity, and associates tolerated but did not approve his inflexible living standards. Among his few close friends was Andrew Jackson, who encouraged and advanced Polk and whose influence carried him from the Tennessee House of Representatives to the United States House of Representatives, where he served from 1825 to 1839. As speaker of the House from 1835 to 1839, Polk acquired a reputation as an undeviating supporter of Jacksonian principles. In 1839 he left the House to become governor of Tennessee. Two defeats for a second term (1841, 1843) by small majorities convinced him that to strengthen his party he should return to Washington in some capacity.

Polk's nomination as the Democratic candidate for president in 1844 was unsought by him, for the party had more prominent sons in Martin Van Buren, Lewis Cass, and James Buchanan. But the Democrats could not reconcile their differences, and a compromise candidate had to be found because the campaign was to be run on issues and not on personalities; it was decided that Polk would do. He is regarded as the first "dark horse" nominee in the history of the presidency. As a party man from what was then the West, and a former member of the House of Representatives, it was thought that Polk would bring about legislative and executive cooperation and understanding in the functioning of the national government. While speaker of the House he had decided many procedural questions and had usually been sustained by majorities composed of the leaders of both parties. His party feeling was intense, but his integrity was unquestioned; he knew the rights and privileges of the House, and he also knew its responsibilities.

During his campaign Polk surprised the country by taking a positive stand on two burning issues of the day. Whereas other candidates hedged on the question of whether to annex Texas, which had been independent of Mexico since 1836, he demanded annexation. Whereas other candidates evaded the problem of joint occupancy of Oregon with England, he openly laid claim to the whole territory that extended as far north as 54°40′ with the campaign slogan "Fifty-four Forty or Fight." His election was close, but it was decisive—a popular plurality of about 38,000 and 170 electoral votes against 105 for Henry Clay.

Presidency

Not yet 50 years of age, Polk was the youngest successful presidential candidate up to that time. He entered the presidency full of vigour and with an expressed zeal to serve his country to the best of his ability. He left it four years later exhausted and enfeebled by his efforts. In office he demonstrated remarkable skill in the selection and control of his official advisors, and in his formal relations with Congress his legislative experience served him well. When his party was firmly united behind a policy he himself opposed, he yielded to the wishes of Congress. When he disagreed strongly with congressional policy and decided to make an issue of it, he fortified his position with recognized executive precedent and practice. His formal disapprovals (in the form of two message vetoes and one pocket veto by which legislation is killed by the failure of the president to sign a bill before the adjournment of Congress) were questioned, but the two returned measures failed to command the necessary two-thirds majority in order to override his vetoes.

The Polk administration was marked by large territorial gains. The annexation of Texas as a state was concluded and resulted in a two-year war with Mexico. As a consequence of that war the Southwest and far West (California), partly by conquest and partly by purchase, became part of the United States domain. During this period the northwest boundary became fixed by treaty, and the continental United States emerged a recognized reality. Additional achievements included a treaty with New Granada (Colombia) resolving the problem of right-of-way for U.S. citizens across the Isthmus of Panama; establishment of a warehouse system that provided for the temporary retention of undistributed imports; and the passage of the Walker Tariff Act of 1846, which lowered import duties and did much to pacify British public opinion that had been inflamed over the Oregon compromise of 1846 that established the international frontier at the 49th parallel. As these measures helped foreign trade, so the reenactment of the independent treasury system in 1846 helped in the solution of domestic financial problems.

The expansion of the country westward led to the creation of a new agency, the Department of the Interior. The Polk administration should also be credited with the establishment of the United States Naval Academy at Annapolis and the authorization of the Smithsonian Institution, a national foundation for all areas of science.

Polk's influence over his congresses may be gauged from

Assessment

the results of the recommendations of four annual messages and ten significant special messages to one or both houses. His control of legislative policy in bitterly partisan congresses must be judged in terms of results, not oratory or parliamentary delay. He recommended with a high degree of success settlement of a trade dispute with Great Britain; an increase in U.S. armed forces; war with Mexico, peace with Great Britain over Oregon, making available finances to expedite peace conclusions; organization of the Oregon Territory; peace with Mexico providing for limited conquest; and a revised treasury system. He refused information desired by Congress (on the ground that it was incompatible with the public interest), recognized a new French revolutionary government, and proclaimed the validity of the Monroe Doctrine. Succeeding presidents recognized these pronouncements.

A diary kept by Polk during his term of office stressed the presidential burden. Day after day, week after week, he recounted in his diary his experiences with the hosts of office seekers who infested Washington and who occupied so much of his public time. Again and again there is evident in his writings a note of despair. He knew from experience what an evil an unlimited executive patronage can become, but he felt powerless to change its obligations and too conscientious to avoid its duties. At the close of his term, March 4, 1849, Polk retired to his Nashville home, where he died on June 15, 1849.

The office of chief executive under Polk was well filled. He maintained it with dignity, with integrity, and with an extraordinary sense of duty. His great influence over Congress was due to the justness of his policies and the persistence of his efforts in having the members see the questions not as interests of district or section but as matters of national welfare. His sturdy character and unblemished reputation gave weight to his counsels and strengthened his pleas. History may not rate James K. Polk as one of the greatest U.S. presidents, but his successes in office made his influence considerable.

BIBLIOGRAPHY. No published definitive work on the life and times of James K. Polk is known; however, attention is called to the biographical sketch preceding the published papers of President Polk found in JAMES D. RICHARDSON, *Messages and Papers of the Presidents*, vol. 4 (1897). The restricted accounts on aspects of his life before the presidency, E.I. MCCORMAC, *James K. Polk: A Political Biography* (1922, reprinted 1965); and C.G. SELLERS, *James K. Polk, Jacksonian, 1795–1843* (1957), are highly selective. M.M. QUAIFE (ed.), *The Diary of James K. Polk During His Presidency, 1845 to 1849*, 4 vol. (1910), is a reliable, almost complete printing of the manuscript and a revelation of a president's mind.

(G.C.R.)

Pollen Stratigraphy

Pollen stratigraphy is the record of the chronologic and geographic distribution of fossil pollen and spores in the earth's strata. This distribution provides a basis for the time-stratigraphic subdivision of rock sequences and, hence, for the correlation of strata, or geological events, from one region to another. Pollen stratigraphy is one aspect of palynology (Greek *palynein*, "to strew or sprinkle"; *cf. palē*, "fine meal"; cognate with Latin *pollen*, "flour, dust"), a term coined in 1944 to embrace all aspects of the study of pollen and spores, ranging from examination of pollen morphology by the electron microscope, to a determination of the source or adulteration of honey by analysis of its pollen content, and to a study of fossil spores extracted from ancient coals.

Pollen stratigraphy has primarily geological overtones because it is concerned with the time-space relations of fossils, but it is of great importance in such disparate fields as biology, archaeology, climatology, and anthropology, among others. Most changes in pollen profiles of the Pleistocene Epoch (10,000 to 2,500,000 years ago) reflect the impact of climatic fluctuations on the natural vegetation. But the innovation of primitive agriculture in the late Stone Age introduced a nonclimatic factor into vegetational distribution and development that is conspicuous in many pollen profiles of northern and central Europe. This change is marked by the appearance of cultivated cereal pollen and by the first appearance or greatly increased frequency of pollen of agricultural weeds that are commonly associated with human settlement. Changes in the pollen profiles give clear evidence that Neolithic man cleared the virgin forest by ax felling and burning and then began cultivation of grains. Occasional finds of corn (*Zea mays*) pollen in the southwestern United States and Mexico confirm that corn was used by prehistoric peoples of the New World and was treated as a cultivated or semi-cultivated species.

Because pollen passes through the digestive tracts of vertebrate animals and may be recovered undamaged, the examination of coprolites (fossilized fecal material) can provide precise information concerning the diet of prehistoric peoples and their animal husbandry and cultural practices. In Europe, for example, many plants have been used and valued in historic times as fodder plants, among them ivy (*Hedera*), mistletoe (*Viscum*), and elm (*Ulmus*). Study of the pollen record at Danish and Swiss sites suggests that Neolithic farmers may also have used them as fodder for their domestic animals. The pollen of all three genera showed a simultaneous decline that coincided with the occurrence of farmer cultures and the introduction of agriculture in the immediate vicinity. Climatic change was ruled out as a cause because other pollen in the assemblage indicated that this was unlikely. Direct evidence in support of ivy as fodder was found from examination of the pollen content of cultural layers. Despite the general decline of ivy pollen in the natural pollen rain, it was very abundant (*i.e.*, reaching 40 to 80 percent) in samples from the site of a stable and in feces of goats or sheep. Evidently, Neolithic farmers helped to reduce the population of ivy (and of elm and mistletoe) by feeding the flowering plants to their cattle.

This article treats the properties, transportation, distribution, and deposition of organic microfossils—principally pollen and spores—and their evolution and stratigraphic occurrence. Coverage of the paleoecology and paleogeography implied by the pollen record also is presented. For further information on the general principles of stratigraphic analysis, see STRATIGRAPHIC BOUNDARIES. See also PLEISTOCENE EPOCH; HOLOCENE EPOCH; and CLIMATIC CHANGE for relevant data on the deposits and environmental conditions of the last 2,500,000 years, which are related to the pollen record; COALS, which indicates the existence of a Carboniferous (280,000,000 to 345,000,000 years ago) pollen record; and FOSSIL RECORD for general paleontological information, including the geological occurrence and evolution of plants.

CHARACTERISTICS OF ORGANIC MICROFOSSILS

Nature and diversity of microfossils

Organic microfossils, sometimes called palynomorphs, are microfossils of generally similar size with organic walls composed of carbon, hydrogen, oxygen, and nitrogen compounds. They include pollen and spores, dinoflagellates, acritarchs, chitinozoans, bacteria, fungal spores and fruiting bodies, and fossils such as scolecodonts (jaws of certain worms). Varied plant parts and tissue fragments, such as epidermis, tracheids, trichomes, fern sporangia, and moss leaves, as well as a variety of dispersed organic matter also may be present. Some organic microfossils are whole organisms, whereas others are, or appear to be, some part of the life cycle of a larger organism. Pollen is the male reproductive structure of higher vascular plants (gymnosperms and angiosperms); spores are the asexual reproductive bodies produced by lower vascular plants (ferns, psilopsids, lycopsids, sphenopsids). The reproductive structures of nonvascular plants (bryophytes, algae, fungi) are also called spores, but only those of the bryophytes morphologically resemble spores of the so-called lower vascular plants with which they may be confused in the fossil record.

Many organic microfossils are among the group of "primitive" one-celled organisms that zoologists call animals and botanists call plants. These borderline organisms are sometimes grouped into a separate kingdom called the Protista, which includes the bacteria, algae, and fungi of some botanical classifications and the protozoa of some zoological classifications. Other organic microfossils are of problematical or unknown affinity.

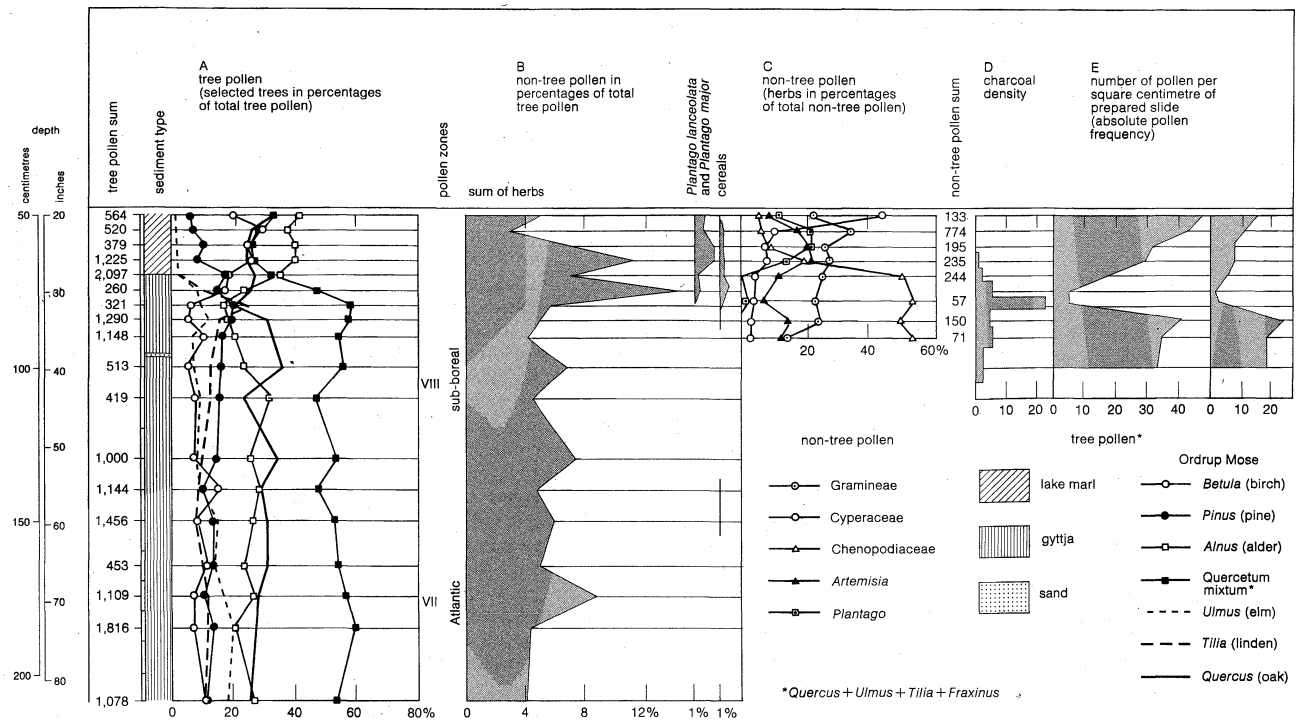


Figure 1: Pollen diagrams used to depict palynological data obtained from a vertical sequence of sediments or sedimentary rocks. They illustrate the vegetational history of an area and, in the case of the Ordrup Mose (Ordrup Bog, Denmark) diagram (E), the kinds of data that bear upon interpretations of Neolithic farming (see text).
After Geological Survey of Denmark, II. Series No. 66. Author Johannes Iversen, painter Ingeborg Frederiksen

Physical and chemical properties. As a group, palynomorphs are nonmineralized, although individuals within some of the groups, for example certain of the dinoflagellates, have mineralized (siliceous or calcareous) tests (hard parts). The chemical composition of organic microfossils does not appear to be the same for all groups and, in general, remains poorly known because the substances involved are relatively inert.

Pollen and spores are complex chemical structures of which only the outer wall, the exine, is preserved in the fossil state. The specific resistant substance in spore-pollen walls has been called sporopollenin. It appears to consist of organic substances the nature and structure of which are still not clear. Other palynomorphs may be composed of a substance similar to sporopollenin; some are described as chitinous or pseudochitinous. Chitin is a complex carbohydrate, a polysaccharide that may be regarded as a derivative of cellulose, which it resembles in many of its properties. Among living plants, chitin is confined to the green algae and the fungi, but it is found in a wide group of invertebrates within which it forms the major organic skeletal substance. Pseudochitin (composed of nitrogen, carbon, hydrogen, and oxygen) is chemically distinct from chitin. As used to describe the tests of chitinozoans, for example, the term indicates primarily that the chemical properties of these microfossils are distinct from those that characterize chitin of modern organisms. This may be the result, in part, of postdepositional modification of the original organic substance of the organism.

Organic microfossils are characteristically measured in microns (one micron equals 0.001 millimetre). Size ranges for representative groups are as follows: pollen and spores, five to 200 microns; dinoflagellates, 20 to 350 microns; acritarchs, seven to 500 microns; chitinozoans, 30 to 1,500 microns. Typical sizes in each group fall between these extremes; most dinoflagellates, for example, are about 60 to 150 microns, and most pollen about 20 to 60 microns. Large spores (megaspores) derived from heterosporous plants such as *Selaginella* may vastly exceed the size limits of most pollen and spores, and some megaspores, including fossil specimens, may be between one and seven millimetres (1,000 and 7,000 microns). The largest organic microfossils can be observed directly without the microscope, or with the use of only minimum magnification (50 ×).

General shape and appearance

Organic microfossils are (as their diverse biological affinities would suggest), varied in their appearance. There are four principal groups.

Chitinozoans are typically rod-, club-, and flask- or trumpet-shaped with a terminal opening. The tests are smooth or partially or entirely covered with short spinelike protrusions. Tests are commonly black and opaque but may be amber and translucent. Chitinozoans occur singly or in chains of as many as six individuals.

Dinoflagellates are commonly spherical, ellipsoidal, or polygonal; the basic shape is sometimes modified by outgrowths. The test is often divided into a system of smooth or variously ornamented plates whose junction is clearly marked or, in the case of cysts of the type generally called hystrichospheres, the plate pattern is indistinct or absent. Hystrichospheres often have long, sometimes bifurcating, spinelike processes radiating from the surface.

Acritarchs are varied, and may be spherical, ellipsoidal, discoidal, elongate or polygonal. Some acritarchs lack ornamentation; others possess spines or raised ridges. Ornamentation, if present, is random or confined to certain parts of the test. Some acritarchs resemble some dinoflagellate cysts.

Spores and pollen—though treated as a group—differ in appearance. Spores are typically bilateral-monolete (kidney-bean shaped) or radial-trilete (rounded triangular). They are smooth or variously ornamented. Pollen are morphologically the most complex organic microfossils. They typically have one to many apertures that are commonly round or ellipsoidal (pores), or elongate (furrows or colpi), or a combination of the two. Pollen grains are symmetrical and possess complex surface patterns.

Extraction of samples from rocks

The extraction of microfossil samples from sediments and sedimentary rocks is accomplished by chemical and mechanical techniques. The chemical techniques are based on the use of acids and bases which dissolve or disaggregate mineral and certain organic fractions of the rock, but to which the microfossils are resistant. Silicates and carbonates are removed by hydrochloric and hydrofluoric acid. Organic materials, other than microfossils, are destroyed by oxidation using, for example, nitric acid, hydrogen peroxide, Schulze's solution (nitric acid plus

potassium chlorate), or by acetylation (sulfuric acid plus acetic anhydride). Oxidation, which attacks the lignin fraction of organic residues, and acetylation, which attacks the cellulose fraction, may be used on the same sample. Various mechanical techniques are effective in removing material that survives chemical processing. The most successful of these involve heavy liquid flotation that will float the organic microfossils but permit heavier, organic or inorganic fractions to sink. Commonly used heavy liquids include zinc chloride and bromoform. The extraction procedure must be done with considerable care to avoid contaminating the sample with airborne modern pollen or with fossils of a different age; selective destruction of some of the microfossils also must be avoided. The occurrence of any of these errors will bias the fossil assemblage and will lead to errors of interpretation of the assemblage.

Environments of accumulation. The principal environments of accumulation of organic microfossils are continental, continental-marine, and marine, which correspond to the depositional environments of sedimentary rocks. Continental environments are terrestrial, such as soils, glacial ice, freshwater lakes, and swamps and bogs. The marine-nonmarine environments include near-shore littoral, deltaic, and estuarine sites. Marine environments extend from immediately below the tidal limits to the deepest oceanic zone.

Organic microfossils are found in each principal environment and in many specialized sub-environments. Pollen and spores, the most widespread palynomorphs, are found in soils, fluviatile deposits, and glacial ice, as well as in transitional and marine environments; their occurrence is by no means uniform in these environments, however. Terrestrial bogs, lake sediments, soils, and cave deposits are important in pollen stratigraphy of Pleistocene deposits and archaeological sites. Transitional marine-nonmarine environments with interbedded continental and marine deposits are abundant sites of pollen accumulation, although the littoral zone is not generally favourable to pollen preservation because of its instability and frequent accumulation of coarse sediments. Richly fossiliferous coal deposits near sea level, with interbedded terrestrial bog and marine shales, accumulated in transitional enviroments. Marine environments, especially those of the continental shelf, are extremely favourable to pollen preservation but pollen-spore density in ocean sediments drops off at the continental slope and into the abyssal (deep sea) zone.

The other principal groups of palynomorphs, the chitinozoans, acritarchs, and dinoflagellates, are entirely or largely confined to marine environments. A few acritarchs have been recovered from nonmarine rocks and some fossil freshwater dinoflagellates are known, but such finds are exceptional.

Problems of interpretation can arise because the contamination potential of a fossil assemblage is high. Organic microfossils are so indestructible that they may survive several generations of deposition and erosion and, thus, migrate from one environment to another. Under some circumstances redeposited specimens are abundant and, depending on the age differential, redeposited and younger microfossils may be indistinguishable. Contamination may also occur when percolating groundwater picks up pollen from the rocks it transgresses and redeposits these alien grains elsewhere. Drilling muds containing pollen and spores foreign to the host rock provide still another possibility for contamination.

STRATIGRAPHY AND SEDIMENTATION

Influence of life habits on occurrence

Transportation and deposition of microfossils. Most organic microfossils, exclusive of pollen and spores, appear to be the remains of planktonic (free-floating) marine organisms. Living dinoflagellates primarily inhabit near-surface layers of the oceans; presumably this was true of the fossil forms when they were living. They are found in marine, brackish-water, and freshwater environments, though they are most abundant in marine environments. Hystrichospheres (encysted dinoflagellates with radiating structures) seem to manifest decreased diversity and abundance in areas of reduced salinity adjacent to shorelines, and a general increase in absolute abundance seaward. Acritarchs, to a large extent extinct, also appear to have been planktonic because their remains are common in a variety of sedimentary rocks of marine origin. A few nonmarine acritarchs occur in Pleistocene and younger strata. Chitinozoans, wholly extinct, are also found in rocks of many lithologies. They too are presumably marine plankton, although some may have been benthonic (bottom dwellers) and attached to the substratum. The frequent association of chitinozoans with corals and crinoids has led to the speculation that these problematical micro-organisms lived in maximum concentration in warm, shallow, well-aerated waters, although the association of chitinozoans with the mineral glauconite suggests that some also lived in oxygen-poor environments. Pollen and spores are derived largely from land-based vegetation; the habitats of individual plants may include freshwater lakes and streams and some few monocotyledons, such as ditch grass (*Ruppia*), grow submerged in salt and brackish water. Pollen and spores dispersed by wind and water transgress widely into marine environments, as well as into terrigenous environments that are alien to their immediate source vegetation. Lacustrine and fluviatile micro-organisms, such as the green alga *Pediastrum*, are also found in marine deposits as a result of transportation. Marine-based micro-organisms, however, seldom transgress into continental environments.

Occurrence in sedimentary rocks

Because they are the remains of free-floating organisms, or wind- and water-dispersed structures, organic microfossils are not necessarily restricted to environments primarily suited to the living organisms. Organic microfossils, thus, occur in sedimentary rocks representing many environmental types. The usefulness of organic microfossils for determining the ecology of depositional environments therefore varies. Some empirical generalizations are obvious. The presence of dinoflagellates and acritarchs, for example, generally indicates deposition in marine environments. The presence of pollen and spores without an admixture of marine microfossils is indicative of continental conditions. Generalizations based on empirical observations such as the ratio of marine to nonmarine microfossils and the association of distinct microfossil assemblages with distinctive rock types may also be useful in determining the ecology of depositional environments. In a vertical sequence of essentially homogeneous rock strata, for example, a mixed assemblage of marine- and land-derived microfossils in one part of the section, compared with an entirely land-derived assemblage in another part of the section, probably indicates a change in depositional environments from open marine conditions distant from shore, to shallow-water, near-shore, or continental environments. When distinctive microfossil assemblages are confined to distinctive rock types, the combined sedimentological-microfossil data may be necessary to determine the relation between the microfossil assemblage and the depositional environment if the affinities of the specimens are unknown. This is true of certain interbedded Carboniferous coals and marine shales of England; the sequence of coals contains similar pollen-spore floras that are distinct from those that occur in the sequence of marine shales. The coal plant assemblage represents a recurring swamp flora that is obliterated by marine transgressions and then re-established following sea regression. The swamp flora is independent of the stratigraphy and will reoccur, at least locally (in restricted areas) in this particular part of the Carboniferous, wherever swamp conditions are re-established. The environmental significance of the different microfossil assemblages would not have been apparent without knowledge that they were restricted to particular and distinctive rock types and sequences.

When botanical relationships of fossils are determinable, they may provide sensitive indicators of depositional environments without resorting to empirical observations. In coals and lignites, for example, a pronounced representation of swamp vegetation is expected. Sometimes vegetation that is environmentally suggestive may

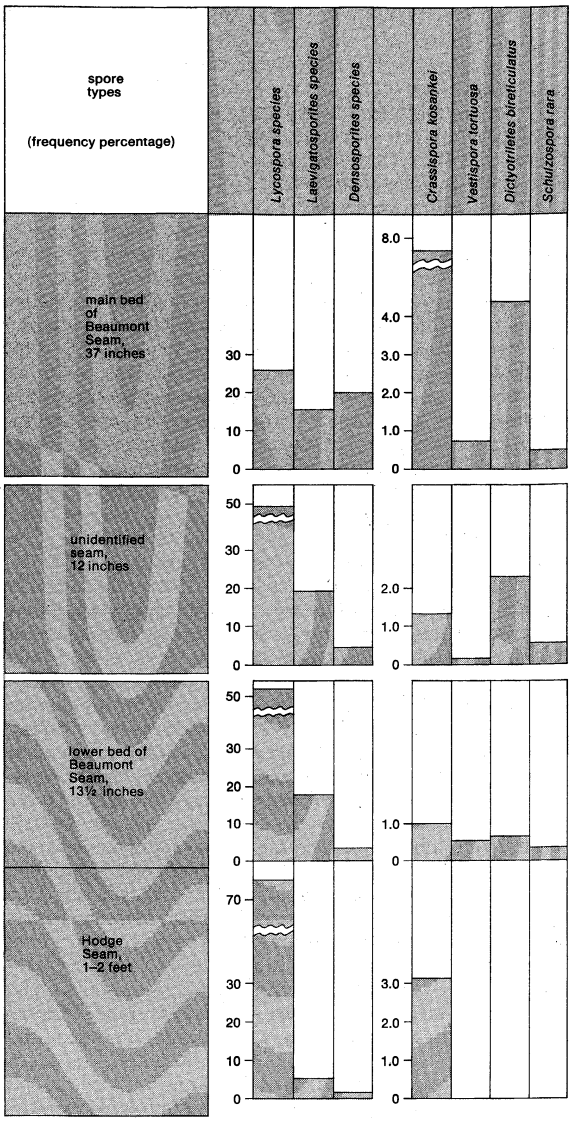


Figure 2: Miospore frequencies in coal seams at Montagu Colliery, Great Britain.

Reprinted with the permission of the Palaeontological Association from the original figure used in *Palaeontology*

occur in situations in which the environment in question is not immediately obvious from the lithology. Some types of coastal vegetation may thus show pronounced quantitative fluctuations dependent on local conditions. This is the case beginning in the Oligocene, in such remote regions as Venezuela, Nigeria, New Guinea, and Borneo, where large quantities of pollen of the *Rhizophora* type point to the introduction and perpetuation of mangrove-swamp vegetation in the tidal zone along these coasts.

Dispersal of pollen by wind and water

Palynomorphs are transported by wind and by water. Agents such as insects and birds also play a part in pollen dispersal, and the long distance transport of pollen may involve hundreds or thousands of miles. Pollen of the Southern Hemisphere beech (*Nothofagus*) has been found in the South Atlantic island of Tristan da Cunha, to which it apparently was borne from South America at least 4,500 kilometres (2,800 miles) to the west. Pollen of the conifers *Podocarpus* and *Dacrydium* have been found in Pacific island peats at least 700 kilometres (435 miles) from their apparent New Zealand source. Pollen from North America is found over the mid-Atlantic and alder (*Alnus*) pollen occurs on the continental shelf near Trinidad; the latter was waterborne from an Andean source 800 kilometres (500 miles) to the west at elevations of 2,000 to 3,000 metres (6,550 to 9,850 feet). Few instances of long-distance pollen transport are quantitatively important.

The distribution of planktonic aquatic palynomorphs largely reflects water currents and turbulence but the restriction of pollen-spore studies to terrestrial environments for many years has emphasized their aerial distribution by the wind as of prime importance. Though it has long been known that buoyant conifer pollen floats on water for extended periods, the potential of rivers and ocean currents as avenues of pollen travel are still poorly explored.

The amounts of pollen suspended in estuarine and open ocean water and contained in surface sediments from deltas, the continental shelf, and the ocean basins have demonstrated that pollen is abundant in marine and transitional marine-continental environments, and that water transport in noncontinental environments may well exceed wind as the primary agent. For example, 2,000 grains of pine (*Pinus*) pollen was concentrated from 100 litres (26 gallons) of water on the Great Bahama Banks as much as five months after pine had flowered. Delaware Bay water contained 200 pollen grains per gallon in June. Pollen density in the open ocean during June was far less, although at a depth of 2,240 fathoms, 100 gallons of water still contained 68 pollen grains, a richer concentration than in surface layers where approximately one to 18 grains per 100 gallons were recovered.

In surface sediments from the Orinoco River Delta and the Gulf of Paria, the abundance of pollen and spores per gram of sediment decreases seaward, with heaviest concentrations opposite the larger delta tributaries. Because the wind direction is predominantly easterly, marine currents must be the main transport medium. Studies on the Great Bahama Bank also demonstrate that wind is not accountable for the distribution of pollen in surface sediments. The distribution of pine pollen, which remains suspended in water for long periods, was related neither to proximity of pine trees nor to wind direction during flowering, but was explained by water turbulence and sedimentation patterns. Studies of surface sediments adjacent to the coast of Israel also demonstrate the effectiveness of the Nile and the currents of the Mediterranean in pollen and spore transport.

It is clear that spores and pollen that settle directly into moving water, or that do so after some wind transport, may be transported great distances from their source areas. Water turbulence keeps pollen and spores in suspension because of their natural buoyancy. The lighter fraction is winnowed from the heavier and is carried with the lighter sedimentary particles an increasing distance from vegetation and sediment sources. Selectivity of pollen on the basis of size occurs with increasing distance seaward; the highest concentrations of small, light pollen are preferentially associated with silt-size particles with which waterborne pollen compares in settling velocity (rate of fall of particles in still water). Size selection and concentrations of pollen where water turbulence is minimal and sediments are uniformly fine-grained emphasize that the transportational and depositional characteristics of pollen are similar to those of sedimentary particles of similar size and density.

Limits of dispersal

The normal limits of effective pollen and spore dispersal by water are still unknown, but they appear to be tens of miles. Pollen and spores are found in workable quantities both on the continental shelf and in some ocean basins, although the abundance and the diversity both drop off with increasing distance from shoreline. In the Gulf of Mexico the absolute abundance of pollen and spores per gram of surface sediment drops off rapidly up to a distance of about 30 miles offshore, after which it tends to become constant. At 80 miles from shore, surface sediments still contain between 2,000 and 3,000 spores and pollen grains per gram. Off the Atlantic Coast of North America the concentration per gram of sample increases eastward from about 20 miles to a maximum of about 185 miles offshore, and then it decreases toward the ocean basins. The greatest diversity of types is found on the continental shelf. In the Pacific Ocean pollen and spores are recovered from all bottom sediments of marginal seas (Sea of Japan, Sea of Okhotsk, Bering Sea), but along meridional profiles from Hawaii to New Zea-

land only surface sediments from comparatively nearby islands contain pollen and spores in any quantity. Surface sediments from the South Pacific 100 to 150 kilometres (60 to 90 miles) east of New Zealand have conifer pollen and spores but almost no angiosperm pollen, whereas sediments 300 to 400 kilometres east of New Caldeonia have mainly spores and almost no pollen.

Palynomorphs behave like sedimentary particles in another way. They may survive the erosion of sedimentary rocks and retransportation and redeposition. Recycled organic microfossils may involve several generations of rocks.

Pollen and spores travel at least a short distance in air before settling on land and water surfaces. As in water, pollen is suspended in air according to turbulence and natural buoyancy. Pollen from vegetation at any one locality, therefore, is not necessarily deposited at the same time within a limited geographic region. Experiments show that great quantities of the most buoyant pollen may be updrafted in warm, dry, high-velocity air currents beneath cumulus clouds or along mountain fronts to be transported horizontally for great distances at high altitudes. Pollen fallout from the upper air masses occurs during periods of relative calm. The heavy pollen drops first—the light pollen may remain continuously suspended for varying times. Transport distance and suspension time are largely determined by pollen density and by meteorological conditions such as wind strength and direction and relative humidity. Rain, especially of small droplets, will collect and eliminate airborne spores and pollen almost immediately. In areas of daily rains the effectiveness of aerial transport of pollen and spores is greatly restricted. Fifty to 100 kilometres (30 to 60 miles) is generally accepted as the effective limit of aerial pollen dispersal. The aerial pollen density drops off rapidly away from vegetation-clothed landmasses although both tree and herb pollen have been trapped across the Atlantic from Göteborg to New York during the early summer.

Eh–pH conditions favouring preservation

Preservation and alteration of pollen and spores. The best natural environments for pollen-spore preservation are said to be those in which the potential oxidation of substances (Eh) is low or negative (reducing, anaerobic environments) and the hydrogen ion concentration (pH) is low (acidic environments). Anaerobic, acidic environments are common in bogs and in the deepwater layers of some lake and ocean bottoms. Some alkaline (high pH), anaerobic Pleistocene lake deposits yield well-preserved, abundant pollen except where the Eh rises; then the pollen is scarce or absent. Alkaline evaporite deposits such as salt and gypsum also contain well-preserved pollen and spores as their only significant fossil. High pH carbonate deposits (limestones) are also said to be unfavourable for pollen-spore preservation, but plant microfossils are found in modern carbonate sediments from the Grand Bahama Banks and the Persian Gulf, and dinoflagellates, acritarchs, chitinozoans, and others are common and well-preserved in calcareous rocks. In some Upper Cretaceous limestones it has been found that laboratory oxidation destroys the available pollen. This has given rise to the postulate that the resistance to oxidation of pollen depends on the pH of environments in which the pollen are preserved. Pollen and spores have low concentration in both limestones and saline rocks, and neither is found in dolomites (calcium magnesium carbonate), although chitinozoa occur in this rock type.

The highest concentrations of well-preserved palynomorphs are found in unaltered, slowly deposited, fine-grained sedimentary rocks where: (1) distortion due to sediment compaction and sedimentary overburden is minimum, and (2) post-depositional weathering (oxidation) also is minimum. Excessive overburden of several thousand or more feet of rock produces a low-grade metamorphism (rock alteration) with the gradual carbonization of plant fragments and pollen and spores. With the increasing weight of overburden these microfossils lose their structure and eventually are converted to amorphous carbon. Some low-grade metamorphic coals do yield pollen and spores, but in the highest rank coals (*e.g.*, anthracites), they may be difficult to isolate from the coal matrix without chemically attacking the pollen and spores. In such floras only durable, resistant types of microfossils are preserved and liberated. Porous rocks long exposed to atmospheric oxygen or groundwater may be weathered to great depths and their original pollen and spore content partially or entirely destroyed. Red beds, for example, whether deposited in an already oxidized state or whether subjected to *in situ* oxidation following deposition, are commonly free of pollen and spores and other organic material, although interbedded nonred shales may contain plant microfossils.

Stratigraphic correlation. The application of pollen and spores to the solution of stratigraphic and other geologic problems has been extensively explored in some parts of the geological column; in problems involving Carboniferous coals, German Tertiary brown coals, and Quaternary history, for example, their use has become essentially routine. The use of pollen and spores and other organic microfossils for stratigraphic correlation of a variety of sedimentary rocks of all ages has gained increasing popularity among petroleum companies in recent years.

Utility of organic microfossils

Organic microfossils are stratigraphically useful because of their extensive geographic distribution, limited stratigraphic ranges of genera and species, great abundance of some groups throughout much of the geological column (the rock record of all geological time), independence from local environmental conditions, and their distinctive and diverse morphology that permits characterization of individual genera and species. Accordingly, many organic microfossils are useful for local, regional, interregional, and long distance correlation.

Because of their general resistance to biochemical degradation and their easy dispersal, organic microfossils occur in certain types of sediments and sedimentary rocks that lack other fossils. Black shales deposited in reducing (anaerobic) environments, or saline deposits that are generally unsuitable for the preservation of other fossils, are good repositories for palynomorphs, for example. Because of their small size, thousands of specimens may occur in a few grams of rock; this is a special advantage in studying drill core samples. Organic microfossils may be especially valuable if fossils traditionally used for correlation and age determinations are absent. Organic microfossils are among the only Precambrian (older than 570,000,000 years) fossils with stratigraphic promise, although they are still imperfectly known and do not yet provide an independent basis for stratigraphic correlation and dating of strata. Organic microfossils may complement information from other fossil groups when these are present, as in the case of many marine and brackish water environments. In particular, pollen and spores, either alone or in conjunction with other fossils, are especially useful in predominantly continental stratigraphic sequences such as bogs, lake deposits, and archeological sites, or where marine and continental rocks alternate within a single depositional environment as in many Carboniferous (280,000,000 to 345,000,000 years ago) coal deposits. In mixed or transitional environments pollen and spores are the most useful microfossils, for they are widely and simultaneously deposited and thus provide a means of correlating sedimentary rocks across marine, brackish water, and continental environments. Most of the traditional stratigraphic fossils are restricted to one of these environments or have only limited stratigraphic use when they occur in others.

Chronology of floral changes

Pollen stratigraphy has become the cornerstone of Pleistocene stratigraphy by providing a chronological scale against which to measure events of the Quaternary Period (the last 2,500,000 years). In Quaternary deposits, pollen zones are established by significant changes in the character of the vegetation, such as the appearance and disappearance of some forest type, or changes in dominance of some of the principal pollen types. Quaternary vegetational changes are primarily the result of regional climatic fluctuations; sometimes they are the result of some factor not climatically predicated, such as the introduction of agriculture by Neolithic man in Europe. The

initiation of purely local environmental changes may result in the introduction of local plant communities. Changes of the latter type have local stratigraphic significance of potential paleogeographic and paleoecologic interest. When the botanical affinities of most of the specimens are determinable, as in the Quaternary, it is possible to recognize what vegetational changes are due to nonclimatic or local factors. For most of pre-Cenozoic time (prior to 65,000,000 years ago), pollen and spores are arbitrarily designated without regard to their taxonomic and phylogenetic relationships. Under these circumstances, the ability to recognize purely local environmental changes and their specific vegetational assemblages is limited.

The correlation and age determination of pre-Quaternary strata by pollen and spores also depends on major differentiation of floras through time on a regional basis. But in comparison with the Quaternary, the order of magnitude differs as do some of the mechanisms that initiate floral changes. The Quaternary zonation (subdivisions) involves rock units that represent only a few thousand years; pre-Quaternary studies may involve rock units that span several million years. Climatic fluctuations that are by and large the "normal" determinant of floral change in the Quaternary in practice occur at a reduced magnitude and pace during much of the rest of geologic time. Most traceable pre-Quaternary floral changes are determined principally by organic evolution and secondarily by major topographic or climatic changes.

Vegetational changes that are determined by evolution involve the morphological modification of specific types or the replacement of one floral assemblage by another because of the introduction of new plant groups and the extinction of old groups. Floral changes useful in stratigraphic work thus involve changes in taxonomic composition (qualitative) and in relative abundance (quantitative) of members of plant assemblages. Similar floral successions of a regional nature provide the basis for correlation of sedimentary rocks from area to area by their pollen and spore content. In a later section of this article (*Stratigraphic distribution and evolution*), some floral changes useful in stratigraphic work are traced.

Correlation of coals and coal seams

One of the first applications of plant microfossils in stratigraphic work was to Carboniferous coals in the United States and Great Britain in the 1920s and 1930s. Within the Carboniferous coals, variation in spore content in individual coals was found to differ. Once individual coals were characterized, it was possible to correlate coal beds by means of their spores. Many of the changes from coal to coal, depend on: (1) striking changes in generic and specific content, with some genera and species having long stratigraphic ranges and some having very restricted ranges; (2) characteristic assemblages of spore types; and (3) pronounced changes in the abundance of genera and species. One bed is correlated with another if assemblages in the two are the same and the abundance or frequencies of types is at least roughly similar. The precision of correlations made by this means is improved if the ecology of the coal swamp flora and the environmental conditions relating to the deposition are also utilized as fully as possible.

In favourable instances individual, adjacent coal seams can be identified and correlated by their spore assemblages. During mining operations at a colliery in England, for example, the question arose whether a 12-inch coal seam was the lower bed of the Beaumont Seam, which overlay it or whether it was a part of the upfaulted Hodge Seam, which occurred stratigraphically beneath the Beaumont Coal. Comparisons of the relative frequencies of certain spores in composite samples of each of the coals showed that the Lower Beaumont and the unidentified seam were correlative.

Stratigraphic distribution and evolution. *Pollen and spores.* The evolution and stratigraphic distribution of pollen and spores are traced primarily from morphological types. It is possible, however, to relate morphological types of pollen and spores to major plant groups whose times of appearance, dominance, and extinction are known from paleobotanical data, and sometimes to determine the relationship of microfossils to megafossil groups. The major evolutionary events of the palynological record generally parallel those of the plant kingdom traced from megafossil remains.

Plant microfossils called spores may be functionally different. Spores of one size, sexually undifferentiated, and produced by homosporous plants are homospores or isospores. Spores produced by heterosporous plants are commonly of different sizes and are called microspores (male) and megaspores (female). Isolated spores the affinities of which are unknown can seldom be distinguished as homospores or microspores and not always as microspores or megaspores. Pollen are the microspores of seed plants, living and fossil. The microspores of some primitive Paleozoic gymnosperms appearing to have the germinal aperture in the proximal rather than the distal position of modern pollen are sometimes referred to as prepollen.

Spores commonly have one or two principal aperture types and symmetries—monolete bilateral (kidney-bean shaped) or radial-trilete (rounded triangular). Some are inaperturate (alete). Gymnosperm pollen, including primitive types, are broadly saccate (with wings, bladders, or sacci) or nonsaccate. Primitive types may possess a trilete aperture. Saccate types have the general configuration of modern conifer pollen: they are monosaccate, bisaccate, or polysaccate (three or more sacci). Gymnosperm pollen of nonsaccate type may have one to many furrows or no apertures. Angiosperm pollen are classified primarily on the number and distribution of pores and furrows.

The Paleozoic record

Sporelike fossils found by Soviet workers in pre-Silurian strata (older than 430,000,000 years) are not accepted as constituting proof of the existence of terrestrial vascular plants of such antiquity, nor have they been substantiated by comparable records elsewhere. No land plants that could have borne these spores are known.

Trilete spores assigned to genus *Ambitisporites*, which probably represent vascular land plants, first occur in Silurian rocks of Libya, Spain, and Great Britain. Silurian spores are of limited numbers and variability—only 24 small, comparatively simple types, representing at most eight genera—are known by the end of the period (Lower Downtonian). The relationship of dispersed Silurian spores to Silurian plants is unknown.

By the end of the Lower Devonian (the Devonian Period began 395,000,000 years ago and extended over 50,000,000 years in time), however, abundant, diverse trilete spores represent most of the genera that characterize the rest of the period. Because of their small and essentially uniform size, Silurian and earliest Devonian spores are judged to be homospores. An increase in spore size accompanied the introduction of new spore types during the Devonian. Some large Lower Devonian spores, in excess of 200 microns (one micron equals 0.001 millimetre), may be megaspores and indicate the condition of heterospory, although plant megafossils clearly demonstrated to be heterosporous are unknown prior to the Middle to Late Devonian, when the spore evidence is also more convincing. Rapid diversification of spore types during the uppermost Upper Devonian and the Lower Carboniferous again introduced many new types. Uppermost Devonian spores are more similar to those of the Lower Carboniferous than to spores of the rest of the Devonian or of the Silurian. Several genera of dispersed small spores and one dispersed megaspore genus can be related to Devonian plant genera.

Microfloras of Carboniferous age (from 280,000,000 to 345,000,000 years ago) include homospores, megaspores, and microspores, and the first specimens classified as pollen. Carboniferous spores and pollen provide many significant guide fossils (fossils with restricted stratigraphic ranges) that are useful in correlation and age determination. Carboniferous assemblages are commonly dominated by spores and pollen of radial-trilete configuration. Monolete spores (a morphologic type rarely recorded in the Devonian) are first prominent in the Pennsylvanian (approximately correlative with the Upper Carboniferous). Monolete spores with coarsely ornamented surfaces are an innovation of the Middle to Late Pennsyl-

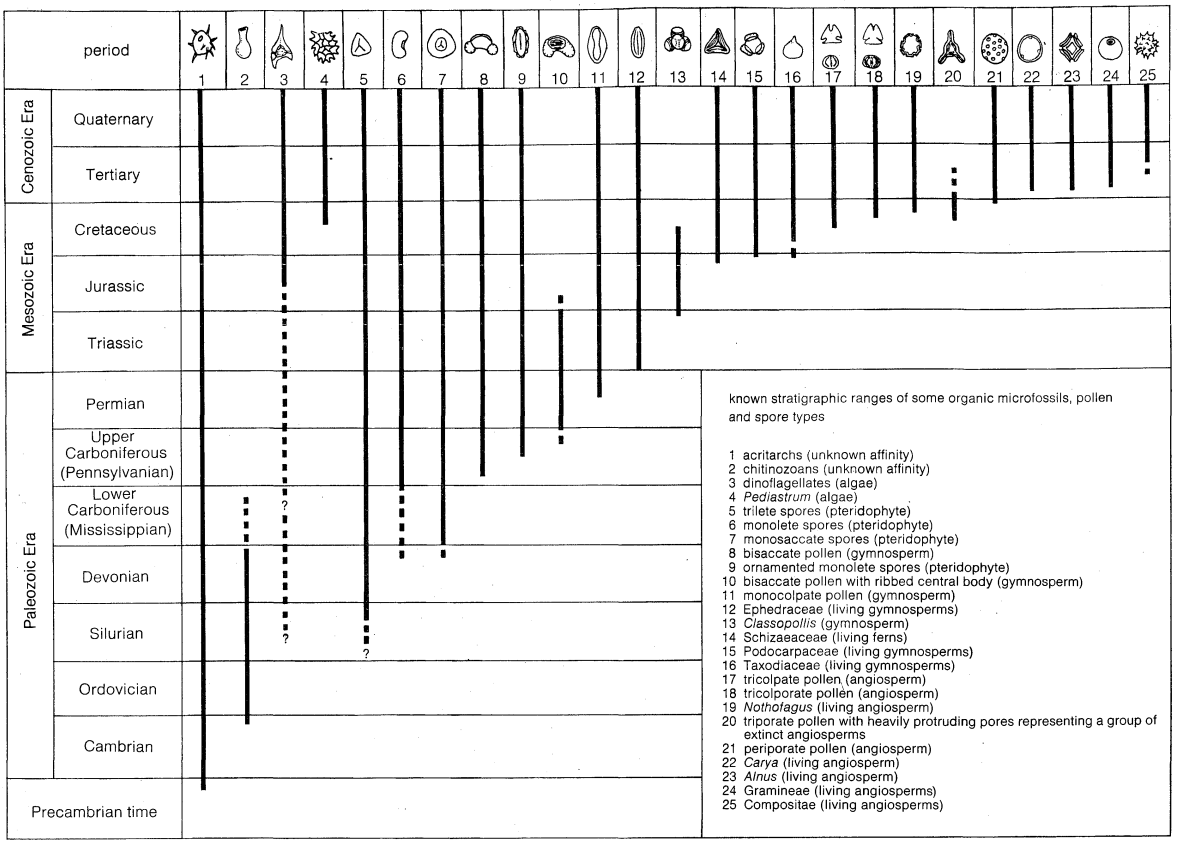


Figure 3: Known stratigraphic ranges of some organic microfossils, pollen, and spore types.

vanian. A few Carboniferous spores and pollen are alete. Carboniferous spores greatly exceed those of the Silurian and the Devonian in the diversity and complexity of their surface ornamentation. Many possess a variety of equatorial structures. Characteristic sacci or wings appear for the first time on spores and on pollen. Carboniferous saccate genera are predominantly monosaccates, some having a lycopsid source, others a gymnosperm source. Bisaccate genera proliferate in the Late Pennsylvanian and are common to dominant by the Early Permian, largely replacing monolete and trilete spores. Bisaccates are considered gymnosperm pollen or prepollen even when their provenance is unknown. One trilete, multisaccate spore genus is known. Three to 16 (depending on classification) Carboniferous trilete megaspore genera are recognized and are among the largest found; *Cystosporites*, for example, may reach a diameter exceeding one centimetre.

Spore-pollen floras of Permian age (from 225,000,000 to 280,000,000 years ago) are dominated by bisaccate pollen. Some prominent Permian bisaccate genera bore thickened longitudinal ribs on the central body. This type, the parent source of which is unknown, became extinct or essentially so in the early part of the Mesozoic Era (from 65,000,000 to 225,000,000 years ago). Bisaccates without ribs and types with a proximal monolete mark are also characteristic of Permian spore-pollen floras. Monosaccates are less diverse and abundant than bisaccates. Nonsaccate, single-furrowed and multifurrowed gymnosperm pollen are forerunners of these abundant types in the Mesozoic. The single-furrowed *Cycadopites* is the most common; it may have been produced by a variety of plants allied to the cycadophytes, the ginkgoes, or the pteridosperms (seed ferns). Permian trilete and monolete spores differ from Carboniferous genera mainly at the species level; trilete types are more abundant than monoletes.

Permian pollen floras are gymnospermous and they are dominated by prominent Mesozoic groups (conifers, cycadophytes, ginkgoes, and pteridsperms) in Upper Permian strata. Lycopsids, sphenopsids, and the ferns declined from their Carboniferous maximum occurrence, judged by diminished numbers of trilete spores, especially in the Late Permian. The prominence of Mesozoic groups and the decreased importance of significant Paleozoic groups give the Permian a Mesozoic rather than a Paleozoic stamp.

The Mesozoic record

Palynological changes between the Late Permian and the Early Triassic (the Triassic Period began 225,000,000 years ago and lasted 35,000,000 years in time) are less pronounced than those that occurred during the Permian. Many Triassic spore and pollen genera occur either in the Permian or in the Jurassic (from 136,000,000 to 190,000,000 years ago), although some are restricted to the Triassic. Bisaccate gymnosperm pollen are prominent, and nonribbed forms are more abundant and diverse than ribbed types; only three Permian genera with ribbed bodies survive into the Triassic and only one into the Lower Jurassic. Also occurring in the Triassic, but gaining prominence in Jurassic and Cretaceous floras, are nonsaccate, single-furrowed gymnosperm pollen of the type first recorded in the Permian and multifurrowed specimens possibly representing the living Ephedraceae. One such furrowed type, *Eucommiidites*, first appearing in the Middle Triassic, was originally regarded as angiosperm pollen. Trilete spores rival or outnumber bisaccate pollen in prominence only in the Late Triassic. Most of these types are thought (on megafossil evidence) to represent ferns, whereas trilete spores of the Late Paleozoic had lycopsid, sphenopsid, and fern affinities. A number of Triassic spores are assigned to living fern families (Osmundaceae, Gleicheniaceae, and others), and some are attributed to the bryophytes.

Major plant groups of the Triassic continue into the Jurassic and Cretaceous (the Cretaceous Period began 136,000,000 years ago and lasted 71,000,000 years). Bisaccate gymnosperm pollen are important and sometimes dominant in Jurassic microfloras but decline in diversity and size during the period. The abundance of *Parvisaccites*, a bisaccate genus characterized by small sacci, in the Early Cretaceous temporarily reverses the general Jurassic decline in bisaccates. Some Cretaceous bisaccates are morphologically inseparable from pollen of modern conifers. Monosaccate gymnosperm pollen

decline after the Triassic, although one such type, *Tsugaepollenites*, is a conspicuous element in some European Middle and Late Jurassic floras. Trisaccate gymnosperm pollen are important in the Early Cretaceous of the Southern Hemisphere. Such types are allied to the Podocarpaceae, a prominent Southern Hemisphere family. Nonsaccate, multifurrowed pollen, first recorded in the Triassic and similar to pollen of the living genus *Ephedra*, are prominent in Cretaceous strata from equatorial regions such as Venezuela, Colombia, and Nigeria, where bisaccate conifer pollen are absent. Single-furrowed nonsaccates with ginkgoalean and cycadophyte affinities are present in both Jurassic and Cretaceous pollen-spore floras, though seldom abundant. Spores of the Jurassic and Cretaceous commonly resemble those of a variety of modern fern and fern-allied families. The abundant trilete spores often display elaborate and diversified sculpture patterns and equatorial structures. Widespread and common, especially during the Early Cretaceous, were types with heavy ribs, fine to coarse warts, and radial equatorial appendages. Many of these have affinities with the living fern family Schizaeaceae. Jurassic and Early Cretaceous megaspores are more diversified than the known heterosporous land plants that could have been their source. Some of them appear comparable to megaspores of some living members of the Selaginellaceae; others may have been related to the water ferns, Salviniaceae and Marsileaceae. Spores of bryophyte types are also found during the Jurassic but reach greater diversity in the Early Cretaceous.

Toward the end of the Early Cretaceous, rare angiosperm pollen appear in the fossil record. No evidence suggests an earlier occurrence in either the tropical regions or the Southern Hemisphere, both among the postulated sites of initial angiosperm evolution, although some limited palynological evidence suggests that angiosperm pollen may be relatively more abundant in Upper Cretaceous strata of tropical and subtropical regions. The first simple tricolpate (three-furrowed) types are harbingers of the enormous elaboration of angiosperm pollen to follow in the Late Cretaceous and Early Tertiary (the Tertiary Period began 65,000,000 years ago and lasted 62,500,000 years) during the palynological transition from archaic pteridophyte-gymnosperm floras of the Mesozoic to angiosperm floras of the Cenozoic. In early Upper Cretaceous floras angiosperm pollen are limited in diversity and abundance or absent entirely; they are seldom assignable to living families or genera, including those identified from leaf remains. By contrast, megafossil floras of similar age have a large number of angiosperms attributed to living families. This discrepancy has not been satisfactorily explained; it may hinge on ecological factors or on the forced assignment of many fossil leaf genera to extant genera and families on superficial similarities in their leaf characters. The gradual increase in the abundance of angiosperm pollen during the Late Cretaceous is accompanied by increase in their diversity. Many of the types elaborated became extinct in a relatively short time during the Late Cretaceous and the Early Tertiary. Certain of them are thus ideal for stratigraphic correlation and zonation. In some Southern Hemisphere pollen-spore floras—as, for example, those from New Zealand—Podocarpacean pollen continue to dominate microfloras into the latest Cretaceous and earliest Tertiary. In South America (Colombia), an influx of palm pollen and dicotyledonous types, foreshadowing the modern flora, replaced the predominantly fern floras of older and more primitive stamp by the latest Upper Cretaceous. Some latest Cretaceous floras from western North America are markedly different from European and Australian floras of the same age, indicating definite provincialism (geographic restriction).

The Cenozoic record

The transition from the Mesozoic to the Cenozoic Era (the last 65,000,000 years) is not without floral change, but the dramatic change in the world's vegetation marked by the onset of angiosperms occurred within the Cretaceous Period. Angiosperms, primarily dicotyledons, dominate Tertiary floras, but palm pollen was common in some strata of Early Tertiary age, and gymnosperm pollen of *Taxodium* and *Sequoia* types were locally abundant.

Pollen floras of the Paleocene, Eocene, and Oligocene (the first three epochs of Tertiary time) gradually assume a more modern aspect, and the major morphological classes of angiosperm pollen are recognized. Because the elaboration of new forms and the rate of extinction among angiosperms appear to have been high during this interval, Early Tertiary pollen commonly are unassignable to modern taxa. Nevertheless pollen of a number of living genera including alder (*Alnus*), hickory (*Carya*), linden (*Tilia*), maple (*Acer*), sweet gum (*Liquidambar*), and holly (*Ilex*) can be recognized at least by the Eocene (38,000,000 to 54,000,000 years ago). By the end of the Oligocene (26,000,000 to 38,000,000 years ago), major evolution and dispersal within most angiosperm families appears to have been completed.

Miocene and Pliocene pollen floras are progressively modernized as related to modern plant communities. Most Late Tertiary and all Quaternary (the last 2,500,000 years) pollen and spores are related to living genera and species of plants. With decreasing age, pollen floras take on the cast of the local vegetation—a decreasing proportion of the flora is exotic. Increasing provincialism makes it more difficult to make generalizations about the stratigraphic distribution of specific pollen-spore types. Few evidences of organic evolution have been recognized in Late Tertiary and Quaternary vegetation. Changes mainly involve plant redistribution related to environmental factors, but a significant mid-Tertiary evolutionary event is the introduction of the sunflower family (Compositae). This predominantly herbaceous group is the largest angiosperm family and has global distribution. Most taxonomists agree that composites are the most advanced dicotyledonous plants. The family lacks authentic megafossil remains, possibly because of its largely herbaceous nature; its pollen first appears in the Miocene or the latest Oligocene. Pollen attributable to other herbaceous plants also assumes increasing importance in the Late Tertiary pollen floras, although a number of herbs have Early Tertiary records. Pollen assigned to the Chenopodiaceae, for example, is recorded in the Paleocene and grass pollen in the Eocene. Corn (*Zea mays*) is a member of the grass family that apparently evolved sometime during the Late Pleistocene.

Dinoflagellates, acritarchs, and chitinozoans

Other organic microfossils. Dinoflagellate fossils consist mainly of resistant encysted specimens. Dinoflagellates are known from Silurian deposits, but they are uncommon Paleozoic and early Mesozoic fossils with limited stratigraphic usefulness. Dinoflagellates become abundant in Middle Jurassic time when a variety of cyst types appeared. During the rest of the Mesozoic and Tertiary, the group appears to have evolved rapidly and many species have worldwide distribution. The dinoflagellates are an important stratigraphic group from mid-Mesozoic onward because of the occurrence of many cosmopolitan (widespread) species, the progressive introduction of new types that gain dominance, and the decline and disappearance of early types.

The acritarchs were rapidly evolving, abundant forms during the early Paleozoic, when most other organic microfossils have no fossil records or are too scarce to have any stratigraphic impact. Acritarchs are at least locally abundant in Precambrian strata and may prove useful for stratigraphic subdivision during that interval. From middle Paleozoic onward, most acritarchs are too long-ranging to be stratigraphically useful. In Upper Jurassic and younger deposits, acritarchs are scarce fossils, although Cretaceous, Tertiary, and nonmarine Quaternary forms are known. *Tasmanites*, a genus locally abundant in Paleozoic marine black shales, is sometimes included in the acritarchs, although some workers regard *Tasmanites* and its possible allies as closely similar to certain living planktonic green algae. *Tasmanites* and related Paleozoic forms are too long-ranging to have stratigraphic potential.

Chitinozoans appear to be ideally suited both for intercontinental and local stratigraphic work because of restricted stratigraphic ranges, rapid evolution, and commonness in a wide variety of sedimentary rocks of ap-

propriate age. Chitinozoans first appeared in the Late Cambrian or Early Ordovician and apparently had become largely extinct by Late Devonian. A single specimen has been reported from Mississippian strata of North America.

PALEOECOLOGY AND PALEOGEOGRAPHY

Fossil palynomorphs were once living organisms or parts of living organisms that influenced their environments and were influenced by them. Among organic microfossils, Cenozoic pollen and spores provide the most significant source of paleoecological inferences. Whether concerned with pollen and spores or with other fossils, paleoecology suffers from a variety of limitations of evidence. The principal of these for pollen and spores are: (1) the incompleteness of the fossil record, either as preserved or as known; (2) the limited geologic time during which spores and pollen can be related to living species; (3) the assumption that fossil plants represented by spores and pollen and their living relatives have maintained similar ecological requirements through geologic time; (4) the difficulty of deducing the environments of organisms with no living relatives; and (5) the presence of pollen and spores in assemblages and environments quite different from those in which their source plants lived, due to long-distance transport or reworking from older deposits.

Environmental change. The reliability of paleoecological inferences deduced from pollen and spores drops off rapidly as their geologic age increases. Only during the Cenozoic and the latest Mesozoic are pollen and spores assignable to living genera and not always then. The most valid inferences are based on fossils or subfossils of the Pleistocene where pollen may be assigned to living species of plants. Since pioneering work in Scandinavia in the early 1900s, pollen studies (pollen analysis) have served as the principal guide to climatic and environmental change of the Quaternary, particularly the late Quaternary. Thousands of pollen studies throughout the world have traced the sequence of former vegetation types and deduced from them fluctuations of Quaternary climate. On the basis of significant geographic change in vegetation it is possible to divide much of Pleistocene time into pollen zones each with a climatic basis. Because the concern is with existing plant communities and with familiar species or genera whose ecology is determinable, Pleistocene pollen analysis is a highly sensitive paleoclimatological technique.

The fact that Pleistocene as well as late Tertiary pollen floras seldom contain disharmonious mixtures of genera lends an air of reliability to the ecological conclusions drawn from them. The reconstruction of plant communities and their environments even from subfossil pollen assemblages is hampered, however, by the difficulty of relating pollen fallout to vegetation in some quantitative way. Part of this difficulty is overcome by seeking "good matches" between fossil pollen fallouts and those of modern communities of plants. By such direct comparisons it may be possible to determine the significance of pollen frequencies in fossil assemblages.

Paleoecological inferences from single fossils or small groups of fossils are strengthened by data from other fossil groups and from sedimentology. Thus, the presence of the freshwater alga *Pediastrum* is no proof of terrestrial environments because *Pediastrum* is found in bottom sediments of the continental shelf off the Pacific Coast and in marine Cretaceous rocks. Pollen and spores from land-based plants including those restricted to fresh water, such as cattail (*Typha*) and water lily (*Nymphaea*), also occur in ocean sediments where they have been borne by wind and water. Dinoflagellates, acritarchs, chitinozoans, and other marine or brackish-water microorganisms seldom transgress into terrestrial environments.

Pollen as a climatic indicator

A classic example of the use of pollen as a climatic indicator is provided by a study of holly (*Ilex aquifolium*), whose pollen is identified in interglacial and postglacial deposits of northern Europe. The presence of holly pollen has enhanced significance; the low pollen productivity related to insect pollination virtually guarantees nearby occurrence of the plant. When a detailed study of the climatic requirements of holly in northern Europe was made, using the mean temperatures of the warmest and coldest months at meteorological stations near holly sites to establish a thermal correlation curve, the curve suggested that holly is intolerant of winter frost, but not especially demanding of high summer temperatures. During hard winters in Denmark, holly was badly damaged and a large proportion was killed. For the intervals when holly pollen is recorded, mild, oceanic winters based on the previously established climatic index were deduced. A German interglacial site of holly pollen now lies beyond the winter temperature limits of *Ilex aquifolium*; winter climate at that site was clearly milder than it is today. Companion studies of ivy (*Hedera helix*) and mistletoe (*Viscum album*), also recorded as pollen, amplify and reinforce this climatic interpretation. Ivy is only slightly more tolerant of winter cold than holly. Mistletoe can tolerate lower winter temperatures than either holly or ivy, but its present northern range is restricted by higher summer temperature requirements than now obtain in northern Europe. The parameters controlling the distribution of these species probably operated in the past. Together holly and ivy are regarded as the best indicators of winter temperatures and mistletoe of summer temperatures in northern Europe. When found together during postglacial periods they indicate higher winter and summer temperatures than those of today.

Plants of the more remote geologic past, including those assigned to living genera, provide a greater challenge to interpretation, for habitat requirements may have changed through time. Commonly, this seems the best explanation for disharmonious palynological assemblages, barring misidentification of some of the members. Some fossils, by their morphology, provide clues to changing environmental adaptations. Thus some Tertiary leaves gradually reduce in size through time with increasing drought. Pollen and spores show no such changes. Fossil pollen that are virtually identical to pollen of living plants with specific habitat requirements can be traced for millions of years. In the absence of supporting evidence, the assumption that these plants always occupied the same ecological niche is questionable. The gymnosperm *Ephedra*, for example, is presently an arid land shrub of deserts and dry woodlands; as dispersed fossil pollen, named *Gnetaceaepollenites* and *Ephedripites*, it has a pollen record that extends into the early Mesozoic. *Ephedra* pollen is frequent in some Tertiary floras where the rest of the assemblage provides little indication of conditions favourable to the genus. Pollen of *Ephedra* is distinctive and not readily confused with that of other plants. Local conditions of dryness of the soil in highly insolated sites may have favoured the occurrence of *Ephedra* for millions of years in habitats virtually identical to those it now occupies. But the ecological requirements of some of the fossil species may have been quite different from those of the living species.

Location of ancient shorelines. The potential usefulness of pollen distribution for locating shoreline position was noted in a patent issued to an oil company palynologist in 1954, in which it was claimed that nearshore sediments might contain as many as 7,500 pollen grains per gram of sediment and that the ratio of large to small pollen was about 1:4. At increasing distances from the shoreline, small pollen was more common and grains per gram of sediment progressively decreased. Studies of the Orinoco River Delta and continental shelf in the vicinity of Trinidad generally support the offshore decrease in pollen and spores, but locally the nature of the shoreline, wind direction, river supply, presence of coarse-textured sediments, mineral sedimentation rates, and current patterns upset the picture so that the position of the coastline is not invariably reflected by the palynology of the sediments. In environments of carbonate rock deposition such as the Great Bahama Bank, pollen was finally sedimented largely as a result of fluctuations in turbulence and other hydrographic factors not clearly related to land proximity.

Considerable work is needed to substantiate the reliabil-

ity of pollen and spores alone to define shoreline position, although in marine Silurian rocks from England, it was noted that rocks coming from sites consistent with proximity to the supposed shoreline yielded among their components 90 percent spores (of presumed land plants) whereas more offshore samples contained only about 1 percent spores.

The potential of redeposited palynomorphs as shoreline indexes is suggested by study of the Eocene London Clay of southern England. There, 17 percent derived Jurassic and Cretaceous acritarchs, dinoflagellates, and pollen and spores were found nearest the reconstructed Eocene shoreline, but dropped off to about 6 percent at the locality most distant from the shoreline. A Silurian assemblage of acritarchs that consists almost entirely of redeposited pre-Silurian specimens was also found in close proximity to the position of the postulated Silurian shoreline. Redeposited Eocene pollen was similarly found in greatest abundance in the Orinoco Delta, close to its source, but decreased rapidly toward the sea.

Evidence of water depth fluctuation

Some evidence suggests that morphology, taxonomic diversity, and population density are environmentally controlled in dinoflagellates and acritarchs, and that these microfossils are therefore useful as indexes to depositional environments and to shoreline position and water depths. Rich varied assemblages of Jurassic acritarchs from Great Britain, for example, are found in strata indicative of open sea conditions, whereas inshore, basin environments are marked by a reduction in diversity and abundance; a group of small acritarchs that belong primarily to a single genus is dominant. Changes in the diversity and morphology of the acritarch assemblages appeared to trace cycles of sedimentation corresponding to transgression and regression of the sea. Oligocene dinoflagellates from the southeastern United States also appear to show distinctive morphology related to shoreline position and water depth.

In general, dinoflagellate frequency drops off from the open ocean toward land, suggesting the usefulness of the ratio of these microplankton to pollen and spores. Such a ratio was used to interpret water-depth fluctuations and, hence, periods of sea transgression and regression during deposition of the Upper Cretaceous Mancos Group of Utah. In the Upper Cretaceous Frontier Formation of Wyoming, the ratio of pollen and spores to marine microplankton appeared to delineate marine and nonmarine sedimentation based on other information.

The geographic distribution of fossil assemblages may suggest major changes in landmass distribution, some of which is of global significance. The areal distribution of Permian spore assemblages on a worldwide basis delimits distinct paleobotanical provinces that conform to the classical concept of the landmasses of Gondwanaland and Laurasia. These distinct assemblages seldom transgress the limits of the paleobotanical provinces originally defined by leaf genera in establishing Permian biogeography. Typical Gondwanian pollen-spore assemblages are found in Australia, Africa, India, South America, and Antarctica. They are palynologically distinct from Laurasian pollen-spore assemblages known from Europe, North America, the Soviet Union, Iraq, Turkey, and southern China. Not all authorities agree that continental drift is the only possible mechanism by which to explain such striking geographic differentiation of floral assemblages. Other hypotheses, however, must postulate the existence of major geographic barriers to account for the varied and disparate Permian floras.

BIBLIOGRAPHY. A.T. CROSS (ed.), *Palynology in Oil Exploration* (1964), a symposium discussing the use of pollen and spores and other organic microfossils in stratigraphic correlation; G. ERDTMAN, *An Introduction to Pollen Analysis* (1943), a general treatment of several aspects of palynology, with concise descriptions and drawings of many pollen and spore types, and *Handbook of Palynology: An Introduction to the Study of Pollen Grains and Spores* (1969), emphasis on pollen morphology, with numerous photomicrographs and drawings and some treatment of general aspects of palynology; K. FAEGRI and J. IVERSEN, *Textbook of Pollen Analysis*, 2nd rev. ed. (1964), discussion of theoretical and practical problems, primarily as they pertain to Pleistocene pollen analysis; H. GODWIN, *The History of the British Flora: A Factual Basis for Phytogeography* (1956), a comprehensive account of Pleistocene and Recent history of the British flora including the contribution made by pollen analysis and discussion of the use of pollen in archaeology; J. GRAY, "Techniques in Palynology," pt. 3, pp. 469–706, in B. KUMMEL and D.M. RAUP (eds.), *Handbook of Paleontological Techniques* (1965), description of laboratory and other techniques used to secure and study organic microfossils; J. IVERSEN, "Land Occupation in Denmark's Stone Age: A Pollen Analytical Study of the Influence of Farmer Culture on the Vegetational Development," *Danm. Geol. Unders.*, 2nd series, no. 66 (1941), first recognition by pollen study of the influence of prehistoric man on vegetation, and "*Viscum, Hedera* and *Ilex* as Climatic Indicators," *Geol. Fören, Förh.*, 66:463–483 (1944), an example of the use of pollen records in making paleoecological deductions about Pleistocene climate; G.O. KREMP, *Morphologic Encyclopedia of Palynology* (1965), comprehensive dictionary of morphologic terms used to describe pollen and spores, with drawings and photomicrographs illustrating morphological features; P.S. MARTIN and F.W. SHARROCK, "Pollen Analysis of Prehistoric Human Feces: A New Approach to Ethnobotany," *American Antiquity*, 30:168–180 (1964), use of pollen to determine dietary and cultural practices of early man in the American Southwest; J. MULLER, "Palynology of Recent Orinoco Delta and Shelf Sediments: Reports of the Orinoco Shelf Expedition," *Micropaleontology*, 5:1–32 (1959), comprehensive study of distribution of pollen and other organic microfossils in modern sediments of the continental shelf, offshore Venezuela, with brief discussion of applications to older marine sedimentary rocks; V. POKORNÝ, *Grundzüge der zoologischen Mikropalaeontologie*, 2 vol. (1958; Eng. trans., *Principles of Zoological Micropalaeontology*, 1963), comprehensive account of microfossils including some organic microfossils such as chitinozoans, scolecodonts, dinoflagellates, and acritarchs; G. SHAW and A. YEADON, "Chemical Studies on the Constitution of Some Pollen and Spores Membranes," *J. Chem. Soc. (C)*, pt. 1, pp. 16–22 (1966), recent research on the chemistry of pollen and spore membranes plus review of previous work; A.H.V. SMITH and M.A. BUTTERWORTH, *Miospores in the Coal Seams of the Carboniferous of Great Britain* (1967), comprehensive account of spore floras of the Carboniferous coals of Great Britain with discussion of theoretical and practical aspects of using pollen and spores in stratigraphic correlation; R.H. TSCHUDY and R.A. SCOTT (eds.), *Aspects of Palynology: An Introduction to Plant Microfossils in Time* (1969), comprehensive account of the stratigraphic and geographic distribution of pollen, spores, and other organic microfossils; R.P. WODEHOUSE, *Pollen Grains: Their Structure, Identification and Significance in Science and Medicine* (1935), classic work, providing one of the first detailed accounts of pollen morphology.

(J.Gr.)

Pollination

Pollination is the transference of pollen grains from the stamens, the flower parts that produce them, to the ovule-bearing organs or to the ovules (seed precursors) themselves. In conifers and cycads, in which the ovules are exposed, the pollen is simply caught in a drop of fluid secreted by the ovule. In flowering plants, however, the ovules are contained within a hollow organ called the pistil, and the pollen is deposited on the pistil's receptive surface, the stigma. There the pollen germinates and gives rise to a pollen tube, which grows down through the pistil toward one of the ovules in its base. In an act of double fertilization, one of the two sperm cells within the pollen tube fuses with the egg cell of the ovule, making possible the development of an embryo, and the other cell combines with the two subsidiary sexual nuclei of the ovule, which initiates formation of a reserve food tissue, the endosperm. The growing ovule then transforms itself into a seed. As a prerequisite for fertilization, pollination is essential to the production of fruit and seed crops and plays an important part in programs designed to improve plants by breeding. Furthermore, studies of pollination are invaluable for understanding the evolution of flowering plants and their distribution in the world today. As sedentary organisms, plants usually must enlist the services of external agents for pollen transport. In flowering plants, these are (roughly in order of diminishing importance): insects, wind, birds, mammals, and water.

Importance to agriculture

Pollination by insects probably occurred in primitive seed plants, reliance on other means being a relatively recent evolutionary development. Reasonable evidence indi-

cates that flowering plants first appeared in tropical rain forests during the Mesozoic Era (about 65,000,000 to 225,000,000 years ago). The most prevalent insect forms of the period were primitive beetles; no bees and butterflies were present. Some Mesozoic beetles, already adapted to a diet of spores from primitive plants, apparently became pollen eaters, capable of effecting chance pollination with grains accidentally spared. The visits of such beetles to primitive flowering plants may have been encouraged by insect attractants, such as odours of carrion, dung, or fruit, or by sex attractants. In addition, visits of the insects to the plants could be made to last longer and thus potentially be more valuable to the plant as far as fertilization was concerned, if the flower had a functional, traplike structure. Nowadays, such flowers are found predominantly, although not exclusively, in tropical families regarded as ancient; *e.g.*, the water lily (Nymphaeaceae) and the arum lily (Araceae) families. At the same time, other plants apparently began to exploit the fact that primitive gall-forming insects visited the flowers to deposit eggs. In the ancient genus *Ficus* (figs and banyan trees), pollination still depends on gall wasps. In general, Mesozoic flowering plants could not fully rely on their pollinators, whose presence also depended on the existence of a complete, well-functioning ecological web with dung, cadavers, and food plants always available. More advanced flowers escaped from such dependence on chance by no longer relying on deceit, trapping, and tasty pollen alone; nectar became increasingly important as a reward for the pollinators. Essentially a concentrated, aqueous sugar solution, nectar existed in certain ancestors of the flowering plants. In bracken fern even nowadays, nectar glands (nectaries) are found at the base of young leaves. In the course of evolutionary change, certain nectaries were incorporated into the modern flower (floral nectaries), although extrafloral nectaries also persist. Flower colours thus seem to have been introduced as "advertisements" of the presence of nectar, and more specific nectar guides (such as patterns of dots or lines, contrasting colour patches, or special odour patterns) were introduced near the entrance to the flower, pointing the way to the nectar hidden within. At the same time, in a complex pattern of parallel evolution, groups of insects appeared with sucking mouthparts capable of feeding on nectar. In extreme cases, there arose a complete mutual dependence. For example, a Madagascar orchid, *Angraecum sesquipedale*, with a nectar receptacle 20 to 35 centimetres (eight to 14 inches) long, depends for its pollination exclusively on the local race of a hawkmoth, *Xanthopan morganii*, which has a proboscis of 22½ centimetres (nine inches). Interestingly enough, the existence of the hawkmoth was predicted by Charles Darwin and Alfred Russel Wallace, codiscoverers of evolution, about 40 years before its actual discovery.

Development of nectar

TYPES OF POLLINATION

An egg cell in an ovule of a flower may be fertilized by a sperm cell derived from a pollen grain produced by that same flower or by another flower on the same plant, in either of which two cases fertilization is said to be due to self-pollination (autogamy); or, the sperm may be derived from pollen originating on a different plant individual, in which case the process is called cross-pollination (heterogamy). Both processes are common, but cross-pollination clearly has certain evolutionary advantages for the species: the seeds formed may combine the hereditary traits of both parents, and the resulting offspring generally are more varied than would be the case after self-pollination. In a changing environment, some of the individuals resulting from cross-pollination still may be found capable of coping with their new situation, ensuring survival of the species, whereas the individuals resulting from self-pollination might all be unable to adjust. Self-pollination, or selfing, although foolproof in a stable environment, thus is an evolutionary cul-de-sac. There also is a more direct, visible difference between selfing and outbreeding (cross-pollination): in those species where both methods work, cross-pollination usually produces more, and better quality, seeds. A dramatic demonstration of this effect is found with hybrid corn (maize), a superior product that results from cross-breeding of several especially bred lines.

Self-pollination and cross-pollination

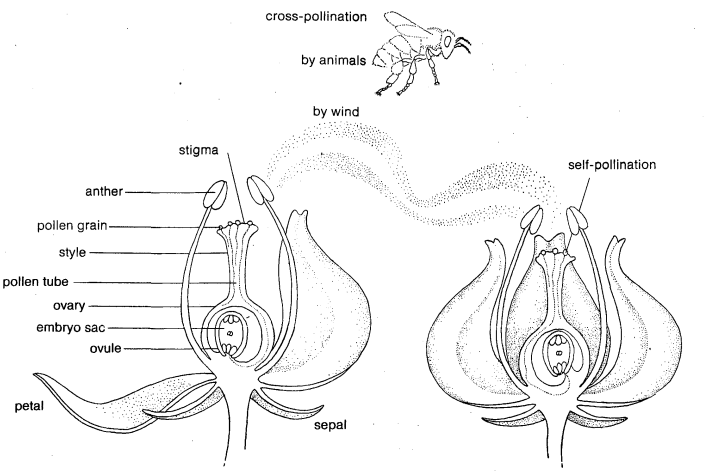


Figure 1: Types of pollination.
Drawing by M. Moran based on B.J.D. Meeuse, *The Story of Pollination*

Mechanisms that prevent self-pollination. *Structural.* Not surprisingly, many species of plants have developed mechanisms that prevent self-pollination. Some—*e.g.*, date palms (*Phoenix dactylifera*) and willows (*Salix* species)—have become dioecious; that is, some plants produce only "male" (staminate) flowers, with the rest producing only "female" (pistillate or ovule-producing) ones. In species in which staminate and pistillate flowers are found on the same individual (monoecious plants) and in those with hermaphroditic flowers (flowers possessing both stamens and pistils), a common way of preventing self-fertilization is to have the pollen shed either before or after the period during which the stigmas on the same plant are receptive, a situation known as dichogamy. The more usual form of dichogamy, which is found especially in such insect-pollinated flowers as fireweed (*Epilobium angustifolium*) and salvias (*Salvia* species), is proterandry, in which the stamens ripen before the pistils. Proterogyny, the situation in which the pistils mature first, occurs in arum lilies and many wind-pollinated plants, such as grasses—although several grasses are self-pollinated, including common varieties of wheat, barley, and oats. Avocado has both proterogynous and proterandrous varieties, and these often are grown together to encourage cross-fertilization. A structural feature of flowers that discourages selfing is heterostyly, or variation in the length of the style (neck of the pistil). This occurs in the common primrose (*Primula vulgaris*) and species of wood sorrel (*Oxalis*) and flax. In most British primrose populations, for example, approximately half the individuals have so-called "pin" flowers, which possess short stamens and a long style, giving the stigma a position at the flower's mouth, whereas the other half have "thrum" flowers, in which the style is short and the stamens are long, forming a "thrumhead" at the opening of the flower. Bees can hardly fail to deposit the pollen they receive from one type of flower onto the stigmas of the other type. The genetic system that regulates flower structure in these primroses is so constituted that cross-pollination automatically maintains a 50:50 ratio between pins and thrums. In the flowers of purple loosestrife (*Lythrum salicaria*), the stamens and styles are of three different lengths to limit self-fertilization.

Chemical. Chemical self-incompatibility is another device for preventing self-fertilization. In this phenomenon, which depends on chemical substances within the plant, the pollen may fail to grow on a stigma of the same flower that produced it or, after germination, the pollen-tube may not grow normally down the style to effect fertilization. The process is controlled genetically; it need not be absolute and can change in degree during the flowering season. Not surprisingly, chemical incompatibility usually is not found in those plants that have strong structural or temporal barriers against self-pollination.



Wind pollination: pollen of the lodgepole pine *(Pinus contorta)* and mountain hemlock *(Tsuga mertensiana)* on the surface of Crater Lake, Oregon.



Beetle pollination: pollen-eating longhorn beetle (family Corambycidae) in a thimbleberry (*Rubus parviflorus*).



Butterfly pollination: the extended proboscis of the heliconian butterfly *(Heliconius)* carries pollen between flowers of *Lantana camara*.



Bee pollination: pollen-laden honeybee *(Apis mellifera)* on a wall flower *(Erysimum)*.



Wind pollination in grasses: yellow free-hanging anthers (pollen producers) and white feathery stigmas (pollen collectors) of meadow fescue *(Festuca pratensis)* provide maximum wind exposure.



Bird pollination: Anna's hummingbird (*Calypte anna*) feeding on nectar of a fuchsia.



Hawkmoth pollination: hawkmoth (Sphingidae) hovering near a honeysuckle *(Lonicera caprifolium)*.

Agents of pollen dispersal

Plate 1: (Top right, centre) E.S. Ross, (top left) Joe Van Wormer—Bruce Coleman Inc., (centre left) Rudolf Schmid, (centre right) Mary W. Ferguson, F.P.S.A., (bottom left) Walter E. Harvey—National Audubon Society, (bottom right) Ingmar Holmasen



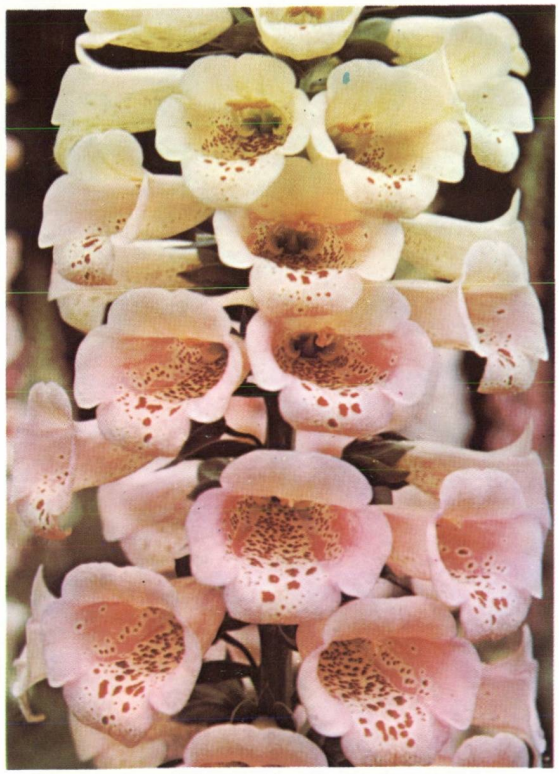
Trap flower: pollinating psychodid flies trapped in the floral chamber of cuckoopint (*Arum maculatum*; vertical section with spathe cut away).



Deception flower: carrion flower *(Stapelia)* has the appearance and odour of decayed meat. Flies laying eggs crawl over flower's surface and come into contact with stigmas and stamens.



Trap flower: when certain bees land on the hairy, upright petal of the flower of grass-pink orchid (*Calopogon pulchellus*), the petal topples forward, and a visiting bee falls on its back onto the sticky surface of the column.



Nectar guides: spots in throat of foxglove *(Digitalis purpurea* variety *gloxiniaeflora)* provide insects with a visual pathway to nectar glands.







Explosive flower: (top) flower of Scotch broom *(Cytisus scoparius)*. (Bottom left) The bumblebee *(Bombus vosnesenskii)* triggers release of the stamens (bottom right) which forcibly strike underside of bee and dust it with pollen.

Flower adaptations related to pollination

Plate 2: (Top left, top right) B.J.D. Meeuse, (centre) Rudolf Schmid, (centre left, bottom left, bottom right) E.S. Ross, (centre right) Derek Fell

Formation of one such mechanism during evolution apparently was enough for most plant species.

Circumstances favouring self-pollination

Mechanisms that permit self-pollination. In many instances, successful self-pollination takes place at the end of a flower's life-span if cross-pollination has not occurred. Such self-pollination may be achieved by curving of stamens or style as occurs, for example, in fireweed. It can be an evolutionary advantage when animal pollinators are temporarily scarce or when the plants in a population are widely scattered. Under such circumstances, selfing may tide the species over until better circumstances for outbreeding arrive. For this reason, selfing is common among annual plants; these often must produce an abundance of seed for the rapid and massive colonization of any bare ground that becomes available. If, in a given year, an annual plant were to produce no seed at all, survival of the species might be endangered. A persistent habit of self-pollination apparently has been adopted successfully by some plant species whose natural pollinators have died out. Continued selfing also is practiced by many food-crop plants. Some of these plants are cleistogamous, meaning that the flowers fail to open, an extreme way of ensuring self-pollination. A similar process is apomixis, the development of an ovule into a seed without fertilization. Apomixis is easily demonstrated in lawn dandelions, which produce seeds even when stamens and styles are cut off just before the flowers open. Consistent apomixis has the same pros and cons as continued selfing. The offspring show very little genetic variability, but there is good survival if the species is well adapted to its habitat and if the environment does not change.

AGENTS OF POLLEN DISPERSAL

Insects. *Beetles and flies.* The ancient principle of trapping insects as a means of ensuring pollination was readopted by some advanced families (*e.g.*, orchids and milkweeds), and further elaboration perfected the flower traps of primitive families. The cuckoo pint (*Arum maculatum*), for example, attracts minute flies, which normally breed in cow dung, by means of a fetid smell. This smell is generated in early evening, along with considerable heat, which helps to volatilize the odour ingredients. The flies visiting the plant, many of which carry *Arum* pollen, enter the floral trap through a zone of bristles and then fall into a smooth-walled floral chamber from which escape is impossible. Gorging themselves on a nutritious stigmatic secretion produced by the female flowers at the base of the chamber, the flies effect cross-pollination. Late at night, when the stigmas no longer function, the male flowers, situated much higher on the floral column, proceed to bombard the flies with a rain of pollen. The next day, when smell, heat, and food are gone, the prisoners, "tarred" with stigmatic secretion and "feathered" with pollen, are allowed to escape by a wilting of the inflorescence (flower cluster). Usually the escaped flies are soon recaught by another inflorescence, which is still in the smelly, receptive stage, and cross-pollination again ensues. Superb timing mechanisms underlie these events. The heat-generating metabolic process in the inflorescence is triggered by a hormone, calorigen, originating in the male flower buds only under the right conditions. The giant inflorescences of the tropical plant *Amorphophallus titanum* similarly trap large carrion beetles.

Trap flowers

In general, trap flowers victimize beetles or flies of a primitive type. Although beetles most likely were involved as pollinators when flowering plants as a group were born, their later performance in pollination has been disappointing. Some modern beetles do visit smelly flowers of an open type, such as elderberry and hawthorn, but with few exceptions they are still mainly pollen eaters. Flies as a group have become much more diversified in their habits than beetles have. Female short-tongued flies may be deceived by open-type flowers with carrion smells; *e.g.*, the flowers of *Stapelia* and *Rafflesia*. Mosquitoes with their long tongues are effective pollinators of certain orchids (*Habenaria* species) in North American swamps. In Europe, the bee fly (*Bombylius*) is an important long-tongued pollinator. Extremely specialized as nectar drinkers are certain South African flies; for example, *Megistorhynchus longirostris*, which has a tongue that is 60 to 70 millimetres (2.3 to 2.7 inches) long.

Pollen flowers

The voraciousness of flower beetles demonstrates the futility of enticing insect pollinators solely with such an indispensable material as pollen. As a defensive strategy, certain nectar-free flowers that cater to beetles and bees —such as wild roses, peonies, and poppies—produce a superabundance of pollen. Other plants—*e.g.*, *Cassia*—have two types of stamens, one producing a special sterile pollen used by insects as food, the other yielding normal pollen for fertilizing the ovules. Other flowers contain hairs or food bodies that are attractive to insects.

Bees. In the modern world, bees are probably the most important insect pollinators. Living almost exclusively on nectar, they feed their larvae pollen and honey (a modified nectar). To obtain their foods, they possess striking physical and behavioral adaptations, such as tongues as long as 2½ centimetres (1 inch), hairy bodies, and (in honeybees and bumblebees) special pollen baskets. The Austrian naturalist Karl von Frisch has demonstrated that honeybees, although blind to red light, distinguish at least four different colour regions, namely, yellow (in-

Drawing by M. Pahl based on (E) photograph courtesy of B.J.D. Meeuse; (A) reprinted with permission of Macmillan Publishing Co., Inc. from *A Textbook of General Botany*, 3rd ed., by G.M. Smith *et al.* Copyright 1935 by The Macmillan Company, renewed 1963 by Genevieve Allan, Helen P. Smith and Katherine N. Bryan; (B,F,G,) B.J.D. Meeuse, *The Story of Pollination* (Copyright © 1961), The Ronald Press Company, New York; (C,D) *An Evolutionary Survey of the Plant Kingdom* by R.F. Scagel, G.E. Rouse, J.R. Stein, R.J. Bandoni, W.B. Schofield, T.M.C. Taylor. © 1965 by Wadsworth Publishing Company, Inc., Belmont, California 94002. Reprinted by permission of the publisher

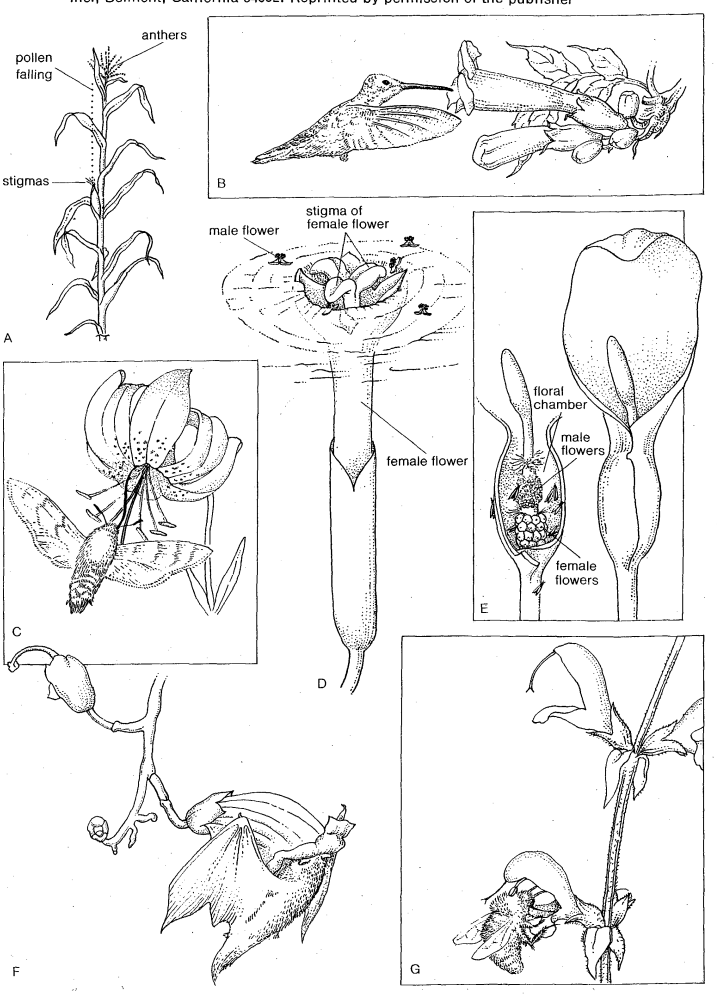


Figure 2: *Examples of different pollination modes:* (A) Pollination by wind in corn (*Zea mays*). (B) Bird pollination of a trumpet-vine flower by a hummingbird. (C) Hawkmoth taking nectar from petal pouch of lily and pollinating at the same time. (D) Water pollination of *Vallisneria americana* in which floating male flowers come in contact with stigmata of larger female flower. (E) A trap flower (*Arum italicum*) pollinated by insects trapped in the flower structure for a time. (F) Bat pollinating a flower of the sausage tree (*Kigelia*). (G) Pollination of *Salvia pratensis* by bumblebees, which are touched on the back by either the stigma or the anthers, depending on which stage the flower is in.

cluding orange and yellow green), blue green, blue (including purple and violet), and ultraviolet. Their sensitivity to ultraviolet enables bees to follow nectar-guide patterns not apparent to the human eye. They are able to taste several different sugars and also can be trained to differentiate between aromatic, sweet, or minty odours but not foul smells. Fragrance may be the decisive factor in establishing the honeybee's habit of staying with one species of flower as long as it is abundantly available. Also important is that honeybee workers can communicate to one another both the distance and the direction of an abundant food source by means of special dances.

Bee flowers

Bee flowers, open in the daytime, attract their insect visitors primarily by bright colours; at close range, special patterns and fragrances come into play. Many bee flowers provide their visitors with a landing platform in the form of a broad lower lip on which the bee sits down before pushing its way into the flower's interior, which usually contains both stamens and pistils. The hermaphroditism of most bee flowers makes for efficiency, because the flower both delivers and receives a load of pollen during a single visit of the pollinator, and the pollinator never travels from one flower to another without a full load of pollen. Indeed, the floral mechanism of many bee flowers permits only one pollination visit. The pollen grains of most bee flowers are sticky, spiny, or highly sculptured, ensuring their adherence to the bodies of the bees. Since one load of pollen contains enough pollen grains to initiate fertilization of many ovules, most individual bee flowers produce many seeds.

Examples of flowers that depend heavily on bees are larkspur, monkshood, bleeding heart, and Scotch broom. Alkali bees (*Nomia*) and leaf-cutter bees (*Megachile*) are both efficient pollinators of alfalfa; unlike honeybees, they are not afraid to trigger the explosive mechanism that liberates a cloud of pollen in alfalfa flowers. Certain Ecuadorian orchids (*Oncidium*) are pollinated by male bees of the genus *Centris*; vibrating in the breeze, the beelike flowers are attacked headlong by the strongly territorial males, who mistake them for competitors. Other South American orchids, nectarless but very fragrant, are visited by male bees (*Euglossa* species) who, for reasons not yet understood, collect from the surface of the flowers an odour substance, which they store in the inflated parts of their hindlegs.

Wasp flowers

Wasps. Few wasps feed their young pollen or nectar. Yellow jackets, however, occurring occasionally in large numbers and visiting flowers for nectar for their own consumption, may assume local importance as pollinators. These insects prefer brownish-purple flowers with easily accessible nectar, such as those of figwort. The flowers of some Mediterranean and Australian orchids mimic the females of certain wasps (of the families Scoliidae and Ichneumonidae) so successfully that the males of these species attempt copulation and receive the pollen masses on their bodies. In figs, it is not the pollinator's sexual drive that is harnessed by the plant but the instinct to take care of the young; tiny gall wasps (*Blastophaga*) use the diminutive flowers (within their fleshy receptacles) as incubators.

Butterflies and moths. The evolution of moths and butterflies (Lepidoptera) was made possible only by the development of the modern flower, which provides their food. Nearly all species of Lepidoptera have a tongue, or proboscis, especially adapted for sucking. The proboscis is coiled at rest and extended in feeding. Hawkmoths hover while they feed, whereas butterflies alight on the flower. Significantly, some butterflies can taste sugar solutions with their feet. Although moths, in general, are nocturnal and butterflies are diurnal, a colour sense has been demonstrated in representatives of both. Generally, the colour sense in Lepidoptera is similar to that of bees, but swallowtails and certain other butterflies also respond to red colours. Typically, colour and fragrance cooperate in guiding Lepidoptera to flowers, but in some cases there is a strong emphasis on just one attractant; for example, certain hawkmoths can find fragrant honeysuckles hidden from sight.

Moth flowers and butterfly flowers

Typical moth flowers—*e.g.*, jimsonweed, stephanotis, and honeysuckle—are light-coloured, often long and narrow, without landing platforms. The petals are sometimes fringed; the copious nectar is often in a spur. They are open and overwhelmingly fragrant at night. Butterfly flowers—*e.g.*, those of butterfly bush, milkweed, and verbena—are conspicuously coloured, often red, generally smaller than moth flowers, but grouped together in erect, flat-topped inflorescences that provide landing space for the butterflies.

Important pollinating moths are the various species of the genus *Plusia*, sometimes occurring in enormous numbers, and the hummingbird hawkmoth (*Macroglossa*), which is active in daylight. A small moth, *Tegeticula maculata*, presents an interesting case. It is totally dependent on yucca flowers, in whose ovules its larvae develop. Before depositing their eggs, the females pollinate the flowers, following an almost unbelievable pattern of specialized behaviour, which includes preparing a ball of pollen grains and carrying it to the stigma of the plant they are about to use for egg laying.

Wind. Although prevalent in the primitive cycads and in conifers, such as pine and fir, wind pollination (anemophily) in the flowering plants must be considered as a secondary development. It most likely arose when such plants left the tropical rain forest where they originated and faced a more hostile environment, in which the wind weakened the effectiveness of smell as an insect attractant and the lack of pollinating flies and beetles also made itself felt. Lacking in precision, wind pollination is a wasteful process. For example, one male plant of *Mercurialis annua*, a common weed, produces 1,250,000,000 grains of pollen to be dispersed by the wind; a male sorrel plant produces 400,000,000. Although, in general, the concentration of such pollen becomes very low about one-fourth mile from its source, nonetheless in windy areas it can cover considerable distances. Pine pollen, for example, which is naturally equipped with air sacs, can travel up to 500 miles (although the grains may lose their viability in the process). Statistically, this still gives only a slim chance that an individual stigma will be hit by more than one or two pollen grains. Also relevant to the number of pollen grains per stigma is the fact that the dry, glueless, and smooth-surfaced grains are shed singly. Since the number of fertilizing pollen grains is low, the number of ovules in a single flower is low and, as a consequence, so is the number of seeds in each fruit. In hazel, walnut, beech, and oak, for example, there are only two ovules per flower, and, in stinging nettle, elm, birch, sweet gale, and grasses, there is only one. Wind-pollinated flowers are inconspicuous, being devoid of insect attractants and rewards, such as fragrance, showy petals, and nectar. To facilitate exposure of the flowers to the wind, blooming often takes place before the leaves are out in spring, or the flowers may be placed very high on the plant. Inflorescences, flowers, or the stamens themselves move easily in the breeze, shaking out the pollen, or the pollen containers (anthers) burst open in an explosive fashion when the sun hits them, scattering the pollen widely into the air. The stigmas often are long and divided into arms or lobes, so that a large area is available for catching pollen grains. Moreover, in open areas wind-pollinated plants of one species often grow together in dense populations. The chance of self-pollination, high by the very nature of wind pollination, is minimized by the fact that many species are dioecious or (like hazel) have separate male and female flowers on each plant. Familiar flowering plants relying on wind pollination are grasses, rushes, sedges, cattail, sorrel, lamb's quarters, hemp, nettle, plantain, alder, hazel, birch, poplar, and oak. (Tropical oaks, however, may be insect-pollinated.)

Wind flowers

Birds. Because the study of mechanisms of pollination began in Europe, where pollinating birds are rare, their importance is often underestimated. In fact, in the tropics and the southern temperate zones, birds are at least as important as pollinators as insects are, perhaps more so. About a third of the 300 families of flowering plants have at least some members with ornithophilous flowers; *i.e.*, flowers attractive to birds. Conversely, about 2,000 species of birds, belonging to 50 families, visit flowers more

or less regularly to feed on nectar, pollen, and flower-inhabiting insects or spiders. Special adaptations to this way of life, in the form of slender, sometimes curved, beaks and tongues provided with brushes or shaped into tubes, are found in 1,615 species of eight families: hummingbirds, sunbirds, honey eaters, brush-tongued parrots, white-eyes, flower-peckers, honeycreepers (or sugarbirds), and Hawaiian honeycreepers. Generally, the sense of smell in birds is poorly developed and not used in their quest for food; instead, they rely on their powerful vision and their colour sense, which resembles that of man (ultraviolet not being seen as a colour, whereas red is). Furthermore, the sensitivity of the bird's eye is greatest in the middle and red part of the spectrum. This is sometimes ascribed to the presence in the retina of orange-red drops of oil, which together may act as a light filter.

Bird flowers

Although other explanations have been forwarded, the special red sensitivity of the bird eye is usually thought to be the reason why so many bird-pollinated flowers are of a uniform, pure red colour. Combinations of complementary colors, such as orange and blue, or green and red, also are found, as are white flowers. As might be expected, bird flowers generally lack smell and are open in the daytime; they are bigger than most insect flowers and have a wider floral tube. Bird flowers also are sturdily constructed as a protection against the probing bill of the visitors, with the ovules kept out of harm's way in an inferior ovary beneath the floral chamber or placed at the end of a special stalk or behind a screen formed by the fused bases of the stamens. The latter, often so strong as to resemble metal wire, are usually numerous, brightly coloured, and protruding, so that they touch a visiting bird on the breast or head as it feeds. The pollen grains often stick together in clumps or chains, with the result that a single visit may result in the fertilization of hundreds of ovules. In America, where hummingbirds usually suck the nectar of flowers on the wing, ornithophilous flowers (*e.g.*, fuchsias) are often pendant and radially symmetrical, lacking the landing platform of the typical bee flower. In Africa and Asia, bird flowers often are erect and do offer their visitors, which do not hover, either a landing platform or special perches in the form of small twigs near the flower. Pollinating birds are bigger than insects and have a very high rate of metabolism. Although some hummingbirds go into a state resembling hibernation every night, curtailing their metabolism drastically, others keep late hours. Thus, in general, birds need much more nectar per individual than insects do. Accordingly, bird flowers produce nectar copiously—a thimbleful in each flower of the coral tree, for example, and as much as a liqueur glassful in flowers of the spear lily (*Doryanthus*). Plants bearing typical bird flowers are: cardinal flower, fuchsia, red columbine, trumpet vine, hibiscus, strelitzia, and eucalyptus, and many members of the pea, orchid, cactus, and pineapple families.

Mammals. In Madagascar, the mouse lemurs (*Microcebus*), which are only ten centimetres (four inches) long, obtain food from flowers, and in Australia the diminutive marsupial honey possums and pygmy possums also are flower specialists. Certain highly specialized tropical bats, particularly *Macroglossus* and *Glossophaga*, also obtain most or all of their food from flowers. The *Macroglossus* (big-tongued) species of southern Asia and the Pacific are small bats with sharp snouts and long, extensible tongues, which carry special projections (papillae) and sometimes a brushlike tip for picking up a sticky mixture of nectar and pollen. Significantly, they are almost toothless. Colour sense and that sonar sense so prominent in other bats, seem to be lacking. Their eyesight is keen but, since they feed at night, they are probably guided to the flowers principally by their highly developed sense of smell. The bats hook themselves into the petals with their thumb claws and stick their slender heads into the flowers, extracting viscid nectar and protein-rich pollen with their tongues. The plants involved have, in the process of evolution, responded to the bats by producing large (sometimes huge) amounts of these foods. One balsa-tree flower, for example, may contain a full ten grams of nectar, and one flower from a baobab tree has about 2,000 pollen-producing stamens. Some bat flowers also provide succulent petals or special food bodies to their visitors. Another striking adaptation is that the flowers are often placed on the main trunk or the big limbs of a tree (cauliflory); or, borne on thin, ropelike branches, they dangle beneath the crown (flagelliflory). The pagoda shape of the kapok tree serves the same purpose: facilitation of the bat's approach. Characteristics of the flowers themselves include drab colour, large size, sturdiness, bell-shape with wide mouth and, frequently, a powerful rancid or urinelike smell. The giant saguaro cactus and the century plant (*Agave*) are pollinated by bats, although not exclusively, and cup-and-saucer vine (*Cobaea scandens*) is the direct descendant of a bat-pollinated American plant. Calabash, candle tree, and areca palm are other examples of plants with bat-pollinated flowers.

Bat flowers

Water. Although pollen grains can be made to germinate in aqueous sugar solutions, water alone in most cases has a disastrous effect on them. Accordingly, only a very few terrestrial plants, such as the bog asphodel of the Faroes, use rainwater as a means of pollen transport. Even in aquatic plants, water is seldom the true medium of pollen dispersal. Thus, the famous Podostemonaceae, plants that grow only on rocks in rushing water, flower in the dry season when the plants are exposed; pollination occurs with the aid of wind or insects or by selfing. Another aquatic plant, ribbon weed, sends its male and female flowers to the surface separately. There, the former transform themselves into minute sailboats, which are driven by the wind until they collide with the female flowers. In the Canadian waterweed, and also in pondweed (*Potamogeton*) and ditch grass (*Ruppia*), the pollen itself is dispersed on the water's surface; it is, however, still water-repellent. True water dispersal (hydrophily), in which the pollen grains are wet by water, is found only in the freshwater hornworts and the marine eelgrasses.

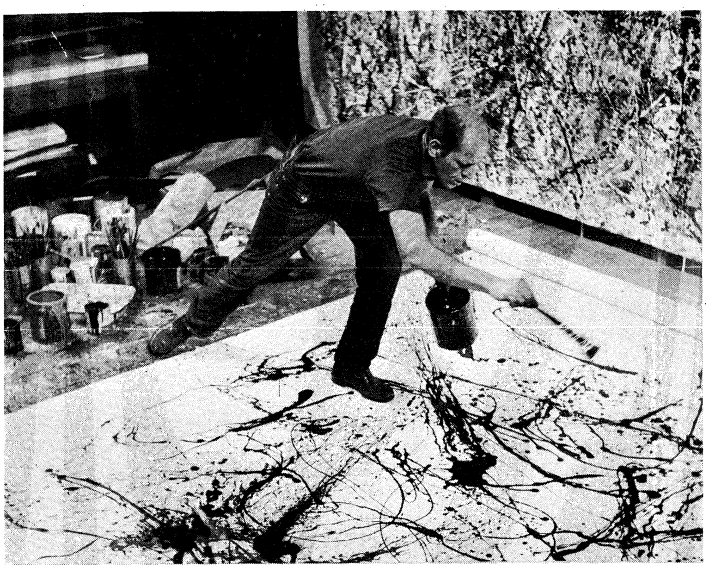
BIBLIOGRAPHY. K. FAEGRI and L. VAN DER PIJL, *The Principles of Pollination Ecology* (1966), a modern, well-balanced work based on extensive experience in both tropic and temperate regions; B.J.D. MEEUSE, *The Story of Pollination* (1961), semipopular, well-illustrated book dealing with general principles of pollination and with the various classes of pollinating agents; M.S. PERCIVAL, *Floral Biology* (1965), a most useful book, in which practical aspects are not ignored; P. JAEGER, *The Wonderful Life of Flowers* (1961), especially valuable for its superb illustrations; L. VAN DER PIJL and C.H. DODSON, *Orchid Flowers: Their Pollination and Evolution* (1967), a discussion of the diverse and highly complex pollination mechanisms in the orchid family; K.A. and V. GRANT, *Hummingbirds and their Flowers* (1968), an account of an important group of pollinators, the hummingbirds.

(B.J.D.M.)

Pollock, Jackson

Jackson Pollock was the leading painter of Abstract Expressionism, an American art movement characterized by free-associative gestures in paint sometimes referred to as action painting. During his lifetime he received widespread publicity and some serious recognition for the radical "drip" or pouring technique he used to create his major works. Among his contemporaries, he was respected for his deeply personal and totally uncompromising commitment to the art of painting. His work and example had enormous influence on them and on most subsequent art movements in America. He is also one of the few American painters to be recognized during his lifetime and after as a peer of 20th-century European masters of modern art.

Paul Jackson Pollock, the fifth and youngest son of Stella May McClure and LeRoy Pollock, was born on January 28, 1912, at Cody, Wyoming. Both parents were of Scottish-Irish extraction (the father's original surname was McCoy before his adoption about 1890 by a family named Pollock), and both were born and raised in Iowa. Jackson dropped his first name, Paul, about the time he went to New York in 1930. The family left Cody 11 months after Jackson's birth. He would know Cody only through family photographs. Over the next 16 years the



Pollock painting in his studio on Long Island, New York, 1950.
Hans Namuth

family lived in California and Arizona, moving nine times during the period. In 1928 it moved to Los Angeles where Pollock enrolled at Manual Arts High School. There he came under the influence of Frederick John de St. Vrain Schwankovsky, a painter and illustrator who was also a member of the Theosophical Society, a mystical religious sect. Schwankovsky gave Pollock some rudimentary training in drawing and painting, introduced him to advanced currents of European modern art, and encouraged an interest in Theosophical literature. At this time Pollock also attended the camp meetings of the former messiah of the Theosophists, Jiddu Krishnamurti, who was a personal friend of Schwankovsky. This experience of occult religion prepared Pollock, who had been raised an agnostic, to later accept the theories of the Swiss psychologist Carl Gustav Jung and the concept of the exploration of unconscious imagery espoused by adherents to Surrealism, a modern literary and artistic movement.

Early influences

In the fall of 1930 he followed his brother Charles, who left home to study art in 1922, to New York, where he enrolled at the Art Students League under his brother's teacher, the regionalist painter Thomas Hart Benton. He studied life drawing, painting, and composition with Benton for the next two and one-half years, leaving the League in the early months of 1933. For the next two years he lived in poverty, first with his brother Charles and, by the fall of 1934, with his brother Sanford. He shared an apartment in Greenwich Village with Sanford and his wife until 1942. In the fall of 1935 he was employed on the WPA Federal Art Project as an easel painter. This provided economic security during the years of the Great Depression and the opportunity to develop his art through early 1943. In 1937 he began psychiatric treatment for alcoholism, and he suffered a nervous breakdown in 1938, during which he was institutionalized for about four months. From 1939 through 1941 he was in treatment with two successive Jungian psychoanalysts, who used Pollock's own drawings in the therapy sessions. In 1943 after the liquidation of the Federal Art Project, he was given a contract by Peggy Guggenheim at her Art of This Century Gallery in New York City, and his first one-man show was held there in November. Thereafter he had one-man shows of new work nearly every year in New York, his work being handled by Peggy Guggenheim through 1947, the Betty Parsons Gallery from 1947 to 1952, and finally by the Sidney Janis Gallery. In 1945 he married the painter Lee Krasner (1908–) and moved to East Hampton on Long Island in New York where he lived until his death in an automobile accident on August 11, 1956.

Pollock's artistic development can be divided into six periods. From his years with Benton to 1938, his work was strongly influenced by the compositional methods and regionalist subject matter of his teacher and by the poetically expressionist vision of the American painter Albert Pinkham Ryder. It consisted mostly of small landscapes and figurative scenes such as "Going West" in which Pollock also utilized motifs derived from photographs of his birthplace at Cody. After his breakdown and hospitalization in 1938 until the time of his first one-man show in 1943, his work was semiabstract and shows the assimilation of motifs from the modern Spanish artists Pablo Picasso and Joan Miró, as well as the Mexican muralist José Clemente Orozco. Jungian symbolism and Surrealist theory also influenced the works of this period. Paintings such as "Bird," "Male and Female," and "Guardians of the Secret" reflect these influences. Very late in 1943, possibly in the early weeks of 1944, Pollock painted his first wall-size work, called "Mural." This painting is his breakthrough into a totally personal style in which Benton's compositional methods and energetic linear invention are fused with Surrealist free-association of motifs or unconscious imagery. His evolution from this point through the 1940s shows a struggle to find a process by which he could translate his entire personality into painting. The figurative character of works such as "The Totem, Lesson I" and "The Blue Unconscious" contrasts with the heavily painted allover design of "Shimmering Substance" and "Eyes in the Heat," indicating the range of imagery and technique he employed during this period. In 1947 he first used the process of pouring or dripping paint onto a flat canvas in stages, often taking weeks of alternating periods of painting and contemplating to finish a canvas. This process permitted him to record the force and scope of his gestures in trajectories of enamel or aluminum paint that he said "veiled the image" that often appeared in his earlier work. The results, in effect, were huge areas covered with complex linear patterns that fused image and form to engulf the vision of the spectator in their scale and intricacy. A whole series of paintings, beginning with "Full Fathom Five" and proceeding through "Summertime" and "Number Ten, 1949" to the mural-size canvasses of 1950 such as "One," "Autumn Rhythm," "Lavender Mist," and the black and white "Number Thirty-two," display the infinite variety of effect and expression this method—poured painting—permitted him. In 1951 and 1952 he painted almost exclusively in black and white, creating works in which his earlier imagery can be seen. The configuration in "Number Twenty-three, 1951" ("Frogman"), for example, relates to "Bird" and also to drawings Pollock did for his second Jungian analyst. Other important paintings of this phase, painted in black enamel on unsized canvas, are "Echo" and "Number Seven, 1952." In 1952 he returned to colour and mural-scale in "Convergence" and "Blue Poles." In 1953 he created his last series of major works. "Portrait and a Dream," "Easter and the Totem," "Ocean Greyness," and "The Deep," among other works, recapitulate many aspects of his former styles and images. Though his production waned and his health deteriorated after 1953, he did produce important paintings such as "White Light" and "Scent" in his last years.

Poured painting or the "drip" technique

From the early 1940s on, Pollock created numerous works on paper, ranging from drawings to mixed-media paintings that parallel the development of his work on canvas. He created very few prints and only experimental sculpture.

As a man Pollock was described by his contemporaries as gentle and contemplative when sober, violent when drunk. These extremes found equilibrium in his art. He was highly intelligent, widely read and, when he chose, incisively articulate. He believed that art derived from the unconscious, saw himself as the essential subject of his painting, and judged his work and that of others on its inherent authenticity of personal expression.

Personality and painting

His critical reception ranged from the discerning art criticism of Clement Greenberg (1909–) in *The Nation* during the 1940s to *Time* magazine's reference to him as "Jack the Dripper" a few months before his death in 1956. Despite occasional attempts in the art press to understand his work seriously, his name became synonymous with extreme artistic caprice, since the novelty of

his "drip" technique overshadowed his obsession with the deeply personal expression that technique permitted. Ironically, he did not profit financially from his fame. He never sold a painting for more than $10,000 in his life and often was hard-pressed for cash. His work was more appreciated abroad. It was seen in Europe, for example, at the Venice Biennales of 1948, 1950, and 1956 and in a one-man show at Paris in 1952. In 1949, the French Abstractionist Georges Mathieu (1921–) stated that he considered Pollock the "greatest living American painter," a judgment still shared by many foreign and American critics.

Art movements following Abstract Expressionism in America, such as Pop art, Op art, and Color-field art, have looked back, with more or less cause, to Pollock's work or the example of his personality as fundamental to their departures. Critical opinion has tended largely to emphasize the formal elements of his art and affinities with such artists as the French Impressionist Claude Monet and Piet Mondrian, the Dutch painter of the de Stijl movement, as well as with modern European art movements, notably Expressionism, Cubism, and Surrealism. The psychologically orientated iconography, so important to Pollock and to a balanced understanding of his contribution to human culture, has generally been ignored.

MAJOR WORKS

Jackson Pollock's works are, for the most part, scattered throughout the United States and abroad in museum and private collections. Major concentrations of his creations are to be found in the collection of the artist's widow, Lee Krasner Pollock and at the Museum of Modern Art, New York. The following are his major paintings. In general, all are painted on canvas except "Number Twenty-nine, 1950," which is on glass.

"Going West" (*c.* 1935; Thomas Hart Benton Collection, Kansas City, Missouri); "The Flame" (1937; Lee Krasner Pollock Collection, New York); "Bird" (*c.* 1941; Lee Krasner Pollock Collection, New York); "Male and Female" (1942; Mrs. H. Gates Lloyd Collection, Haverford, Pennsylvania); "The She-Wolf" (1943; Museum of Modern Art, New York); "Pasiphaë" (1943; Lee Krasner Pollock Collection, New York); "Guardians of the Secret" (1943; San Francisco Museum of Art, San Francisco); "Mural" (dated on canvas 1943, 1944?; The University of Iowa Museum of Art, Iowa City); "The Totem, Lesson I" (1944; Mrs. Emily Walker Collection, West Redding, Connecticut); "The Blue Unconscious" (1946; Mrs. Cecil Blaffer Hudson Collection, Houston, Texas); "Shimmering Substance" (1946; Museum of Modern Art, New York); "Eyes in the Heat" (1946; Peggy Guggenheim Collection, Venice, Italy); "Full Fathom Five" (1947; Museum of Modern Art, New York); "Lucifer" (1947; Mr. and Mrs. Harry B. Anderson Collection, Atherton, California); "Summertime" ("Number Nine, 1948"; American Broadcasting Company Collection, New York); "Number Ten, 1949" (Alfonso A. Ossorio Collection, East Hampton, New York); "One" ("Number Thirty-one, 1950"; Museum of Modern Art, New York); "Autumn Rhythm" ("Number Thirty, 1950"; The Metropolitan Museum of Art, New York); "Lavender Mist" ("Number One, 1950"; Alfonso A. Ossorio and Edward F. Dragon Collection, East Hampton, New York); "Number Twenty-nine, 1950" (National Gallery of Canada, Ottawa); "Number Thirty-two, 1950" (Kunstsammlung Nordrhein-Westfalen, Düsseldorf, Germany); "Mural" (1950; William Rubin Collection, New York); "Number Twenty-three, 1951" ("Frogman"; Martha Jackson Gallery, New York); "Echo" (1951, Museum of Modern Art, New York); "Number Seven, 1952" (Lee Krasner Pollock Collection, New York); "Convergence" ("Number Ten, 1952"; Albright-Knox Art Gallery, Buffalo, New York); "Blue Poles" ("Number Eleven, 1952"; dated on canvas 1953; Ben Heller Collection, New York); "Portrait and a Dream" (1953; Dallas Museum of Fine Arts, Dallas, Texas); "Easter and the Totem" (1953; Lee Krasner Pollock Collection, New York); "Ocean Greyness" (1953; Solomon R. Guggenheim Museum, New York); "Sleeping Effort" (1953; Washington University Gallery of Art, St. Louis, Missouri); "The Deep" (1953; Lee Krasner Pollock Collection, New York); "White Light" (1954; Sidney Janis Collection, New York); "Scent" (1955; Mr. and Mrs. Frederick R. Weisman Collection, Beverly Hills, California).

BIBLIOGRAPHY. The definitive biography to date, along with all known statements by the artist, a detailed history of his exhibitions and their critical reception, and a comprehensive bibliography, is FRANCIS V. O'CONNOR, *Jackson Pollock* (1967). Interpretative monographs have been written by FRANK O'HARA, *Jackson Pollock* (1959); and BRYAN ROBERTSON, *Jackson Pollock* (1960). BERNICE ROSE has written a study of Pollock's drawings and small paintings in *Jackson Pollock: Works on Paper* (1969); C.L. WYSUPH has published the drawings Pollock gave to his first Jungian analyst in *Jackson Pollock: Psychoanalytic Drawings* (1970). All the works cited are heavily illustrated.

Two important articles are: ROBERT GOODNOUGH, "Pollock Paints a Picture," *Art News*, 50:38–41 (1951); and FRANCIS V. O'CONNOR, "The Genesis of Jackson Pollock: 1912 to 1943," *Artforum*, 5:16–23 (1967).

An excellent short colour film of Pollock painting was produced by HANS NAMUTH and PAUL FALKENBERG, with music by MORTON FELDMAN and narration by JACKSON POLLOCK in 1951. It is distributed by Film Images, New York.

(F.V.O'C.)

Pollution Control

Pollution is caused when a change in physical, chemical, or biological conditions in the environment harmfully affects the quality of human life, including effects on other animals and plants, industries, and cultural and aesthetic assets. Most pollution may be characterized as a production excess resulting from manufacturing or growing more than is to be consumed or as a result of discarding products after use. Though pollution is generally attributed to material substances (gases and particulate matter from smokestacks, chemicals in water or solid wastes, paper, glass, used automobiles), pollution may also be nonmaterial, such as an excess of noise and light. Pollution in all its forms is a more serious problem in the heavily industrialized areas of the world than it is in predominantly agricultural regions. Thus, this article is primarily concerned with the control of pollution in North America, Europe, and Japan.

Pollution control consists of the legal, institutional, scientific, and technological arrangements established to avoid or mitigate such excesses in the environment. It can be accomplished by containing pollutants at the source, by devising new technologies of manufacture that eliminate or reduce pollutants, and by reusing (recycling) materials and commodities through reprocessing and resource recovery.

Except for highly toxic substances or for substances concentrated by nature (accumulation in groundwater or concentration in the food chain), pollution control does not mean the elimination of all substances put into air, water, or onto the land. It is notable that, even in the proper uses of water and air, impurities are added, though only in amounts that permit renewal of purity.

Even before man, nature put great quantities of material into the environment and still does (volcanoes, dust storms). Solid material in air, such as salt particles or dust, are essential as the nuclei for raindrops. Yet, when cities add massively to the amount of these nuclei, the excess may diminish rainfall because the drops that form are too small to fall as rain. Because of the multiple uses of air, water, and land, the tolerable amount or level of pollution varies. A ship canal may be allowed a higher level of pollution than recreational water.

Some kinds of pollution, notably radioactive fallout and certain toxic substances from smokestacks, may be widespread and even global. But pollution usually occurs where people and their activities are concentrated. Generally, if all pollutants were distributed equally over the Earth, they would not exceed tolerable limits. Because of his tendency to congregate in huge cities, man himself may be considered a pollutant inasmuch as an excess of people in one place is often the cause of deteriorating human conditions. Thus, one effective way of pollution control is dispersal of its sources.

Interaction of three major types of pollution

Although pollution and pollution control are usually treated in three natural categories—air pollution, water pollution, and land pollution—these three elements of the environment also interact. Many failures in pollution control have resulted from inappropriate division of control responsibilities. Junk piles and dumps pollute the land, but municipalities that burn their solid wastes using incomplete combustion also contribute to air pollution. Pollutants in air or on land, washed or leached out by rain, contribute to water pollution; and substances in wa-

ter settle out in sediments and end up as pollution of the land. What is tolerable depends on when and where the pollution occurs and involves a careful assessment of risks and benefits. DDT, for example, may well be banned or used most sparingly in much of the world, but in Ceylon a worse pollutant is an excess of malaria-carrying mosquitoes; there, DDT may be a proper pollution-control technique.

History of pollution-control problems. Pollution problems had their beginnings with the growth of cities and the massing of people and their related water, industrial, and disposal needs. There are numerous references in ancient history to pollution. Excavations at the site of Troy indicated that inhabitants dropped their refuse on the floor and let it accumulate until the level rose so high that doors would not open, whereupon the doors were repositioned. Ancient Athenians improved somewhat on these sanitation habits, disposing of refuse in dumps at the borders of the city. The Romans, too, left masses of refuse, garbage, and even human corpses in open trenches on the city's outskirts, which historians say contributed to the typhoid, typhus, cholera, and malaria outbreaks that plagued the city.

As urbanization increased, so did pollution. Medieval cities passed many ordinances against refuse dumping in streets and canals, though without much effect. The first smoke abatement law was passed in England in 1273, and by 1306 a man was executed for burning coal in London. In 1388 Parliament forbade disposition of filth and garbage in ditches, rivers, and waters, but the act was no more effective than the edict against burning coal. Pollution problems continued through the Middle Ages, compounded by the practice of throwing waste out of windows onto the streets below.

But it was in the early years of the Industrial Revolution, with the massing of industries in urban areas, that pollution began to produce its staggering environmental effects. A single factory chimney or a single blast furnace might easily have its effluents absorbed by the surrounding landscape, but 20 in a narrow area polluted the air or water beyond technological remedies of the time. Noise levels began rising, and a contemporary account of mid-19th-century Birmingham tells of "hammerings incessantly on the anvil" and of "unending clangs of engines, flame rustles, water hisses, steam roars." More dangerous were the air pollutants, such as chlorine, ammonia, carbon monoxide, and methane, that increased the incidence of bronchitis and pneumonia. Among adults and children living in high-density, highly polluted urban areas in England in the 19th century, the disease and death rates were considerably higher than among those living in low-density and rural areas.

Industrialization contributed to stream deterioration, with wastes thrown into the water. At about the same time, the water-carriage system of sewage disposal was introduced. Large amounts of untreated municipal and industrial wastes began to find their way into streams. By the mid-19th century, water pollution was a serious health problem, particularly in densely populated areas. Odours arising from the Thames River in the late 1850s are reported to have made life in London almost intolerable. In cities elsewhere, conditions were not much better; the Chicago Sanitary and Ship Canal, known as Bubbly Creek, was said to have had in places a scum so thick people could walk on it. Contaminated water supplies led to repeated cholera and typhoid epidemics: one of the most famous was London's Broad Street Pump epidemic in 1854, caused by direct leakage from privies into the hand-pumped well that provided the neighbourhood water supply.

Problems of pollution became steadily more widespread in the 20th century. The automobile, a fledgling in transportation at the turn of the century, became the major air polluter, first in the United States, then in Europe and Japan. Technological advances in industry and agriculture produced thousands of new chemicals, including fertilizers, insecticides, and herbicides, many of which are thrown back or naturally run off into rivers causing eutrophication, a process by which an excess of nutrients stimulates the growth of algae and other plant life to such an extent that they choke waterways.

Air-pollution episodes began to occur. The first recorded disaster was in the heavily industrialized section of the Meuse Valley in Belgium, where, in December 1930, hundreds of people became ill and 60 died during a four-day fog. In the United States during a four-day period in October 1948, the small mill town of Donora, Pennsylvania, was fog-filled with pollutants that made almost half of its 14,000 inhabitants sick. In the 1950s and 1960s both London and New York had air-pollution episodes that resulted in excess deaths and illnesses, chiefly among persons with known heart and respiratory diseases.

Air-pollution disasters

AIR POLLUTION

Pollutants and their sources. Five major classes of pollutants are discharged into the air: carbon monoxide, particulates, sulfur oxides, hydrocarbons, and nitrogen oxides. In addition to the automobile, other major sources are from combustion of fuel in electric-power plants, industrial processes, and space heating, from the combustion in solid-waste disposal, coal-waste fires, and agricultural burning.

Carbon monoxide (CO) is a colourless, odourless, poisonous gas, slightly lighter than air. The only significant source of carbon monoxide is the combustion process when the carbon in fuel is burned incompletely, with the carbon partly oxidized to carbon monoxide instead of fully to carbon dioxide. The largest single source of carbon monoxide is the internal-combustion engine.

Particulates include particles of solid or liquid substances in a number of sizes ranging from those that are visible as soot and smoke to particles too small to detect except under an electron microscope. Particulates are produced primarily by stationary fuel combustion and industrial processes; forest fires account for a small percentage.

Sulfur oxides are acrid, corrosive, poisonous gases produced when fuel containing sulfur is burned. Most of the sulfur oxides come from utility and industrial plants that burn coal and oil, both of which contain sulfur as an impurity. The burning of coal produces about 60 percent of the sulfur oxides annually emitted. Oil burning and industrial processes that use sulfur also are major producers of sulfur oxides, usually concentrated in urban areas.

Hydrocarbons, like carbon monoxide, represent unburned and wasted fuel. Unlike carbon monoxide, gaseous hydrocarbons at concentrations normally found in the atmosphere are not toxic; they are a major pollutant, however, because of their role in forming photochemical smog (see below). Sources of airborne hydrocarbons include incineration, evaporation of industrial solvents, and combustion of coal and wood, but the leading contributor is gasoline, emitted by evaporation and by the internal-combustion engine. Other hydrocarbon emission comes from miscellaneous burning and industrial processes. About 60 percent is produced in urban areas, with their high concentrations of automobiles.

Nitrogen oxides are produced when fuel is burned at very high temperatures. The major sources of this pollutant are transportation vehicles and the combustion process as used in electric-power plants and industrial boilers (which account for about half of the nitrogen oxide emissions). In combustion, the ordinarily inert nitrogen in the air combines with oxygen at high temperatures and tends to stay combined if the exhaust gases are cooled too quickly. Nitrogen oxides combine with hydrocarbons to form a complex variety of secondary pollutants called photochemical oxidants. These oxidants, together with solid and liquid particles in the air, make up what is commonly known as smog. Photochemical-oxidant pollutants include ozone, nitrogen dioxide, peroxyacyl, aldehydes, and acrolein.

Another common air pollutant in urban areas is lead, which comes from manufacturing sources, pesticides, combustion of coal, refuse incineration, and leaded gasoline, which, when burned in automotive engines, is the major source of airborne lead in urban areas. Apart from

the pollutants themselves, certain natural factors affect air pollution. The ultraviolet rays in sunlight combine with hydrocarbons and nitrogen oxides to form photochemical smog, which can result in conditions of atmosphere stagnation during temperature inversions. Such phenomena occur when temperature increases from ground level up to several hundred feet within the troposphere (lower atmosphere); they are called inversions because temperature usually decreases with altitude. Inversions can impede the rise of the air from below, and, if the lower air contains pollutants, the inversion acts as a lid to seal them below. If no significant lateral movement of air occurs, the stage is set for an acute air-pollution episode.

Effect of temperature inversions

Effects on man, plants, and materials. Effects of air pollutants on man are varied. When carbon monoxide is inhaled, it displaces the oxygen in the blood and reduces the amount of oxygen carried to the body tissues. At levels commonly found in city air, mental performances can be dulled and reactions of even the healthiest persons can be slowed, making them more prone to accidents. The gas is believed to impose an extra burden on those already suffering from anemia, diseases of the heart and blood vessels, chronic lung conditions, and overactive thyroid.

Sulfur oxides can cause temporary and permanent injury to the respiratory system, irritating the upper respiratory tract and lung tissue. Photochemical oxidants result in eye irritations, and studies indicate that nitrogen oxides are harmful to human health, specifically increasing children's susceptibility to flu.

Many other pollutants are a growing public health worry even though they may not constitute immediate and direct threats. Studies suggest that lead levels found in the bodies of urban people may interfere with their ability to produce blood. It is being debated whether the lead levels in the environment are of themselves dangerous to health or whether the major danger from them is an increased likelihood of further concentrating lead in some food chains, leading ultimately to toxic doses for man or other organisms.

Air pollution has inflicted widespread damage on plant life, buildings, and materials. Smog in the Los Angeles area contributed to the slow decline of citrus groves south of the city. Sulfur oxides released into the air by phosphate-fertilizer processing in Florida have blighted large numbers of pines and citrus orchards there. Air pollutants also damage man's most durable products: steel corrodes two to four times faster in urban industrial centres than in rural areas where much less sulfur-bearing coal and oil are burned. Sulfur oxides are also accelerating the erosion of statuary and buildings throughout the world; particulates not only settle as dirt but also speed up the corrosive action of other pollutants. Ozone accelerates the cracking of rubber.

Corrosion of steel and erosion of statuary and buildings

Techniques for combatting air pollution. Technological capabilities and control devices are available to abate many air-pollution problems, although some have not become economically feasible. Local air quality standards have long helped alleviate the pollution problem, and in the 1960s and 1970s the trend was toward regional and national enactments.

Automobiles. Specific legislative control of automobile pollution began in California, where devices to control emissions were required on domestic cars starting with the 1961 model year. National standards were enacted in the United States in 1968, requiring complete control of crankcase hydrocarbons and roughly a 50 percent control of exhaust hydrocarbons and carbon monoxide. Standards were later tightened, with the most stringent requiring a 90 percent reduction from 1968 levels of carbon monoxide and hydrocarbon emissions by 1975 and 90 percent reduction of nitrogen oxide by 1976.

By the early 1970s, similar measures had been taken in several other parts of the world, notably by the European Economic Community, which issued directives requiring a harmonization of the national legislation of the member nations. Restrictions of this type can bring a rapid decline in vehicle emissions, but, ultimately, levels will begin to rise again because of the constantly increasing number of automobiles on the world's streets and highways. The technological controls necessary in order for automobiles to comply with emission standards are varied. Emission of hydrocarbons from the crankcase was largely eliminated by ventilation systems that recycle crankcase ventilation air and blow-by gases to the engine intake instead of venting them to the atmosphere. For controlling evaporation of fuel, techniques have been developed to recycle the vapours to the engine. Evaporative emissions could also be decreased by changing the properties of gasoline.

The automobile exhaust problem

To meet the early exhaust-emission standards for hydrocarbons and carbon monoxide, auto manufacturers developed ways to induce further oxidation, and they redesigned cylinders and adjusted the fuel–air ratio and other variables to reduce the amounts of these pollutants in the exhaust. To meet more stringent standards, new techniques are being developed to give even higher degrees of oxidation. Methods of markedly reducing emission of nitrogen oxides in automobile exhaust are still in the development stage.

Motor-vehicle pollution is the product of a complex combustion system of engines, fuels, and fuel additives. Effective control cannot be achieved solely by redesigning engines and installing control devices, but it also will depend on improvements in fuel and fuel additives. The key additive in air-pollution control is tetraethyl lead in gasoline, which accounts for a significant portion of particulate pollution from cars. Major gasoline companies in the early 1970s introduced low-leaded or lead-free gasoline, the manufacturing cost of which was higher than the leaded variety.

Alternatives to the internal-combustion engine have also received attention. The most promising low-pollutant power sources for 1975–85 are the reciprocating steam engine and the gas turbine. Gas turbines are well along in development for vehicular use, but initially they will be economically feasible only in heavy-duty vehicles; research in developing steam engines, on the other hand, is aimed at mass-produced automobiles. As increased electric power becomes available, electrically powered automobiles may become attractive. Electric vehicles today perform a variety of special-purpose, short-range tasks (golf carts, internal-use and short-range trucks). More effective devices are needed for storing and converting electrical energy before such cars can be mass-produced. Another alternative is the hybrid car, with electric power for in-town driving and an internal-combustion engine for long-distance travel (see also AUTOMOBILE).

Electric-power plants. A consideration of electric power for automobiles brings up the question of pollutants emitted in power production. Electric-power plants fired by coal or oil emit sulfur oxides, nitrogen oxides, and particulates. In industrialized countries such plants account for up to 75 percent of the total of sulfur oxides, and, since the electric-power industry is rapidly proliferating, the potential increase of sulfur-oxide emissions is tremendous. Nuclear-power plants, which do not emit sulfur-oxide pollutants, produced only a modest amount of the world's electrical power in the 1970s. Thus, despite ambitious development programs in many countries, nuclear-power production will contribute little to the alleviation of the problem of pollution by sulfur oxides in the remainder of the century. The greatest source of energy will continue to be the burning of coal and oil. The amount of coal burned for power by the year 2000 will be several times greater than in the early 1970s, with total fossil fuel usage tripling in the same period.

Controlling sulfur oxide

A number of alternatives are available to control sulfur-oxide pollution; restrictive standards have been widely adopted. In Britain a construction program of nuclear plants brought their energy output to 6 percent of all electric-power production by 1972. That program, however, did not actually reduce sulfur-oxide pollution in the country, but it undoubtedly slowed its increase. The British had notable success however, in reducing such pollution in large cities and industrial areas by the Clean Air Act of 1956. That law gave local authorities the right to

set up "smokeless zones," which, in effect, banned the use of smoky fuels in the areas and forced a shift to oil, gas-firing, or nuclear plants. In Sweden and the United States, air quality standards that affect electric-power plants have been set on a national level for sulfur oxides, particulates, and nitrogen oxides.

A number of technical problems stand in the way, however. In many existing power plants low-sulfur coal cannot be burned without operational difficulties or incurring high capital costs for furnace modifications. Sulfur can be removed from coal before burning, but the procedure is costly. Content can be cut in half by pulverizing the coal to the consistency of talcum powder and removing the pyrites (sulfur compounds) or by one-third by washing the coal and removing noncarbonaceous material. But, even with as much as 70 percent of the sulfur removed, the final coal product might still be classified as a high-sulfur fuel.

Some oils are also low-sulfur fuels with less than 1 percent sulfur, but they are not always within economic transport range of oil-using industry. Technology available for desulfurizing fuel oil includes removing some of the sulfur at the refinery, diluting high-sulfur oil with low-sulfur oil, or both. The cost of reducing sulfur content from 2.6 percent to an acceptable 0.5 percent adds 35 percent to the oil's cost.

Natural gas, which is very low in sulfur content, is relatively scarce and is used primarily for nonpower purposes, such as space heating and in industrial processes.

Several methods of removing sulfur from stack gases were under study in Europe and the United States in the early 1970s. In one promising method pulverized limestone or dolomite is added to the boiler charge, creating oxides that react with the sulfur oxides to form solid sulfite and sulfate particles that can be removed by electrostatic precipitation. In another process, catalytic conversion, the sulfur dioxide is converted to sulfur trioxide, which combines with water in the stack gas to form a sulfuric acid mist that can be trapped and eliminated. Yet another method is to produce sulfuric acid, which can readily be removed from the stack gas by the addition of an activated char, a carbonaceous material. A similar process, using manganese dioxide, produces ammonium sulfate, which again can be removed. None of these processes, however, had reached the stage of large-scale use in the early 1970s.

Also emitted from the stack are particulates and fly ash; electrostatic precipitators and mechanical collectors can control such pollutants with 99 percent efficiency. The disposal problem that large quantities of fly ash poses is being attacked by researchers who are trying to convert the fly ash into a useful product, such as an additive to cement.

Work on abatement of nitrogen oxides emitted by power plants has been limited, and practical technology for achieving significant control is not yet available. Emissions can be reduced 40–50 percent in some furnaces by two-stage combustion, which limits peak combustion temperatures and thus the amount of nitrogen oxides formed. Efficiency is little affected. Only a few power boilers have been designed or modified to use two-stage combustion.

Industrial and other sources of air pollution. Industrial sources generate a range of air pollutants specific to the process involved. Integrated steel mills emit particles, smoke, carbon monoxide, and fluorides; cement plants produce particles and sulfur compounds; refineries emit sulfur compounds, hydrocarbons, smoke particles, and odours. The major industrial air polluters are inorganic and organic chemical plants, iron and steel mills, petroleum refineries, pulp and paper mills, and nonferrous metal smelters.

The chief industrial polluters

A wide variety of air-pollution-control technology is available to industry: filters, vapour collectors, electrical precipitators, cyclones (centrifugal devices for separating solid material from gases), scrubbers, absorbers, and combustion equipment such as afterburners and catalytic units.

In many industrial operations, however, air pollution cannot be controlled with current technology: nitrogen oxides formed in combustion processes and emissions from coke ovens in the steel industry cannot be adequately controlled. Odour abatement is an unsolved problem in kraft pulp mills. It does not seem likely that sulfur oxides formed by industrial combustion of fossil fuels can be controlled to any extent by stack gas treatment, although some limited successes have been reported. The solutions are substitutions of low-sulfur fuels, development of new combustion processes with reduced sulfur-oxide emissions, or removal of sulfur from fuels before combustion.

A growing contributor to air pollution is domestic and commercial space heating, the effect of which on urban air quality is proportionally greater than its aggregate weight, because the emissions enter the air in crowded areas at relatively low heights above the ground. Space heating causes a particularly difficult abatement problem because of the large numbers of its relatively small sources. Solutions include reducing the sulfur-oxide emissions by using a low-sulfur fuel, reducing particle emissions by not using high-ash-content coals, and upgrading the efficiency of combustion equipment.

Refuse disposal also contributes to air pollution. Few of the world's disposal sites on land are sanitary landfills (defined as having daily cover, no open burning, and no water-pollution problems); and few municipal incinerators have adequate air-pollution-control devices. Air quality requirements beginning to appear on local levels make control technologies for incinerators mandatory. Devices most commonly used are settling chambers and wetted baffles, cyclone collectors, direct-impaction scrubbers, electrostatic precipitators, and bag filters, all of which to some degree separate out and collect emissions of fly ash and other solid particles; only the last three have operating efficiencies of over 95 percent.

Requirements of sanitary landfills

WATER POLLUTION

Water, an almost universal solvent, even in its natural state contains dissolved materials, carries particles, and is the home of living organisms. When water is used for domestic and industrial purposes even more materials, of a wide range, are added to it. Although the composition of wastes in water varies widely, its main inorganic constituents are sodium, potassium, ammonium, calcium, magnesium, chloride, nitrate, nitrite, bicarbonate, sulfate, and phosphate. Organic compounds are numerous and not well-known, with the exception of pesticides, detergents, phenolic substances, and carboxylic acids in streams.

Water pollution by nitrates, toxic substances, bacteria, and viruses affects human health, and taste and odours can be unpleasant, even in water not harmful to human health. More subtle are long-term changes in the aquatic life of a water body: the decline of sport fish and increase in carp, sludge worms, and other life forms tolerant of pollution. Water pollution also speeds eutrophication, with resultant aging and clogging of lakes. In these ways, water pollution degrades the multiple uses of water for human supply, recreation, aquatic life, and other wildlife, agriculture, and industry. Fish kills often result from pollution. In 1969 in the United States, it is estimated that 70 percent of the kills were caused by industrial wastes that consume dissolved oxygen in the water or by excess of pesticides or other poisonous wastes.

Degradable and nondegradable wastes. The complex mixtures in the waters of rivers and lakes include many unknown chemicals. For practical purposes they may be divided into degradable and nondegradable wastes.

Degradable wastes in water are those that are reduced in quantity by natural processes. These include organic wastes and thermal discharges. A great source of organic materials in water is domestic sewage. If these materials do not constitute too heavy a load, they are converted to stable inorganic materials by bacteria and other organisms. The degradation process uses the oxygen in the water and constitutes natural "self-purification." If, however, the waters are too loaded with organic matter, degradation without sufficient free oxygen produces offensive odours, such as hydrogen sulfide and methane.

Nondegradable salts, gases, and particulates

Nondegradable wastes are salts—soluble in water—soluble gases, or particulate matter (tiny, suspended particles). These wastes may include a wide range of toxic metals arising from complex manufacturing processes; cadmium, mercury, and lead are examples. Generally, the effects of these substances on life is not fully understood, and permissible levels are not properly known. There have been incidences, however, of permanent neurological impairment and death among persons eating fish from waters heavily polluted by mercury.

Settleable and suspended solids that are nondegradable represent primarily an industrial problem. Suspended particles—of various sizes—absorb other substances, act as nuclei for bacterial growth, and, by sedimentation, deposit many substances, including insecticide residues such as DDT, in the bottom muds of bodies of water.

Municipal wastes. Municipal wastes come from homes, commercial establishments, and industries. With the widespread shortage of adequate treatment systems for normal sewage flow, the additional inflow of storm water, which collects rapidly in volume too great for treatment plants to handle, intensifies the disposal problem. In the usual combined sewer system, the wastewater cannot be separated from the large volume of rainwater. Therefore, when total flow exceeds the capacity of the system, most of the wastewater is included in the bypass flow that is discharged directly into the outlet river or stream. Separate storm-water systems, or holding reservoirs, are under study, but such solutions for large urban areas are costly.

If municipal wastes produced by industries contain toxic substances, the pollution is combatted by setting standards limiting the maximum concentrations of toxic effluent that may leave the plant. To comply with such standards, an industrial plant must either clean the effluent water before discharge or redesign the industrial process so that certain highly toxic substances are not discharged into the municipal system.

Primary, secondary, and tertiary treatment

Municipal water pollution is combatted under one of three levels of sewage treatment.

Primary treatment is simply the separation of solid materials by screens or by settling in still tanks, sometimes aided by chemical precipitation. Removal of grits and solids is accomplished by screening, grinding, flocculation (causing small particles to combine to form larger particles), and sedimentation. In primary treatment most of the settleable solid materials are removed.

Secondary treatment entails oxidizing and dissolving colloidal organic material by action of bacteria and other organisms with oxygen. This stage is accomplished by trickling filters, aeration, and passage through sludge activated with bacteria.

Advanced, or tertiary, treatment is a term used for all treatments beyond the secondary up to the purification of the water to a point where it is once again potable. At least one city in the world, Windhoek, South West Africa, reclaims one-third of the total municipal water after use and returns it to its drinking-water supply.

After sewage is treated by any of the three methods, the remaining sludge is digested with the help of bacteria and concentrated by drying in beds or in lagoons with filters or in centrifuges or by heat. The final sludge may be incinerated, used for landfill or to condition soil, or may be discharged into the sea (see also SEWAGE SYSTEMS).

Industrial wastes. Water-using industrial plants add pollutants that have a variety of effects. With the steady spread of industrialization, the wastes of industry have been growing several times faster than sewage. Four major world industries account for a large portion of wastewater discharge: paper, chemical, petroleum, and steel. In addition, great amounts of waste heat are discharged into water by power generation (see below).

Because industrial wastes are generated in bulk at the plants, they can more easily be controlled, and much has been done at the source either to change production to minimize waste or to treat wastes. Modern sulfate pulp and paper plants have reduced their wastes to only 7 percent of the volume under older methods. Many of the new complex chemicals are difficult to detect and control, however, and even small amounts can be dangerous. More than 9,000 organic chemicals are now manufactured. Commercial metallic compounds number in the hundreds, and hundreds of new chemicals are marketed each year. Substances such as compounds of mercury and cadmium and polychlorinated biphenyls (PCB) in plastics have known, or potential, hazards to human, animal, and plant life. Levels of tolerance are often not known, but it is certain that, because they accumulate in living tissues and are concentrated in the food chain before its use by man, originally small amounts can lead to concentrated hazard levels. Some chemical substances are regulated by approval prior to their being marketed in pesticide, food additive, and drug products. The release of other chemicals into the environment is subject to controls imposed by effluent standards. Toxic substances that are difficult to detect are not controlled by waste-treatment methods and may only be identified after damage has occurred. Both research and new controls are needed on such substances.

In early 1972, Britain, the Scandinavian countries, and eight other European governments signed the first major international agreement to prevent pollution of the seas by dumping from aircraft and ships. Covering the northwest Atlantic, this agreement bans dumping of many poisonous substances, including pesticides, mercury, cadmium, and persistent plastic waste. At the time of signing, other nations, specifically those of eastern Europe, were expected to join.

Agricultural wastes. While agricultural wastes contribute considerably to water pollution through the runoff from pesticides, herbicides, and fertilizers, they are generally less concentrated and have less aggravating effects than industrial and municipal sewage wastes originating in or near cities.

Where the land area is available, the wastes from farm animals, many times greater in volume in nearly all countries than that of the human population, are best disposed of by field spreading. Lagoons for anaerobic degradation may be constructed near feedlots.

Acid mine drainage. Another serious non-urban water pollutant is the drainage of acids from mines, mostly abandoned coal diggings, into neighbouring streams. The best method of control is at the source by sealing the mine to prevent air and water from entering to react with the substances that form the acids. Apart from prevention, most of the treatment is confined to neutralizing acids with lime at the source.

Waste heat. In addition to using water for flushing and carrying off wastes, industry also uses huge amounts for cooling. The largest such user is the electric-power industry. Steam-generating plants, which draw water from a natural source to cool their condensers, discharge it back at an average of about 13° F (7° C) higher temperature. Nuclear plants require almost 50 percent more water than equivalent fossil-fuel plants. With the growth of the electric-power industry certain to continue, the cooling of power generation plants is clearly becoming a major pollution problem. Such thermal pollution, as it has been termed, accelerates biological and chemical reactions and aggravates the diminution of dissolved oxygen. Water temperature may rise above the level tolerable for the fish and, at the same time, may promote the growth of undesirable algae, adding to the effects of other pollutants: excess nutrients, detergents, fertilizers, and human wastes. Waste heat, therefore, speeds the eutrophication of lakes.

Effects of thermal pollution

Technological innovations include using contained lakes and ponds or cooling towers to recycle the water. Ponds (2,000 acres per 1,000,000-kilowatt plant), either manmade or isolated from natural waters, do not harm natural ecologies. Furthermore, they may be turned to economic use by developing useful warm-water ecologies with fish that do well in warmer water and types of plants and algae that may be harvested for protein animal feed.

In cooling towers, heated water is dripped over a lattice while air enters from the bottom of the tower, passes upward through the lattice, evaporates part of the water, and cools the rest. Natural draft or fans carry the warm moisture-laden air out of the top. Though some of the

water is thus lost through evaporation, most is cooled and collects in a basin to be recirculated to the condenser.

Another type, the dry cooling tower, uses air alone to cool the hot water, which is contained in tubing exposed to the upward flow of air. Dry towers are more expensive, but many have been built in Europe.

Cooling towers, rising 500 feet (150 metres) in the air, are generally considered unsightly. Another drawback is that the heat and moisture added to the air may affect the nearby weather and climate by stimulating cumulus clouds, fog, or ice, although this effect is believed limited. The same amount of heat must be dissipated, whether in air or water, but it may be less damaging to the ecology in air. Only a fraction of the nuclear-power plants planned or operating in the early 1970s had cooling towers or cooling ponds; most drew in natural water and returned it to the source.

LAND AND SOIL POLLUTION

Agricultural wastes and mineral wastes, principally from mining, make up a very large proportion of the total solid wastes in all countries but constitute a proportionately less harmful pollutant because they are spread over wide areas, not concentrated as are most residential and industrial wastes.

Agricultural solid wastes. This largest single source of solid wastes includes animal manures and wastes from slaughterhouses and from all forms of crop harvesting. Food animals, principally cattle, hogs, sheep, and chickens, generate 1,000 times more manure than does the total human population. In the past, nutrients were returned to the soil by the livestock in pastures or on rangeland. Modern feedlots, with increasing numbers of food animals, concentrate these quantities, however. Corrective measures are to spread the wastes on areas where they will help crop production and minimize water pollution.

Sediment from soil erosion is another factor in the deterioration of the land and the adding of turbidity to the waters. Sediment production is an increasing problem from poorly managed forest rangelands and agricultural crops. Mining operations and construction add to the sediment bulk.

Fertilizer salts on the land also contribute to eutrophication of waters, either when more fertilizer is applied to the land than the crops can take up or when heavy rains wash off the salts. While fertilizer is essential to the growing of food, its side effects increase with increase of population. The densely populated Netherlands, for example, in 1965 used 10 times as much nitrogen per acre as the United States. Technical controls include better understanding of the movement of fertilizers in various soils, the innovation of slow-release, nitrogenous fertilizers, and regimes of application to minimize escape of fertilizers from the land.

The pesticide problem

Pesticides (including insecticides, nematocides, herbicides, and fungicides) contain hundreds of major chemicals; their impact on soil and water contamination is great. To feed a rising population, pesticide usage is not likely to decrease. Much work needs to be done to learn more about their persistence, their levels of tolerance in man, animals, and plants, and to understand the changes they undergo in soils. Basic chemical and soil research and studies of biological tolerance levels are especially relevant.

Mineral solid wastes. Mineral solid wastes come mostly from mining of minerals and fossil fuels and the associated milling and processing industries. Strip mining causes a particularly difficult form of pollution in that it not only produces mineral wastes that pollute streams and destroy watersheds but also leaves huge scars on the land. Power shovels that can take more than 200 cubic yards at a bite are used to remove the overburden of earth and expose coal or other valuable minerals beneath.

Efforts to repair the scars on the countryside have been attempted by some companies but are expensive. They include building parks, planting vetch on the spoil banks, and turning these areas into ranches for cattle raising. So far such constructive moves have been only partly successful. West Germany has adopted strict legislative controls for strip-mine operations, requiring advance planning for land renewal of the site.

Municipal solid wastes: residential and industrial. Individual householders and commercial enterprises in cities and communities generate a vast amount of solid wastes, including paper, food, metals, glass, wood, plastic, clothing, rags, rubber, leather, and dirt. Disposal has been largly into open dumps, with a fraction going into sanitary landfills and another fraction burned in incinerators. The remaining small proportion is flushed, composted, or salvaged.

Open dumps burn, contributing to air pollution; breed rats, a severe health hazard; pollute water sources; and are offensive eyesores. Incinerators are expensive and, if not properly designed, contribute to air pollution. Sanitary landfill is satisfactory if marginal land is filled and turned into useful residential, industrial, or recreation areas.

Technological means for meeting the municipal waste problem include improved collection, improved incineration, using the heat for power generation, and improvement of sanitary landfill. While all of these measures will be needed, it should be noted that none recovers the resources of the solid wastes.

Resource recovery by recycling

There is further hope for waste management in resource recovery by recycling. Such a comprehensive approach envisages the separation of the wastes so that some may be either converted to fuel or burned directly as fuel, some may be transformed by chemical processing into new by-products, while others may be reused or recycled in their original form. Paper, a major constituent of most municipal waste, is an excellent candidate for recycling; Japan recovers and reuses half of all the paper it consumes. In the early 1970s, progress was made in collecting and reusing glass products, principally beverage bottles.

Recycling, recovery, and reuse. Recycling, recovery, and reuse constitute what has been called the secondary materials industry, growing rapidly in the 1970s. The metals that are reused to the greatest extent are iron and steel, followed by copper, zinc, and aluminum. Reclaimed rubber constitutes an important fraction of the total consumption. The growing reuse-and-recovery industry is hampered, however, by the lack of economy of scale. Large photographic companies, for example, reclaim silver from the film they develop; but for small processors it is uneconomical to do so, and large amounts of silver are therefore lost.

Recycling plans and programs have been advanced in several countries, with both private and governmental support. Three basic steps are involved: collection, separation, and processing. The significance of recycling in the future from both the pollution-control and conservation viewpoints has been pointed out as one of the advantages of the "new city" concept. Such experimental cities, built from scratch, can plan all types of pollution control far better than can old, established cities, and extensive recycling can be economically realized.

Junked automobiles. The disastrous aesthetic impact of automobile graveyards makes this component of solid waste conspicuous in the public eye, but most automobiles were being scrap-processed for recycling in the early 1970s, with technology improving. Bundles of auto hulks compressed by the older methods contain up to 90 percent iron and 0.5 percent copper, the worst contaminant in the reuse of the iron. New types of shredders separate the ferrous from the nonferrous materials and increase the iron yields to 97 percent, reducing the copper to 0.2 percent or less. It is believed that the junk-car problem will be solved by reuse soon because of technological changes in steelmaking techniques and improvement of scrapping and shredding efforts.

Tires. A considerable proportion of the tires discarded annually can be reclaimed, but most are unfit for reclamation and constitute a disposal problem. They may be incorporated into asphalt for paving, but further work is required to identify and arrange for economical reuse of scrap rubber.

OTHER KINDS OF POLLUTION

Noise. Noise, unwanted sound, is one form of environmental contamination that disappears when the source is turned off. Yet, it is a form of environmental degradation and has implications for health that may be as serious as air or water pollution. It can change man's physiological state by speeding up pulse and respiratory rates, and it can impair hearing either permanently or temporarily; millions of industrial workers are threatened with hearing damage. There is medical evidence that noises can cause heart attacks in individuals with existing cardiac injury and that continued exposure to loud noises could cause such chronic effects as hypertension or ulcers. Noise is measured by intensity in decibels in a range from barely audible to shatteringly loud (see Table).

Weighted Sound Levels and Human Response

sound source	decibels*	response criteria
	150	
carrier deck jet operation	140	
		painfully loud
	130	limit amplified speech
jet takeoff (200 ft)	120	
discotheque		
auto horn (3 ft)		maximum vocal effort
riveting machine	110	
jet takeoff (2,000 ft)		
shout (0.5 ft)	100	
New York City subway station		very annoying
heavy truck (50 ft)	90	hearing damage (8 hours)
pneumatic drill (50 ft)		
	80	annoying
freight train (50 ft)		
freeway traffic (50 ft)	70	telephone use difficult
		intrusive
air-conditioning unit (20 ft)	60	
light auto traffic (50 ft)		
	50	quiet
living room		
bedroom	40	
library		
soft whisper (15 ft)	30	very quiet
broadcasting studio	20	
	10	just audible
	0	threshold of hearing

*Weighted sound levels taken with a sound-level meter and expressed as decibels on the scale approximating the frequency response of the human ear.
Source: Department of Transportation.

The rapid growth of air transportation after World War II and the development of jets created a major noise problem in and around airports. Millions of people live near enough to airports to be subjected to intense aircraft noise. In spite of acoustical advances in engine design, modified flight procedures, and design improvement, the problem of aircraft noise is becoming more serious because every year there are more and larger aircraft. Aircraft noise is abated by delaying flap retraction, holding speeds relatively low while making turns and while climbing, directing aircraft traffic away from communities, and soundproofing and air-conditioning residences near airports. But the most effective abatement is to control the source of the noise, the engine. With the switch from turbojet to turbofan engines, engine noise has been cut by new designs that increase mass flow and reduce velocity through the engine. Engineers are now working to reduce fan noise.

The sonic boom

One noise that cannot be eliminated is the sonic boom created by planes flying faster than the speed of sound. Sonic boom is a sudden pressure disturbance or shock wave in the air. The exposure zone ranges from eight to 50 miles (13 to 80 kilometres) on each side of the centre line of the flight path of a plane travelling at supersonic speeds at an altitude of 65,000 feet (20,000 metres). The French–British and the Soviet commercial supersonic planes promised to bring the problem to the fore in the mid-1970s. Supersonic military planes, though kept away from urban areas, have caused booms that have shattered glass, twisted metal window frames, and caused rockslides.

A considerable amount of legislation has been enacted to limit noise pollution. Most cities of the world have codes to govern noise levels in industrial, commercial, and residential areas. Curfew provisions often prohibit particular noisemaking activities or operations during certain days or hours.

Technological methods for reducing industrial and city noises include machines that can cut through asphalt and concrete at equal or greater speeds and much lower noise levels than conventional jackhammers, a quiet air compressor, layers of polymeric materials and metal in railroad-car wheels and saw blades that reduce their screeching noise, sound absorption materials on plant ceilings and walls, and sound-absorbing partitions around noise equipment. For high-level-noise work situations, personal protective devices include plastic-foam earmuffs and earplugs made of rubber, plastic, or disposable materials such as wax-impregnated cotton.

Radiation. Low-level, man-made radiation comes from X-rays, radioactive materials, and electronic devices, including television sets. Medical uses constitute 94 percent of all man-made radiation and 30 percent of all radiation to which the average individual is exposed.

High doses of radiation have been shown to be related to incidence of leukemia and other cancers, and low-level radiation may cause genetic damage and increase natural mutation. Radiation levels in the environment have declined markedly since the early 1960s because of an international agreement banning nuclear-weapons tests in the atmosphere, outer space, and underwater.

Nearly all of the radionuclides escaping into the environment do so at the site of nuclear-fuel reprocessing plants. The significant effluents discharged are krypton-85 and tritium, which move into the air and water system but which are not expected to increase the world population's radiation exposure significantly.

High-level and low-level radiation waste

Radioactive waste is divided into low-level and high-level wastes. Materials with low levels of radioactivity are the gaseous, liquid, and solid wastes from reactors and other nuclear facilities. These are packaged and buried in licensed burying grounds. High-level wastes, the by-products of the reprocessing of used fuel elements from nuclear reactors, will become an increasing problem as their volume increases with the world's expanded uses of radiation sources. Technology has been developed to reduce the high-level liquid wastes to solid form, thereby reducing the final volume to one-tenth that of the original liquid material. Nevertheless, dumping in the sea, storing in salt mines, and other suggested practices all carry with them hazards because of the radioactive materials' very long half-lives.

Other nonmaterial pollution. Several types of nonmaterial pollution annoy or offend the aesthetic sense rather than do physical or physiological harm. In the category

of what has been called visual pollution are such eyesores as billboards, neon signs, electric-power transmission lines, and garish roadside strip developments. Solutions include underground transmission lines and combining electricity, gas, oil, and rail traffic along the same corridors.

Light pollution, the glare of streetlights and advertising signs from the city and its environs, among other things, seriously hampers the observations of astronomers. Light pollution is closely related to air pollution in that the ultraviolet light and general overall glare are scattered by dust and other particles in the air. To the naked eye, nights are still black, yet, to the sensitive electronic eyes, the quality of darkness is strained. Astronomers' telescopes are sensitive enough to pick up stars 100,000,000,000 miles away, but flashes from a neon sign from a roadside 20 miles away can easily block out light from these distant stars. Compounding the problem, many of the celestial wavelengths are nearly identical with the ultraviolet rays emitted by mercury lamps along nearby streets and highways. Astronomers have pressed for regulation of future outdoor lighting.

GLOBAL ASPECTS OF POLLUTION

Pollution of the atmosphere. The Earth's atmosphere contains 21 percent oxygen and 0.032 percent carbon dioxide by volume, a ratio of 650 to 1. Man's activities, including burning of fossil fuels, land clearing, and making cement from limestone, are increasing the carbon dioxide content of air. In spite of its relatively small concentration in the air, carbon dioxide plays an important role in determining the temperature of the planet: the temperature rises when the amount of carbon dioxide in the atmosphere rises, for the gas absorbs sunlight and infrared radiation, preventing a portion of the heat from radiating back into space.

Worldwide temperature increase

The possibility of a worldwide increase in temperature from the addition to the atmosphere of carbon dioxide from the burning of fossil fuels was first posed in 1899. The problem has since become magnified, with predictions for larger yearly increases in rates of carbon-dioxide emissions until nuclear power provides more of the needed energy. Assuming an annual worldwide increase of 5 percent in fossil-fuel consumption, it is estimated that by 2020 the increase in carbon dioxide in the atmosphere would increase world temperatures by about 4° F (2° C) from the level of 1860, when mankind began to burn fossil fuels in appreciable quantities.

Techniques to reduce the carbon-dioxide content of the air or speed up its absorption by ocean waters are not economically feasible. A substantial increase in forest areas would reduce the carbon-dioxide content, but possibilities for this solution are limited. Means of preserving the status quo of carbon-dioxide levels in the atmosphere include using new power sources other than fossil fuels, such as nuclear energy, direct solar energy, and geothermal steam.

The upper atmosphere is also being threatened by man's technological activities. Supersonic planes could increase water vapour in the stratosphere, and nitric-oxide emissions from such planes could possibly reduce layers of ozone, the gas that absorbs the sun's ultraviolet rays. Other particles from supersonic aircraft could cause climatic changes, increasing the temperature of the stratosphere with possible changes in the Earth's surface temperatures.

Ocean pollution. The oceans, which cover more than 70 percent of the Earth and are a critical element in maintaining the world's environment, are the final receptacle for many wastes. Most wastes sink to the bottom, but solid wastes, oil, and garbage contain many floatable materials that are carried great distances by currents and winds. Pollution affects marine life directly through toxicity, which kills some plants and animals and causes reproductive failure in others; it also causes oxygen depletion, biostimulation, and economic losses in shellfish, fishing, and recreation industries.

Wastes dumped into the ocean include dredge spoils, industrial wastes, sewage sludge, construction and demolition debris, solid wastes, explosives, chemical munitions, and radioactive wastes. Spoil materials from dredging constitute the largest single ocean pollutant, followed by industrial wastes, such as acids and refinery, pesticide, and paper-mill wastes. Most are dumped from barges; highly toxic wastes are packed in drums.

In most cases, alternatives to ocean dumping exist. Industrial wastes can be treated on land or recycled, and solid wastes may often be used as landfills, for example, to fill in areas left by strip mining. Disposing of dredge spoils on land is not possible because of the enormous volume.

Ocean pollution from oil spills

A prominent source of ocean pollution comes from large oil spills, some of which have killed marine birds and plants, left layers of encrusted oil along shores, and have cost large sums to clean up. Such spills, while concentrated and disastrous, account for less than 20 percent of the large amounts of oil that man introduces directly into the world's waters every year. Most of the chronic pollution originates from the normal tank cleaning at sea of oil-carrying tankers, from normal operations of refineries and petrochemical plants, and by fallout of airborne hydrocarbons emitted by motor vehicles and industries. The danger of large-scale accidents increases, however, with the steadily increasing size of tankers; 300,000-ton ships are already in operation, and plans for 800,000-ton tankers have been projected. A single spill from one of the new large tankers could add 20 percent to the amount of petroleum entering the oceans in a single year.

A number of measures have been taken to reduce chronic oil pollution. Most tanker operators have stopped dumping wastewaters from tank cleaning into the oceans. Methods for combatting large accidental spills include the use of absorbants such as straw, sawdust, and shredded polyurethane foam; gelling agents that can solidify petroleum materials; burning the slicks; and using chemicals and detergents, most of which, however, are to some degree toxic to aquatic life. Some nations have made illegal the grossly negligent or willful distribution of oil from a vessel in navigable waterways, and international agencies have discussed establishing global protection and cleanup procedures.

Other aspects of international pollution. Prevailing currents of air may carry sulfur dioxide or particulates across national boundaries, and a river system often extends along or beyond national borders. The Rhine, rising above Lake Constance, in Switzerland, flows through West Germany and France before it reaches Rotterdam in The Netherlands, and the North Sea; the river is polluted for most of its length. In June 1969, millions of fish died in the Rhine as a result of a pesticide spill from a chemical industry in Frankfurt Am Main that imperilled the drinking-water supply in The Netherlands. The four riparian countries have agreed on methods for cleaning up and protecting the river.

Among other international efforts at pollution control in the early 1970s were the Environmental Committee for the Organisation for Economic Co-operation and Development, which studied international economic implications of domestic environmental quality standards; the Committee of the North Atlantic Treaty Organization on the Challenges of Modern Society, which is studying ocean oil pollution, low-polluting vehicles, and air-pollution monitoring; the Inter-Governmental Maritime Consultative Organization, which is working on conventions of rights and liabilities of oil spills; and the International Joint Commission, which completed a Canadian–U.S. agreement for measures to control Great Lakes pollution by 1975. The first worldwide meeting on environmental quality was the United Nations Conference on the Human Environment, held in Stockholm, in June 1972. The conference's aim was to develop methods to achieve international environmental quality through control, monitoring, resource management, and education. For developing countries, the aim of the conference was to find ways to advance economically with environmentally sound practices.

BIBLIOGRAPHY. The AMERICAN CHEMICAL SOCIETY, *Cleaning Our Environment: The Chemical Basis for Action* (1969);

and the NATIONAL RESEARCH COUNCIL, COMMITTEE ON POLLUTION, *Waste Management and Control* (1966), are basic reports on the technological state of the art of pollution control, covering air, water, and land pollution. For further details on water pollution, see CHARLES E. WARREN, *Biology and Water Pollution Control* (1971); and ALLEN V. KNEESE and BLAIR BOWER, *Managing Water Quality: Economics, Technology, Institutions* (1968). A worldwide analysis of pollution problems may be found in the STUDY OF CRITICAL ENVIRONMENT PROBLEMS, *Man's Impact on the Global Environment* (1970). *Environmental Quality*, the first and second annual reports of the Council on Environmental Quality, Washington, D.C. (1970–71, as well as additional annual reports), give a yearly analysis of pollution problems and pollution control in the United States. Specialized areas of pollution control are discussed in depth in the following: *Ocean Dumping: A National Policy*, Report to the President prepared by the Council on Environmental Quality (1970); *Oil Pollution: Problems and Policies*, Bureau of National Affairs Management Series (1969); and *Noise As a Public Health Hazard*, Proceedings of an American Speech and Hearing Association Conference, Washington, D.C. (1969). References to the most current literature on problems of pollution are given in *Air Pollution Abstracts* (1970–); *Selected Water Resources Abstracts* (1968–); and *Pollution Abstracts* (1970–).

(A.Sp./J.J.St.)

Polo, Marco

Marco Polo, the most outstanding traveller and explorer from the West to journey to Asia during the Middle Ages, was born in Venice in about 1254; he died in the same city in 1324. His importance in European history is due not so much to the length and variety of his itineraries (both of which were indeed considerable) but to a lucky chain of events that made it possible for him to publish a profuse account of things seen, people met, and information gathered during some 25 years (1271–95) in Asia. The civilizations of Europe and of the Far East—walled off from each other by a formidable wedge of Muslim nations—had flourished for centuries in reciprocal ignorance of each other's existence. Though some travellers had bridged the physical gap between the two worlds, their voices somehow failed to reach a wide audience. Marco's account, on the contrary, created a sensation. This had its good and bad effects, however; on the one hand, new vistas opened to the medieval mind, but, on the other hand, much of the news appeared so fantastic that many considered the work a romance rather than an essentially factual work. But time has vindicated Marco and proven him a keen and generally reliable observer; as Western horizons expanded, Marco's influence became greater and greater. His description of Cipango (or "Zipangu," in Marco Polo's nomenclature; now Japan) set a definite goal for Columbus in his journey toward the setting sun in 1492, while his detailed localizations of spices encouraged Western merchants to seek out these areas and break the age-old Arab trading monopoly.

Influence of his book



Polo, title page of the first printed edition of *The Travels of Marco Polo*, 1477.
By courtesy of the Columbia University Libraries, New York

YEARS OF TRAVEL

Travels of the Polo family. Marco found his way paved by the pioneering efforts of his ancestors and especially of his father, Niccolò, and his uncle Maffeo. The family had traded with the Near East for a long time, apparently reaching a position of considerable wealth and prestige. Although it is not quite clear if the family was actually of the nobility, the point lost much of its typically medieval relevance in a city of republican and mercantile traditions, such as Venice.

The Polos appear to have been shrewd, alert, and courageous; in about 1260 they foresaw a political change in Constantinople, liquidated their property there, invested it in jewels, and set off for the Volga River, where Berke Khan, sovereign of the western territories in the Mongol Empire, held court at Sarai or Bulgar, according to the season. The Polos seem to have managed their affairs well at this provincial Mongol court, doubling their assets. When political events prevented their return to Venice, they travelled eastward to Bokhara, eventually ending their journey four years later (1265) at the capital of the Mongol Empire, probably the summer residence of the Grand Khan—called Shang-tu (the Xanadu of the English poet Samuel Taylor Coleridge). Establishing friendly relations with the great Kublai Khan himself, they were eventually sent back to Europe as Kublai's ambassadors to the Pope, carrying letters asking the Pope to send Kublai one hundred intelligent men "acquainted with the Seven Arts"; they also bore gifts and were asked to bring back some oil from the lamp burning at the Holy Sepulchre in Jerusalem.

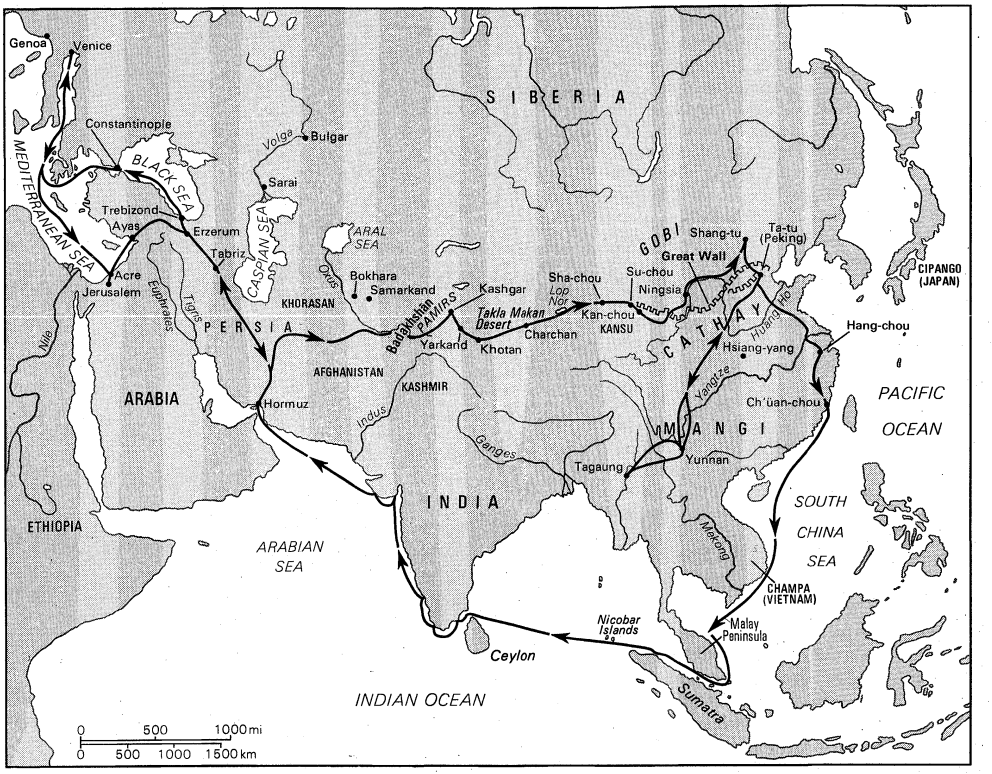
Marco himself had meanwhile been born in or around 1254. (This date, like nearly all the others concerning major events in his life, is conjectural.) Nothing is known of his early years in Venice, although it seems fairly certain that he learned little or no Latin. He was a lad of 15 or 16 when he first met his father, on the latter's return from the East in 1269.

Marco's journey to Asia. When Niccolò and Maffeo returned, they found that the Pope (Clement IV) had recently died. They waited patiently in Venice for a new pontiff to be elected, but after two years a successor had not yet been chosen. Becoming restive, the Polos started off on their journey, taking the young Marco with them (1271). In Palestine the papal legate, Teobaldo of Piacenza, gave them letters for the Mongol emperor; the Polos had been only a few days on the road when they heard that their friend Teobaldo had been elected pope (Gregory X). Back they went to Acre, where they were given proper credentials and the company of two friars. The small expedition departed once more; the two friars, however, soon lost heart, and the Polos continued alone.

The journey across western Asia

From Acre the travellers proceeded to Ayas ("Laiazzo" in Marco's writings, now Yumurtalik, on the Gulf of İskenderun, also called Gulf of Alexandretta, in southeast Turkey). During the early part of 1272 they probably passed through Erzurum, in what is now eastern Turkey, and Tabriz, in what is now northern Iran, later crossing inhospitable deserts infested with brigands before reaching Hormuz on the Persian Gulf. There the Polos decided not to risk a sea passage to India and beyond but to proceed overland to the Mongol capital.

Soon they were on the road again, travelling through deserts of "surpassing aridity" toward the Khorasan region in what is now eastern Iran. Turning gradually to the northeast they reached more hospitable lands; Badakhshān ("Balascian"), in Afghanistan, in particular, pleased the travellers. Marco seems to suggest that they remained here for a year; a long illness (possibly malaria) was cured by the benign climate of the district. It has also been supposed that Marco, during this period, visited territories to the south (other parts of Afghanistan, Kafiristan in the Hindu Kush, Chitral in what is



Routes travelled by Marco Polo between 1271 and 1295.

now Pakistan, and perhaps Kashmir); it is, however, difficult to establish which districts he traversed and which he may be describing from information gathered en route.

Leaving Badakhshān, the Polos proceeded up the Vākhān ("Vocan") valleys toward the Pamirs, which were eventually crossed by an itinerary that has long been a subject of discussion and conjecture. Descending on the northeastern side of the chain, they reached Kashgar ("Cascar") in what is now the Sinkiang Uighur Autonomous Region of China. By now the Polos were on the main Silk Road, and their advance can be followed along the oases to the south and east of the Takla Makan Desert—Yarkand ("Yarcan"), Khotan ("Cotan"), Charchan ("Ciarcian"), and Lop Nor (Lop Lake)—stepping-stones on the way to Sha-chou ("Saciu") on the borders of China, a place now called Tun-huang (Kansu province).

Before reaching Sha-chou, the Polos had travelled primarily among Muslim peoples—also encountering sprinklings of Nestorian Christians, Buddhists, Manichaeans, and Zoroastrians; now, however, they entered the vast province of Kansu (called "Tangut" by Marco), where an entirely different civilization, mainly Buddhist in religion but partly Chinese in culture, prevailed. Their itinerary most probably took the travellers to Su-chou ("Sukchu") and Kan-chou ("Campiciu") before entering the Ningsia area. It is not clear whether the Venetians reached the Mongol summer capital of Shang-tu ("Cian-du") directly or after a detour; at all events, some time in 1275 (1274, according to recent researches by a Japanese scholar, M. Otagi), the Polos were again at the Mongol court, presenting the sacred oil from Jerusalem and the papal letters to their patron, Kublai Khan.

His sojourn in China. For the next 16 or 17 years the Polos lived in the Emperor's dominions—which included, among other countries, Cathay (now North China) and Mangi, or "Manzi" (now South China). Possibly they moved with the court from the summer residence, Shang-tu, to the winter one, Ta-tu or "Taidu"—modern Peking.

Unfortunately Marco's book *Il milione* ("The Million"), known in English as the *Travels of Marco Polo*, is only incidentally a biography and autobiography; it is therefore exceedingly difficult to ascertain where the Polos went and what they did during these years. It is well known that many foreigners were in the employ of the Mongol state, since Chinese were mistrusted; it was therefore natural for the Polos to fit in most honourably and successfully with this motley society.

But precisely how successfully? In what specific capacities did they make themselves useful? These and similar points have been discussed for centuries without definite conclusions being reached. The elder Polos were probably employed in some technical role. Once and very abruptly, a glimpse is obtained of them acting as military advisers during the siege of "Saianfu" (historically and presently Hsiang-yang), a city that was finally taken, says Marco, thanks to some "great mangonels" (missile-throwing engines) built according to their specifications. The whole episode is, however, dubious.

Marco himself reached Cathay very young (at the age of 20 or so). Although he knew little or no Chinese, he spoke some of the many languages then used in East Asia —most probably Turkish (in its Coman dialect) as spoken among the Mongols, Arabized Persian, perhaps Mongol, and Uigur Turkish. Marco was noticed very favourably by Kublai Khan, who took great delight in hearing of strange countries; the Emperor repeatedly sent him on fact-finding missions to distant parts of the empire. One such journey took Marco to Southwest China, to Yunnan, and perhaps as far as Tagaung in Burma; on another occasion he visited Southeast China, later giving an enthusiastic description of the city of "Quinsay" (now Hang-chou) and of the populous regions recently conquered by the Mongols. There is ample evidence that Marco considered himself an adoptive son of his new country.

The Polo legend

Marco's reticence concerning personal matters has left a void that has been eagerly filled in by romantic hero-worshippers, who see him as a brilliant young courtier enamouring princesses and governing provinces: indeed, a Marco Polo myth has been current for centuries and has often provided a theme for novelists, film makers, or dramatists. On the other hand, more sober critics point out that contemporary Chinese records show no trace of Marco. (But under what name was he known: who would recognize the 16th- and 17th-century Italian missionary Matteo Ricci under Li Matou or the 18th-century painter Giuseppe Castiglione under Lang Shih-ning?) They also observe that his reports sound more like echoes of bazaar gossip than gleanings from inner councils. The

truth probably lies somewhere in between. Apart from the missions he undertook for the Emperor, Marco's evident competence in matters pertaining to salt and the administration of its monopoly suggest that he may have held some responsibility in this branch of government. According to some texts of *Il milione*, Marco governed the city of Yang-chou (in Kansu) for three years, sometime between 1282 and 1287; but this seems hardly credible and hinges entirely on the reading of one word.

The return to Venice. Some time around 1292 (1290 according to Otagi), a Mongol princess was to be sent by sea to Persia, there to become the consort of Arghun Khan, and the Polos offered to accompany her. It seems that Kublai Khan was unwilling to let them go but finally granted permission. Kublai Khan was then nearing his 80th year, and his death (and the consequent change in regime) might well have been dangerous for a small group of isolated foreigners. They also, naturally, longed to see once more their native Venice and their families.

The princess, with some 600 courtiers and sailors, and the Polos boarded a fleet of 14 ships, which left the port of Ch'uan-chou ("Zaiton") and sailed southward. The fleet touched Champa ("Ciamba," modern Viet Nam), as well as a number of islands and the Malay Peninsula. On the island of Sumatra ("Lesser Giaua") a stop of five months was made to avoid monsoon storms. There, Marco was much impressed by the fact that the North Star appeared to have dipped below the horizon. The fleet then passed near the Nicobar Islands ("Necuveran"), touching land again in Ceylon ("Seilan"). The Chinese ships subsequently followed the west coast of India and the southern reaches of Persia, finally anchoring at Hormuz. The expedition then proceeded to Khorasan, handing over the princess not to Arghun, who had died, but to his son Maḥmūd Ghāzān.

The Polos eventually departed for Europe, but their movements at this point are not clear; possibly they stayed for a few months in Tabriz. Unfortunately, as soon as they left the Mongol dominions and set foot in a Christian country, at Trebizond in what is now Turkey, they were robbed of most of their hard-won earnings. After further delays, they reached Constantinople and finally Venice (1295). The story of their dramatic recognition by relatives and neighbours, who had thought them long since dead, is a part of Polo lore that is universally known.

"IL MILIONE" AND ITS INFLUENCE

Compilation of the book. Soon after his return to Venice, Marco was taken prisoner by the Genoese—great rivals of the Venetians at sea—during a skirmish or battle in the Mediterranean; Marco was then sent to Genoa and locked in one of the local prisons. There a most felicitous encounter occurred. The Venetian traveller met a prisoner from Pisa, who had possibly lost his liberty some ten years earlier at the Battle of Meloria—one Rustichello (or Rusticiano), a fairly well-known writer of romances and a specialist in chivalry and its lore, then a fashionable subject. Marco may have had it in mind to write a report of his 25 years in Asia but possibly did not feel sufficiently at home in either Venetian or Franco-Italian; but now Rustichello was at hand, and the traveller began dictating his tale. The language employed was Franco-Italian—a strange composite tongue fashionable during the 13th and 14th centuries. Page by page a great book was compiled, later to be known as *Il milione*.

Imprisonment in Genoa

Fortunately Marco was soon freed and returned to Venice. His subsequent life can be reconstructed, in part, through the testimony of some legal documents that have reached us. He seems to have led a somewhat retired life, managing a not too conspicuous fortune. He died in 1324, at the age of 70. A famous story relates how Marco was asked on his deathbed to retract the "fables" he had invented in his book; his answer was that he barely told half of what he actually saw. It is known that in his will he set free a "Tartar slave," who may possibly have followed him all the way from East Asia.

Nature and content of "Il milione." The actual personality of Marco is somewhat elusive. Practically all that is known about him derives from the text of his book; the traditional view is that Marco dictated the text, and certainly many portions sound like transcriptions from a tape recording. But was Rustichello always such a passive instrument? Scholarly research has shown that he often asserted his personality and familiar phraseology, especially in the somewhat standardized description of battles.

The book itself was an instant success. "Tutta Italia in pochi mesi ne fu piena" ("In a few months it spread throughout Italy"), wrote Giovanni Battista Ramusio, the 16th-century Italian geographer. Unfortunately, however, people read the book for the wrong reasons—not as history, geography, or travel but as a fantastic romance in which Kublai Khan was ranked with King Arthur and Cathay became a new province in the dreamy topography of chivalry.

Since all this happened long before the invention of printing, professional scribes or amateurs made dozens of copies of the book, as well as free translations and adaptations—often adding to or subtracting from the text with little or no respect for authenticity. There were also many unfamiliar names that rarely passed unharmed from one copy to another; as a result, modern commentators throw up their hands in despair when confronted with the task of reconstruction. In fact, an authentic and original copy of *Il milione* does not exist; there are some 140 different manuscript versions of the text, in a dozen different languages and dialects—an immensely complex and controversial body of material representing one of the most obdurate philological problems inherited from the Middle Ages.

Different versions of the book

Marco seems to have made additions and corrections himself, on various copies of the work, during the last 20 years or so of his life. Some editors—for instance, the friar Pipino, who made a good Latin translation of the original—found many of Marco's descriptions or interpretations impious or dangerously near to heresy and therefore heavily bowdlerized the text.

The book seems to have been conceived by Marco as a vast cosmography based on first-hand experience—the book to end all books on Asia. The original title—*Divisament dou Monde* ("Description of the World") is significant; Marco all too often sacrifices personal recollections to doctrinal ambitions. Details concerning travel, stages (*i.e.*, distances covered), seasons are rarely stated; the panorama is observed from an impersonal distance with a powerful wide-angle lens. In *Il milione* Marco follows a definite itinerary but often branches off into descriptions of places probably visited not by himself but by his relatives or people he knew. Typical digressions are those on Mesopotamia, the Assassins and their castles, Samarkand, Siberia, Japan, India, Ethiopia, and Madagascar. *Il milione* is better understood if considered as forming a part of the vernacular didactic literature, of which the Middle Ages offer many examples.

The origin of the popular title, *Il milione*, is not quite clear; it may have source relation with the idea of a "tall story," but it may also derive from a nickname running in the family, possibly traceable back to a corruption of Aemilione (Big Emil). The complete silence of the poet Dante Alighieri regarding his illustrious contemporary seems to be significant; he evidently considered the book a fable, an invention, and a dangerously heretical one at that. Marco's descriptions of peoples living under the Equator, where according to the orthodox views of the times there should have been nobody, may have appeared particularly alarming.

Marco Polo's subsequent reputation. As a result of such controversies, which have continued for nearly seven centuries, Marco's reputation has suffered dramatic ups and downs. For some he was a genius, a man of prodigious memory, a most conscientious observer, and a successful official at the Khan's court—a man at home in the cosmopolitan Asia of the great Mongol rulers. For others he was a braggart, who made too much of himself; a drifter ready to believe the gossip of ports and bazaars; a man with little culture, scant imagination, and a total lack of humour; a man who failed, among other things,

to mention the Great Wall of China, the use of tea, and the ideographic script of the Far East.

A more balanced view must take into account many factors, especially the textual problem and medieval ideas of the world. Modern scholarship and research have, however, given a new depth and scope to his work. It is now generally conceded that he reported faithfully what he saw and heard, even though much of what he heard was fabulous or distorted. His information about Japan and South China, for instance, reflects what the Mongols may have said and thought, rather than the actual facts.

Marco must have been endowed with a constitution of exceptional sturdiness to survive the stresses of his long journeys and must also have been well-balanced emotionally to adapt to the repeated changes of social environment. Certain stresses, however, seem to have left definite traces. The Polos had long been merchants, but events turned them into ambassadors and officials. Marco's book frequently reads like the recollections of a retired ambassador surveying men and events, and only secondarily like the report of a merchant. He may also have felt uncomfortable about his origins. Tatars despised merchants, employing mostly Muslims for trade. Marco's inner conflict about this is possibly revealed in his somewhat grudging mentions of his elders. Another detail: the wife he chose upon returning home, Donata, belonged to one of the oldest Venetian families (Badoer), and their three daughters married into the aristocracy. Much of what has been considered irresponsible bragging in Marco's writings may be the unconscious expression of great ambitions that were kindled by circumstances but thwarted by limitations of background and education.

BIBLIOGRAPHY. A Franco-Italian text of Marco Polo's book, generally considered very near to the lost original, has been edited, collated with many other early manuscripts, and published by L.F. BENEDETTO, *Il Milione, prima edizione integrale* (1928). This work is fundamental; a long and detailed introduction (in Italian) deals with the complex problems related to the text and its tradition. The French Orientalist M.G. PAUTHIER was the first modern scholar to confront *Il milione* seriously, but his translation, *Le Livre de Marco Polo, citoyen de Venise*, 2 vol. (1865), is in part superseded. The works by SIR HENRY YULE, now somewhat dated, have been standard texts for decades: *Cathay and the Way Thither: Being a Collection of Medieval Notices on China*, new and enl. ed., 4 vol. (1913–16, reprinted 1967) and *The Book of Ser Marco Polo, the Venetian, Concerning the Kingdoms and Marvels of the East*, 3rd ed. rev. and enl. by H. CORDIER (1903, reprinted many times). A.C. MOULE and P. PELLIOT have published a new composite translation based on the most reliable texts: *Marco Polo: The Description of the World*, 2 vol. (1938). A modern, readable, and dependable translation in English is *The Travels of Marco Polo*, trans. and with introduction by Ronald Latham (1958). The itineraries of the Polos are critically examined in N.M. PENZER'S introduction to *The Most Noble and Famous Travels of Marco Polo*, edited from the Elizabethan translation of JOHN FRAMPTON, 2nd ed. (1937); and in MILTON RUGOFF, *Marco Polo's Adventures in China* (1964), Marco Polo's tale followed brilliantly step by step with illustrations from many different sources. For further points of detail, see P. PELLIOT, *Notes on Marco Polo*, 2 vol. (1959–63); and ORIENTE POLIANO, *Studi e conferenze tenute all'Ismeo, in occasione del VII centenario della nascita di Marco Polo, 1254–1954* (1957). A scholarly and very readable introduction to the subject of Marco Polo, his journeys, his book, and his times is L. OLSCHKI, *L'Asia di Marco Polo* (1957; Eng. trans., *Marco Polo's Asia*, 1960). On a more popular level, see H.H. HART, *Venetian Adventurer* (1942, reissued as *Marco Polo, Venetian Adventurer*, 1967); and M. COLLIS, *Marco Polo* (1950). Partial translations, abridgements, adaptations are legion in all major languages, but some of these works are not reliable. EILEEN POWER, *Medieval People*, 10th ed. (1963), contains a first-rate chapter dealing with Marco Polo and his times.

(F.M.)

Polo

Polo, a stick-and-ball game played on horseback, is the oldest of all equestrian sports and, according to some authorities, the oldest organized sport of any kind.

Origins of the game

History. Definite historical references establish that polo was being played in Iran (Persia) during the 1st century AD. Some scholars claim an even earlier birth of the game and credit the Persians with its invention as far back as 2000 BC. In any case, Iran is generally held to be the nation that invented polo; the game is definitely of Oriental origin.

In its early forms, polo was less a sport, in the modern sense of the word, and more a training game for Oriental cavalry units, usually the king's guard or other elite mounted troops. Indeed, to the warlike tribesmen, polo was a miniature battle with as many as 100 "players" to the side.

Polo in time became a national sport and was played extensively by the Persian nobility. Sons of kings were trained in the use of "hawk, cheetah, polo, swordsmanship, and archery." Evidence that women as well as men played the game (in Persia called *Chaugan*, meaning "stick" or "mallet") includes poetical references to the Queen and her ladies playing King Khosrow II Parvez and his courtiers in the 6th century.

By courtesy of the Trustees of the British Museum

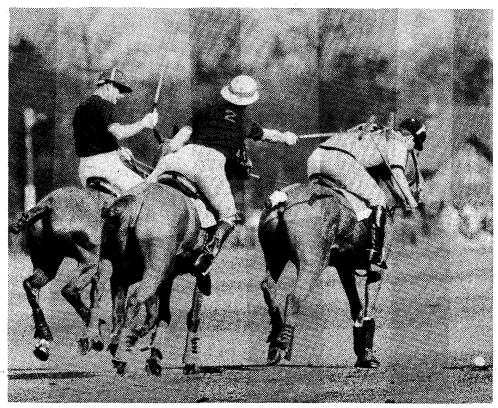
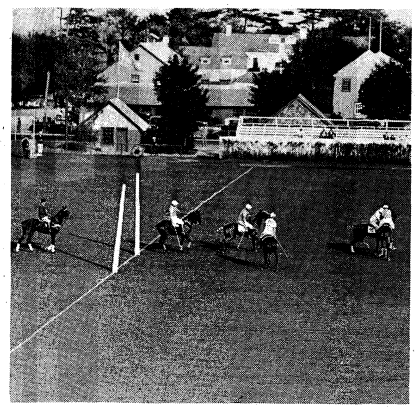
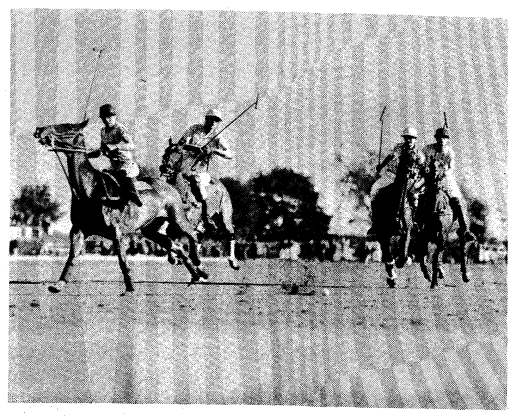


Persian princes playing polo, miniature from an illuminated manuscript, 1472. In the British Museum.

From Iran the game was carried to Arabia, and from there to Tibet (the English word polo is the Balti word meaning "ball"), China, where it found a great following, and Japan. Polo in China often involved more than pure sporting interest; in AD 910, the death of a favoured relative in a game prompted Emperor T'ai Tsu to order all surviving participants in the game beheaded.

Polo was introduced into India by the Mohammedan conquerors in the 13th century. It was there, six centuries later, that the sport was played for the first time by representatives of a Western culture. Although polo had been described in Sir Anthony Sherley's *Travels to Persia* (1613), the first Europeans known to play the game were some British tea planters in Assam, who had learned the game in nearby Manipur and formed the first European polo club in 1859 at Silchar. The Calcutta Polo Club was formed in the early 1860s. The spread of the game was both rapid and dramatic after a captain in the 10th Hussars stationed in India witnessed a match early in 1866 and, being taken with the fast, exciting play, immediately formed a team from among his fellow officers. Before the end of that year informal matches were being staged between British cavalry units in India, and, in 1870, a challenge round was initiated between the 10th Hussars and the 9th Lancers in England. In these early matches there were eight men to the side and almost no rules. Finding a ready interest among prospective players and spectators, polo grew rapidly in England, with matches at Richmond Park and Hurlingham attracting upward of 10,000 fans by 1875.

In 1876, James Gordon Bennett, publisher of the New York *Herald*, witnessed his first polo game, became infatuated, and introduced the game to the United States. By late in that year, informal games were being played in New York, first at Dickel's Riding Academy in Manhattan and then by 1877 at Jerome Park race track in West-

(Left) Play during a tournament at Windsor, Berkshire, in 1965. The centre player hooks his stick under that of the right player in an attempt to interfere with his sweep at the ball. (Centre) East players attacking West team's goal in 1934 East versus West series at Meadowbrook, N.Y. (Right) Manoeuvring for possession of the ball during the U.S.–Argentinian Cup of the Americas series at Westbury, L.I., in 1936, the Argentinians winning the first of their long string of Cup victories.
UPI Compix

chester County. In the summer of 1877, the first North American polo club, the Westchester Polo Club, was formed. One of the early New York polo facilities was the famed Polo Grounds, located on 156th Street. In 1881 the Meadow Brook Club was formed by such outstanding early players as Thomas Hitchcock, Sr., August Belmont, and Benjamin Nicoll.

When polo was first played in the U.S., a team was still made up of eight players. Both in the U.S. and England, this number was later reduced to five and then, in 1881 in the U.S. and in 1883 in England, to four, the present number. Though the rules devised by the Hurlingham Club of England were at first used in the U.S., these were later modified to approximate the rules used today. In 1888 a system of handicapping (see below *The handicap system*) was devised to equalize tournament play, and formal organization was realized in 1890 with the establishment of the Polo Association (predecessor of the United States Polo Association). The Polo Association standardized the rules, and polo spread throughout many parts of the country.

In the 1970s there were more than 100 member clubs (including 10 university and scholastic clubs) in the United States Polo Association playing the game in 30 states, the District of Columbia, Manila, and three Canadian provinces.

Outside the United States, the governing body is the Hurlingham Polo Association, with which the associations of many other countries are associated, including, in the 1970s, India, Pakistan, Assam, New Zealand, Australia, South Africa, Rhodesia, Zambia, Kenya, Nigeria, Malaysia, Jamaica, Malta, Cyprus, the Republic of Ireland, Northern Ireland, Accra, Aden, Tangier, Hong Kong, and the Rhine Army Polo Club. Polo is also actively played in France, Spain, and Italy. There were at the time more clubs and more players participating in polo world-wide than ever before in the sport's history.

International competition. The first of the great international polo competitions was inaugurated in 1886 with the United States challenging Great Britain. The English, at that time the undisputed world leaders of polo, defeated the Americans two straight games in a best-of-three series and won the first Westchester Cup. The U.S. challenged again in 1902 but after winning the opening match lost the two succeeding games. The British again defended in 1909, when the U.S., paced by Harry Payne Whitney, struck hard to win the first two games and the Cup. The Cup was played for nine additional times (the last in 1939) with the Americans winning each time except in 1914. In 1971 the U.S. defeated England for the Coronation Cup in the first international match between the two countries since 1939.

Birth of modern game

The 1909 Westchester Cup series marked the birth of the modern game. Until that tournament polo had been played in the English style, which featured relatively slow play with short, controlled hitting. The Americans had devised a long-hitting, fast-moving, wide-open style of play that devastated the English and revolutionized the sport. Following the Westchester Cup series the rules of the two countries were assimilated. The U.S. adopted the English rule permitting a player to hook an adversary's stick with his mallet, and the English abandoned their offside rules, under which attacking players could not precede the ball, and adopted the handicap system, explained below.

From 1909 until 1950, the U.S. was supreme in polo. In the years before 1920 and throughout that decade and the 1930s, polo had gained in popularity in Argentina, and by 1928 a second great international rivalry had formed. In that year the first Copa de las Americas (Cup of the Americas) was contested between the U.S. and Argentina. The U.S., paced by Thomas Hitchcock, Jr., won that first encounter and repeated again in 1932. After that time, the Copa was played for in 1936, 1950, 1966, and 1969, with Argentina winning each time. In the 1970s Argentina was the unquestioned master of international polo.

With the resumption of the Copa series in 1966 and 1969 combined with the World 30 Goal Championship held in Buenos Aires in 1966 with teams from the United States, Argentina, and Great Britain competing (Argentina won the event) there has been renewed interest in international competition.

To that effect, a privately sponsored British side competed very favourably in the United States in the mid-1960s, and then in 1971 a United States team played in England for the Coronation Cup with the United States side victorious. While the Coronation Cup game (it was a match game, not a series) represents a step toward a possible resumption of the Westchester Cup series, it cannot rightly be placed in the same category as that event.

In another international polo series, between the U.S. and Mexico, the Camacho Cup has been played for three times (the last in 1946), with the U.S. winning on all three occasions.

The game. From its conception as a rough, disorganized, warlike spectacle, polo has evolved into a highly refined, sophisticated sport combining the excitement of horse racing, hockey, and soccer.

Polo is played on an outdoor grass field 300 yards (274.32 metres) long by 160 yards (146.3 metres) wide. There are lightweight goal posts at each end, 8 yards (7.3 metres) apart. A score is made by hitting the ball through these goals. Play commences with the two teams of four men each lined up parallel in the centre of the field. One of the umpires (there are two mounted umpires on the field and a referee on the sidelines) bowls the ball between the line of players. Then, with passes to teammates, speed, and manoeuvring, each team tries to score as the other attempts to prevent a score against it. A game consists of six periods (chukker, chukkar, or chukka) of 7½ minutes each. (Eight chukkers are played in Argentina,

while four is a common number in England and Europe.)

The players and positions

As in most other team sports, each player is assigned a position with certain responsibilities to the strategy of the game. But unlike other team sports, polo players are not known by position name (*e.g.*, forward, halfback, or shortstop) but rather by number. The fine points and intricacies of each position are learned over a period of time with practice and game experience being important factors. The basic duties of the players are as follows:

Number One. This position, usually assigned to the novice or weakest player on a team, is, in fact, one of the most difficult to play. Requiring great anticipation of the play, determination, and self-control, the Number One is theoretically responsible for scoring goals and neutralizing the opposing Number Four (defensive player). He places himself to receive passes from his teammates and generally serves as an offensive spearhead.

Number Two. The Number Two is the "hustler" or "scrambler." Always scrapping for the ball, he requires quick, fast ponies, a keen eye, and an optimistic, aggressive nature.

Number Three. The Number Three, sometimes referred to as the pivot man, "quarterbacks" the team. He must be a long, powerful hitter, capable of tactical leadership. The Three has a dual role: he must feed balls up to the Two and One but also help maintain a solid defense. The Number Three is usually the best player on a team.

Number Four. The Number Four (Back) is primarily a defensive player, who, though he may move anywhere on the field, has as his main function the prevention of the opposition's scoring.

Polo ponies

The polo pony (the term pony persists in polo nomenclature though the mount is really a horse, for there is no longer any height limit on the mounts used) is a unique animal combining the docility of a lady's hack, the quickness of a quarter horse, the speed and endurance of a thoroughbred, and the intelligence of the Arab.

Pure thoroughbreds are used but grade thoroughbreds (horses of mixed breeding—one-fourth or one-eighth quarter horse or other local bloods and three-fourths or seven-eighths thoroughbred) are predominant. Most of the best polo ponies are bred in Argentina or in the U.S., at southwestern or Rocky Mountain ranches where they are broken early and receive thorough work as cow ponies. Most prospective polo ponies begin a training period at about age five that may last from at least six months up to two years. Ponies reach their peak at about age 9 or 10 and with good care and lack of injury may play until age 18 or 20.

Each player is handicapped or rated from 0 to 10 according to his ability and relative worth to a team. (In England minus ratings—*e.g.*, −1, −2—are used, but this practice is not universal.) Players are rated on horsemanship, hitting ability, knowledge of the game, the quality of their horses (the pony is judged to be 60 to 75 percent of a player's ability), and sportsmanship. To indicate the relative difficulty in obtaining a 10 rating, in the early 1970s there were only four such players in the world, all from Argentina. In the U.S., approximately 50 percent of all players (there were about 1,000) were rated 0, 25 percent were rated 1, and the remaining 25 percent were 2 and above—with only 50 players rated at 4 or higher.

Equipment

Players wear a protective helmet, riding boots of a height just below the knee, riding breeches, and a coloured shirt bearing the number of the position each plays. A player may also wear knee pads and spurs (not sharp) and carry a whip.

The ball for outdoor polo is made of bamboo or willow root about 3.25 inches (8.3 centimetres) in diameter and weighing about 4 ounces (113.4 grams). The mallet consists of a bamboo head 9.5 inches in length and weighing about 7 ounces, a flexible bamboo-cane shaft, and a rubber-wrapped grip with a webbed thong for wrapping around the hand. The overall length of the mallet is from 48 to 53 inches (depending on pony size and length of the players' arms). The ball is struck with the side of the mallet head, not the end.

Saddles are English-style with deep seats similar to jumping saddles. The pony's front legs are bandaged from just below the knee to the ankle to prevent injury, and the pony's mane is clipped and his tail braided to prevent interference with the mallet swing.

Variations. Aside from the more common outdoor polo, the game may also be played indoors or in an arena. An indoor field or arena, the standard size of which is 50 yards (45.7 metres) wide by 100 yards (91.4 metres) long, is completely surrounded by wooden boards (called knee boards because they are "knee high" on a mounted player) approximately 4 to 4.5 feet (1.4 metres) high to keep the ball in play. The ball is inflatable leather, 4.5 inches (11.4 centimetres) in diameter and weighing at least 6 ounces (50.1 grams). A team consists of three players rather than four as in outdoor polo. With these exceptions and some minor rule changes, the game is the same as outdoor polo.

BIBLIOGRAPHY. Further information on the history of polo and details on how the game is played may be found in the following references: N. BENT, *American Polo* (1929); J. BOARD, *Polo* (1956); W.C. FORBES, *As to Polo*, 6th ed. (1929); P.G. KENDALL, *Polo Ponies: Their Training and Schooling* (1933); "MARCO" (EARL MOUNTBATTEN OF BURMA), *An Introduction to Polo*, 5th ed. (1965); R.K. MCMASTER, *Polo for Beginners and Spectators* (1954); E.D. MILLER, *Modern Polo*, 6th rev. ed. (1929); and W.G.H. VICKERS, *Practical Polo* (1969). See also publications of the United States Polo Association.

(R.C.L.)

Polybius

Polybius, a Greek who spent 17 years in detention in Rome in the 2nd century BC, is one of the great historians of the ancient world. Combining high moral earnestness and careful research, Polybius told with perception the dramatic story of Rome's rise to world power in the 3rd and 2nd centuries BC. He was born at Megalopolis, in Arcadia, the son of Lycortas, a distinguished Achaean statesman. The dates of his birth and death are inferred at about 200 and 118 BC.

Polybius received the upbringing appropriate for a son of rich landowners. His youthful biography of Philopoemen reflected his admiration for that great Achaean leader; and an interest in military matters found expression in his lost book, *Tactics*. He enjoyed riding and hunting, but his knowledge of literature was rather specialized (apart from the historians) and his acquaintance with philosophy superficial. Between 180 and 170/169, when he was hipparch (cavalry commander) in the Achaean Confederation, nothing is known of his career. But he then became involved in critical events. Encumbered by their war with Perseus of Macedonia, the Romans were watching for disloyalty in the Greek states. Although Polybius declared for open support of Rome and was sent with the cavalry to the consul Quintus Marcius Phillippus, Achaean help was rejected. After Perseus' defeat at Pydna in 168, Polybius was one of 1,000 eminent Achaeans who were deported to Rome and detained in Italy without trial.

Residence in Rome

At Rome, Polybius had the good fortune to attract the friendship of the great Roman general Scipio Aemilianus; he became Scipio's mentor and through his family's influence was allowed to remain in Rome. It is probable that Polybius accompanied Scipio to Spain in 151, went with him to Africa (where he saw the Numidian king Masinissa), and crossed the Alps in Hannibal's footsteps on his way back to Italy. Shortly afterward, when his political detention had ended, Polybius joined Scipio at Carthage and was present at its siege and destruction in 146; and it is likely that he then undertook a voyage of exploration in the Atlantic, which is related in Pliny the Elder's *Historia naturalis* (*Natural History*). Meanwhile, hostilities had broken out between Achaea and Rome, and Polybius was in Corinth shortly after its destruction, in 146. He devoted himself to securing as favourable a settlement as possible for his countrymen and to re-establishing order; and, as the geographer Pausanias states, Achaean gratitude found expression in the erection of statues in his honour at Tegea, Pallantium, Mantineia, Lycosura—where the inscription declared that "Greece

would never have come to grief, had she obeyed Polybius in all things, and having come to grief, she found succour through him alone"—and Megalopolis, where it was recorded that "he had roamed over all the earth and sea, had been the ally of the Romans, and had quenched their wrath against Greece." Of Polybius' life after 146 little is known. At some date he visited Alexandria and Sardis. He is known to have discussed political problems with Scipio and Panaetius of Rhodes. He wrote a history of the Numantine War, evidently after 133 BC, and also a treatise on the habitability of the equatorial region; but when he composed the latter is unknown.

Scope of the history. The history on which his reputation rests consisted of 40 books, the last being indexes. Books I–V are extant. For the rest there are various excerpts, including those contained in the collection of passages from Greek historians assembled in the 10th century and rediscovered and published by various editors from the 16th to the 19th centuries.

Polybius' original purpose was to narrate the history of the 53 years (220–168 BC)—from Hannibal's Spanish campaign to the Battle of Pydna—during which Rome had made itself master of the world. Books I–II form an introduction covering Roman history from the crossing into Sicily against the Carthaginians in 264 and including events in various other parts of the world (especially Achaea) between 264 and 220. In Book III, Polybius sketches a modified plan, proposing to add an account of how the Romans exercised their supremacy down to the destruction of Carthage, in 146.

The events of 168–146 were related in Books XXX–XXXIX. Polybius probably conceived his revision after 146, having by this date completed his narrative down to the end of the Second Punic War. At least Books I–VI seem to have been published by about 150; there is no information as to when the rest of the work, including the revised plan in Book III, appeared.

Conception of history. "All historians," according to Polybius,

> have insisted that the soundest education and training for political activity is the study of history, and that the surest and indeed the only way to learn how to bear bravely the vicissitudes of fortune is to recall the disasters of others.

Practical experience and fortitude in facing calamity are the rewards of studying history and are stressed repeatedly throughout the work. History is essentially didactic. Pleasure is not to be wholly excluded; but the scale comes down sharply on the side of profit. To be really profitable, history must deal with political and military matters; and this is *pragmatiké historia*, in contrast to other sorts of history (IX, 1–2)—genealogies and mythical stories, appealing to the casual reader, and accounts of colonies, foundations of cities, and ties of kindred, which attract the man with antiquarian interests. Its nature is austere, though it may include contemporary developments in art and science. He stands in contrast to the sensationalism of many of his predecessors, who confuse history with tragedy.

Polybius' criticism of other historians

In Book II, in which he attacks the Greek historian Phylarchus, for practices that might be called unprofessional today, Polybius states:

> . . . A historian should not try to astonish his readers by sensationalism, nor, like the tragic poets, seek after men's probable utterances and enumerate all the possible consequences of the events under consideration, but simply record what really happened and was said, however commonplace. For the object of history is the very opposite of that of tragedy. The tragic writer seeks by the most plausible language to thrill and charm the audience temporarily; the historian by real facts and real speeches seeks to instruct and convince serious students for all time. There it is the probable that counts, even though it be false, the object being to beguile the spectator; here it is the truth, the object being to benefit the student.

This attack on Phylarchus is not isolated. Similar faults are castigated in other historians judged guilty of sensationalism (*cf.* II, 16, 13–15; III, 48, 8; VII, 7, 1–2; XV, 34, 1–36). Nor are these their only weaknesses. Many historians are prone to exaggeration—and that for a special reason. As writers of monographs whose subjects are simple and monotonous, they are driven "to magnify small matters, to touch up and elaborate brief statements and to transform incidents of no importance into momentous events and actions" (XXIX, 12, 3). In contrast to such practices, Polybius stresses the universal character of his own theme, which is to narrate "how and thanks to what kind of constitution the Romans in under 53 years have subjected nearly the whole inhabited world to their sole government—a thing unique in history" (I, 1, 5).

Polybius believed that he had a particular reason for adopting a comprehensive view of history, apart from his own predilection for such a view. He wrote:

> Hitherto the affairs of the world had been as it were dispersed . . .; since this date [220 BC] history has formed an organic whole, and the affairs of Italy and Africa have been interlinked with those of Greece and Asia, all tending towards one end (I, 3, 3–4).

The role of Tyche

Indeed, only universal history is capable of adequately treating Rome's rise to world power—the historian's synoptic view matches the organic character of history itself:

> What gives my work its peculiar quality, and is nowadays most remarkable, is this. Tyche [Fortune] having guided almost all the world's affairs in one direction and having inclined them to one and the same goal, so the historian must bring under one conspectus for his readers the operations by which she has accomplished her general purpose. For it was chiefly this consideration, coupled with the fact that none of my contemporaries has attempted a general history, which incited and encouraged me to undertake my task (I, 4, 1–2).

The role here allotted to Fortune is somewhat unusual. For clearly the value of history as a source of practical lessons is diminished if cause and effect are at the mercy of an incalculable and capricious power. Usually, although Polybius uses Fortune to cover a variety of phenomena, ranging from pure chance to something very like a purposeful providence, much of the apparent inconsistency springs from his use of purely verbal elaboration or the careless adoption of current Hellenistic terminology, which habitually made Fortune a goddess. Here, however, Fortune seems to be a real directive power, which raised Rome to world dominion—because Rome deserved it. Normally, Polybius lays great emphasis on causality, and his distinction (III, 6) between the causes of an event (*aitiai*) and its immediate origins (*archai*) is useful up to a point, though it is more mechanical than that of the great Greek historian Thucydides and allows nothing for the dialectical character of real historical situations.

An important place in Polybius' work is occupied by his study of the Roman constitution and army and the early history of the city in Book VI. His analysis of the mixed constitution, which had enabled Rome to avoid the cycle of change and deterioration to which simple constitutional forms were liable, is full of problems, but it has exercised widespread influence, from Cicero's *De re publica* down to Machiavelli and Montesquieu.

Sources of information. Polybius defines the historian's task as the study and collation of documents, acquaintance with relevant geographical features, and, finally, political experience (XII, 25e); of these the last two are the most essential. And he practiced what he preached, for he possessed good political and military experience and had travelled widely throughout the Mediterranean and beyond. Nor did he neglect written sources; indeed, for his introductory books, covering the period from 264 to 220, they were essential. For the main part of his history, from 220 onward, he consulted many writers, Greek and Roman; but following precedent, he rarely names them.

Polybius' private sources

He had access to private sources; for example, Publius Cornelius Scipio's letter to Philip V of Macedonia describing the capture in Spain, in 209 BC, of New Carthage (X, 9, 3), and a letter of Scipio Nasica to some Hellenistic king about the campaigns of the Third Macedonian War (XXIX, 14, 3). He almost certainly consulted the Achaean record office and must have drawn on Roman records for such material as the treaty between Carthage and Philip V (VII, 9). It has not been proved that he had access to the Rhodian records. His detailed

figures for Hannibal's troop formations in Italy came from an inscription left by Hannibal, which he found in the Temple of Juno on the Lacinian promontory.

Polybius regarded oral sources as his most important, and the questioning of witnesses as the most vital part of a historian's task; indeed, this is one reason why he chose to begin his main history at the year 220. Anything else would be "hearsay at one remove," a safe foundation for neither judgments nor statements. Of the thousands whom he must have questioned, few names can be isolated; but at Rome he had opportunities of meeting men from many areas, including internees, ambassadors, and visitors.

Style and qualities as a historian. Writing in the 1st century BC, Dionysius of Halicarnassus reckons Polybius among those who "have left behind them compositions which no one endures to read to the end"; that his successors shared this view of Polybius' style is confirmed by the failure of his works to survive except in an incomplete form. The infelicity of Polybius' Greek (which frequently reproduces the conventional phrases of the Hellenistic chancelleries familiar from contemporary inscriptions) lies in its awkward use of long and cumbersome circumlocutions, vague abstract nouns, and pedantic repetitions. To the scholar his style is, however, no great obstacle; and, though in his anxiety to improve his reader he moralizes and belabours the obvious, the perennial interest and importance of his theme will always ensure him a following among those who can enjoy a historian who is accurate, serious, and sensible, understands the events of which he writes, and, above all, who asks the right questions.

BIBLIOGRAPHY. Two works by FRANK W. WALBANK, *A Historical Commentary on Polybius*, 2 vol. (1957–67), and *Polybius* (1973), contain both general and detailed discussion and comprehensive bibliographies. The standard text is that of THEODORUS BUTTNER-WOBST, 5 vol. (1866–1904); but FRIDERICUS HULTSCH, 4 vol. (1870–92), gives a fuller description of manuscript variants; for the history of the text, JOHN M. MOORE, *The Manuscript Tradition of Polybius* (1965), is invaluable. There is a Greek text with an English translation by W.R. PATON, "Loeb Classical Library," 6 vol. (1922–27); and a translation by E.S. SCHUCKBURGH, 2 vol. (1889, reprinted 1962). Three fascicles (covering about half) of a *Lexicon Polybianum* (1956–) by ARNO HAUERSBERGER have so far been published.

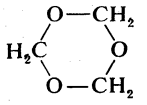
(F.W.W.)

Polymers

Polymers are essential to the existence of man and his civilization. The human body, all animal and plant tissues, and most structural substances of organic nature—such as proteins, wood, resin, and chitin, the hard coating of insects—consist of polymeric (macromolecular) materials. Many minerals, such as diamond, quartz, and feldspar, are inorganic polymers, and numerous products of ancient and modern industry, such as concrete, porcelain, glass, textiles, paper, rubbers, and plastics, are either entirely or substantially polymeric. Only in the first half of the 20th century, however, was it clearly recognized that these substances all possess one essential common feature, namely, that they consist of very large (macro) molecules.

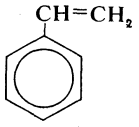
Definition

A polymer (Greek *polys*, "many"; *meros*, "part" or "unit") is a substance the molecules of which are multiples of low-molecular-weight units, molecular weight being, as the name suggests, a measure of the weight of a molecule relative to a chosen standard (for practical purposes, an atom of hydrogen has a weight of one). The low-molecular-weight unit making up the polymer is referred to as the monomer. As long as a polymer is strictly uniform in molecular weight and molecular structure, its degree of polymerization is indicated by the Greek word for the number of monomers it contains; thus, for example, dimers, trimers, tetramers, and pentamers indicate twofold, threefold, fourfold, and fivefold polymers. The term polymer, however, is used to designate a combination of an unspecified number of monomer units. Trioxymethylene, for example, is a (cyclic) trimer of formaldehyde (CH_2O):

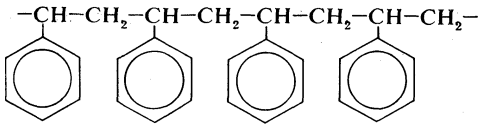


trioxymethylene

and polystyrene is the polymer prepared from the monomer styrene:



styrene



polystyrene

If the number of monomer units in a polymer becomes very large, the latter is sometimes called a high polymer. According to present-day usage, a polymer or high polymer need not consist of individual molecules all having exactly the same molecular weight or precisely the same chemical composition and molecular structure.

In the case of some natural polymers, such as the proteins, the individual molecules all have the same molecular weight and molecular structure. This identity of the polymer molecules results from highly specific cellular processes that produce them according to fixed patterns. With most synthetic and natural polymers, however, significant differences occur in the molecular weight of the individual macromolecules, with the result that the material itself is a mixture of homologous polymeric constituents—that is, of individual molecules differing in the number of monomer units they contain. Methods are not available for preparing synthetic polymers of fixed molecular weight, nor can the prepared mixture be separated completely into fractions with uniform molecular weights. Variability in the composition and molecular structures of polymers results from the presence of different end groups that terminate the chains, occasional branches in the chain, variations in the arrangement of the monomeric units, and irregularity in the sequence of monomeric units of different types if more than one type is present.

A completely rigorous system of classifying and naming polymers, conforming to the practice in naming ordinary organic and inorganic molecules, is impossible because of the common variations in structure between individual molecules of most polymer types—variations that cannot be eliminated, or even determined quantitatively in nature or amount by available methods. Nevertheless, it is convenient to treat a particular type of polymer as a single molecular species for purposes of discussion because these variations in structure do not significantly affect the physical and chemical properties of the substance and do not affect its practical application.

NATURAL POLYMERS

All living beings—animals and plants alike—are composed to a very large extent of organic polymers, which not only provide structural materials for the maintenance of life functions but also participate in carrying out the life functions themselves. The solid portions of all plants consist of several polymers, the most important of which are cellulose, lignin, and various resins.

Cellulose and lignin

Cellulose in wood is intimately connected with lignin; but in other parts of plants, it is found in relatively pure form as fibres such as cotton and flax. Cellulose itself is a polysaccharide, that is to say, a polymer whose monomer units are individual sugar molecules. In cellulose the sugar glucose comprises the monomer units, and they are joined to give molecular weights of up to several millions.

In most fibres the polymer chains are combined to give laterally ordered crystalline units that, in turn, are separated by disordered or amorphous areas. This two-phase structure gives cellulose fibres their strength and, at the same time, their softness and pliability.

In all types of wood cellulose is accompanied by lignin, a second polymer, which shows much less regularity than cellulose. Lignin molecules consist of a complicated three-dimensional network of phenolic residues joined chemically by means of short hydrocarbon chains. Because of this molecular complexity, the molecular weight of lignin is not exactly known, but in any event it is very high. Wood also contains several types of resins, which are polymers of the simple hydrocarbon isoprene and have molecular weights ranging from a few thousand to several hundred thousand.

The most important terpene, as the isoprene polymers are called, is rubber that exists in the sap of many tropical trees, especially of the genus *Hevea*. Rubber is a polyhydrocarbon and consists, in the native state, of long chains of isoprene, which have molecular weights around 500,000. Balata, another natural, rubberlike resin, also is a polyisoprene; it differs from hevea rubber chiefly in the geometry of double bonds, or linkages, between carbon atoms in the molecular chains.

Proteins and nucleic acids

Life on Earth is based largely on the existence, properties, and reactions of two classes of polymers—the proteins and the nucleic acids. The monomeric units of proteins are amino acids, relatively small molecules containing amino ($-NH_2$) and carboxylic acid ($-COOH$) groups. The twenty amino acids that occur normally in proteins are always arranged in precise, but different, orders in different protein molecules. Certain proteins have molecular weights around 50,000; in others the molecular weight reaches several millions. All fibres and membranes of animal tissues consist of high-molecular-weight proteins; the processes of life—such as metabolism, vision, and respiration—are based on the protein materials known as enzymes, which act as catalysts for biochemical reactions. Generally, the exterior parts of the animal body—fur, hair, nails, horn, and skin—consist of proteins with molecular structures of three-dimensional networks of very high molecular weight (*e.g.*, keratin).

Other important natural polymers are the starches, which—like cellulose—are polymers of glucose. Starches are widely distributed in plants and play an important role as a food for animals of all kind. Starches differ from cellulose chiefly in the way the individual monomeric units are joined in the molecules; some varieties—amylose, for example—are simple chains with molecular weights around 100,000; others—called amylopectins—are three-dimensional networks (like lignin) with very high molecular weights. All hard, tough, and strong parts of insects and crustacea—shells, wings, and claws, for example—consist of a nitrogen-containing polysaccharide called chitin. Chitin is composed of long chains of glucose molecules that have been modified by acetylated amino groups; the molecular weights of typical chitin polymer chains are around 100,000.

Many inorganic polymers also are found in nature; these are characterized by their hardness, high melting range, and great mechanical strength. Some inorganic polymers are formed by atoms of the element alone; in others the monomeric units are inorganic compounds. Diamonds are a good example of the first class; this

Diamond, graphite, and silica

substance consists of a three-dimensional network of carbon atoms that are combined with each other exclusively by single bonds. In fact, diamond is the simplest crystalline, three-dimensional polymer that exists. Another polymeric form of the element carbon is graphite, the writing material in a lead pencil. Graphite consists of two-dimensional polymeric layers of carbon atoms joined together alternately by single and double bonds. Between these layers of densely packed carbon atoms are relatively large distances, a factor that causes the typical lamellar (layer-like) character of all graphitic materials. Elemental sulfur and selenium exist generally in crystalline modifications but also in a polymeric form consisting of long chains with molecular weights around 100,000.

The most important inorganic polymer is silica, or silicon dioxide, which exists in several crystalline modifications and in a glossy amorphous form; they are all three-dimensional networks—partly regular, partly random—held together by chemical bonds between silicon and oxygen. The strength of the silicon–oxygen bonds is responsible for the high softening range and the extreme hardness of these compounds.

The three-dimensional silica networks are capable of accepting electrons; in the process they are converted into polymeric ions that can be neutralized and stabilized by positively charged metal ions, such as those of sodium, potassium, magnesium, calcium, and aluminum. The result is a large and important family of silicates that, together with other metals, form most of the hard and high-melting natural rocks—including granite, feldspar, and agate.

SYNTHETIC POLYMERS

Organic. In the course of their studies, organic chemists of the 19th century frequently encountered materials that were neither ordinary liquids nor crystalline solids but of an amorphous semisolid or glassy character. These materials generally defied the classical methods of purification, isolation, and identification. As a consequence, most of these substances were only referred to in passing in scientific writings but were not further investigated. In a few instances, however, the yield of such a material was so considerable and its properties were so attractive for practical use that its preparation was more thoroughly studied, and a larger scale development of the product was carried out. Of these developments, probably the most important stemmed from the discovery in 1909 by Leo Baekeland, the Belgian-born U.S. chemist, of a viscous product from the reaction of phenol with formaldehyde. Baekeland discovered that many species of useful plastics could be prepared by varying the proportions of the components and the conditions of the reaction. This discovery led to the development of the Bakelite resins, which occupy an important position in the plastics industry. In the 1920s, when the structure and character of the most important natural polymers—cellulose, proteins, and rubber—were elucidated, a new, broadly based series of systematic studies of polymer formation was undertaken. These studies led to the production of a large number of synthetic polymers, which in turn permitted an amazing variety of industrial applications of polymers as fibres, plastics, synthetic rubbers, coatings, and adhesives.

This rapid development in the chemistry of synthetic polymers was aided by the fact that most building units for the preparation of synthetic polymers are products of the petroleum-refining industry, which also was undergoing rapid development at the same time.

It is convenient to classify the synthetic organic polymers according to the chemical character of their monomers, beginning with the simplest variety, the hydrocarbons.

Polyhydrocarbons. Many simple and inexpensive unsaturated hydrocarbons, such as ethylene, propylene, butadiene, and styrene, can be readily transformed into polymeric materials by simple addition polymerization, the repeated addition of one monomer unit after another to the growing polymer chain. In this way important thermoplastic and elastomeric (polymers with elastic properties) products are obtained.

Polyethylenes and polypropylenes

There are two ways to prepare polyethylene on an industrial scale: (1) polymerization with oxygen as catalyst at high pressure (1,000–2,000 atmospheres) and high temperatures (190–210° C) and (2) polymerization with organometallic catalysts at ordinary pressure and ordinary temperatures. The high-pressure polyethylenes have specific gravities between 0.92 and 0.93 (water = 1.000), melting points around 110° C, and molecular weights from 20,000 to 40,000. They represent relatively soft, but tough, thermoplastic materials with excellent electrical and mechanical properties and with outstanding resistance against most acids, bases, and solvents. Such high-pressure polyethylenes are widely used for the production

of synthetic fibres and for the preparation of bottles and small containers. The low-pressure polyethylenes have specific gravities between 0.95 and 0.97, melting points around 134° C, and molecular weights up to many hundred thousands. They represent hard, tough, strong thermoplastic materials with excellent electrical, mechanical, and chemical properties. They are widely used for the production of larger bottles, containers, pails, and trash cans.

There are two varieties of polypropylene: a soft, rubbery type of minor importance and a hard, crystalline, high-melting type, which is a valuable thermoplastic material of wide applications in certain areas of the textile industry and also used to make molded objects. The second type of polypropylene has crystalline character, melts at around 178° C, and has a density of about 0.90, and a molecular weight ranging from 50,000 to 200,000.

Polybutadienes and polystyrenes

Most polybutadienes are typical elastomers; they have many valuable properties. Butadiene can be polymerized in solution, suspension, or emulsion with different types of catalysts. Some of the products are crystalline, others are amorphous. All of them remain soft and resilient down to low temperatures and can be cured into elastomers of great strength and excellent abrasion resistance. The molecular weights range from 80,000 to 200,000. Most polybutadienes are soluble in aromatic solvents, such as benzene.

Styrene polymerizes readily in pure form or in solution, suspension, or emulsion—usually under the influence of peroxide catalysts. The polymer is a colourless, transparent, thermoplastic resin (polymer product used to make objects of plastic with the application of heat), which has a specific gravity around 1.08 and a molecular weight range from 50,000 to 200,000. Commercially produced polystyrene is amorphous, has useful electrical and mechanical properties, and softens around 90° C. It is highly resistant to moisture, acids, and bases but is easily attacked by organic solvents. Polystyrene and closely related polymers are the most popular resins for certain molding and extrusion procedures in the forming of plastic objects.

Particularly important are copolymers (polymers composed of several monomer units) of butadiene and styrene, which are made in emulsion with the aid of peroxide catalysts. During the 1940s and 1950s these amorphous elastomeric materials were the most important synthetic rubbers and, subsequent to that time, continued to be valuable and useful products. They can be compounded and used like natural rubber, and they exhibit excellent resistance to abrasion and deterioration caused by oxygen, heat, and light.

Polyvinyls and polyacrylics. Many derivatives of ethylene still retain the double bond and are able to polymerize. Among the materials produced in this way are several important families of polymers, the polyvinyls and polyacrylics.

Polyvinyl chlorides and polyvinyl esters

The replacement of one hydrogen atom in ethylene by chlorine leads to vinyl chloride, which polymerizes readily under the influence of peroxide catalysts in suspension or emulsion. The product, polyvinyl chloride, is a colourless, hard, tough thermoplastic material, which softens only at a high temperature. Its specific gravity is around 1.55; it is substantially amorphous and resistant to most destructive chemical agents, including hot organic solvents. The molecular weight ranges from 50,000 to 150,000. It is a very useful material prepared as sheets and as extruded or molded products.

Vinyl acetate and vinyl butyrate are liquid esters of vinyl alcohol (the hydroxide equivalent of vinyl chloride). They polymerize with peroxide catalysts in solution, suspension, and emulsion to give amorphous, soft resins with molecular weights between 30,000 and 100,000. The products find many uses as coatings and adhesives. Both vinyl acetate and vinyl butyrate copolymerize readily with vinyl chloride to give a large family of thermoplastic materials. In these substances the vinyl chloride contributes hardness and high-melting characteristics, whereas the vinyl esters act as internal (*i.e.*, chemically bonded) plasticizers, additives that impart favourable qualities to plastics. These copolymers exhibit a wide range of properties; hence they find many applications, such as for coatings, cohesives, films, and extruded or molded articles of many kinds.

Polyacrylics

The basic units for the large and important family of polymeric materials known as polyacrylics are acrylic acid (propenoic acid) and methacrylic acid (a methylacrylic acid with the methyl and carboxyl groups joined to the same carbon atom). The methyl esters of these acids are readily polymerized with the aid of peroxide catalysts in the pure material or in solution, suspension, or emulsion. Polymethyl methacrylate is a hard, colourless, brilliantly transparent thermoplastic material that is widely used to manufacture extruded and molded articles. Equally important is the polymer of acrylonitrile (vinyl cyanide) that is usually made in suspension or emulsion with the aid of peroxide catalysts. The product has a molecular weight in the range 70,000 to 120,000; it represents a white, hard, moderately crystalline, material, which softens at a high temperature and is soluble only in a few selected high-boiling organic liquids, such as *N*,*N*-dimethylformamide and dimethyl sulfoxide. Copolymers of acrylonitrile (80–90 percent) and methyl acrylate or vinyl acetate occupy an important place in the man-made fibre industry. Copolymers of acrylonitrile (10–15 percent) and styrene represent thermoplastics for extrusion and molding. The products possess superior hardness, higher softening ranges, and better impact resistance than the polymers prepared from either monomer alone. Finally, copolymers of acrylonitrile (20–30 percent) and butadiene, prepared in emulsion, represent a valuable species of synthetic elastomer. These so-called nitrile rubbers have excellent resistance to hot oils, oxygen, ozone, light, and heat; therefore, they find important applications in the construction of airplanes, automobiles, and electrical equipment.

Polyacetals and polyethers. In addition to the polymers whose backbone chains consist of a continuous sequence of carbon–carbon bonds, there are many important polymers with oxygen or nitrogen atoms in the backbone chain along with carbon atoms. Among those with oxygen atoms in the backbone are the polyacetals and polyethers. One of the simplest polyacetals is polyformaldehyde, a hard, high-melting, crystalline material with excellent resistance against abrasion and against the action of solvents. It is prepared by polymerization of formaldehyde—a very simple and inexpensive monomer—in the gaseous state with anionic catalysts. It is used for making machine parts in place of, or in conjunction with, certain metals. Polyoxyethylene, a polyether, is prepared in bulk or in solution from ethylene oxide with cationic catalysts. The product is a white, crystalline, water-soluble material with a melting point of about 75° C; it is a particularly useful thickener of aqueous solutions and a material used for coatings for textiles and paper.

Polyesters. When a bifunctional carboxylic acid, such as adipic acid, $HO_2C(CH_2)_4CO_2H$, reacts with a bifunctional alcohol, such as ethylene glycol, $HO(CH_2)_2OH$, under conditions that lead to the elimination of water, a linear macromolecule is formed that is characterized by a repetition of ester groups (–COO–) along the chain and is called, therefore, a polyester. Each combination of a specific dicarboxylic acid and a specific glycol leads to a particular polyester.

Nomenclature of linear polymers

In a convenient naming system for linear polyesters, the number of carbon atoms in the glycol and in the dicarboxylic acid are used after the word polyester. For example, hexamethylene glycol, $HO(CH_2)_6OH$, with six carbon atoms, and adipic acid, $HOOC(CH_2)_4COOH$, also with six carbon atoms, give a polyhexamethylene glycol adipate commonly referred to as polyester-6,6. Similarly, polydecamethylene succinate is polyester-10,4. Polyesters can also be formed by the self-condensation of hydroxycarboxylic acids, such as hydroxyundecanoic acid, $HO(CH_2)_{10}COOH$. The resulting polymers need only one figure to characterize them; in this case, the product polymer is called polyester-11.

All linear polyesters made with bifunctional open-chain glycols and dicarboxylic acids are colourless, crystalline,

thermoplastic materials. Their melting points and solubilities vary over a wide range and depend on the length of the methylenic ($-CH_2-$) segments between the ester bonds in the chain. Because of the polarity (separation of electrical charge) of these bonds, they are mainly responsible for the intermolecular attraction between the chains, and hence for their softening and solubility behaviour. The closer the ester bonds, the higher is the melting point of the polyester and the less is its solubility. Polyethylene glycol oxalate, polyester-2,2, melts around 200° C, whereas polyester-6,6 melts around 80° C. The former is soluble in only a few selected organic liquids; the latter dissolves readily in most ordinary organic solvents, such as benzene, ethyl methyl ketone, and butyl acetate.

Low-molecular-weight polyesters (molecular weight up to 5,000) are excellent plasticizers for many natural and synthetic resins; the high-molecular-weight species (above 10,000–15,000) are valuable thermoplastics themselves and are used for the manufacture of films and molded objects. Polyesters with high melting points (250° C and above) are obtained from aromatic (cyclic) constituents. Examples are the polymer made from ethylene glycol and terephthalic acid (benzene-1,4-dicarboxylic acid)—the so-called polyester-2T—and the material made by the self-esterification of 4-hydroxybenzoic acid —the HOB-polyester—both of which melt above 250° C and do not dissolve in most organic solvents. Because of their high melting points and insolubility, they represent valuable materials for the formation of films used in electrical insulation, photography, and packaging, and also for fibres used in fabrics and tire cord. When the polyesterification is carried out in the presence of a polyfunctional component—such as pentaerythritol, a tetrahydroxy compound, or a tricarboxylic acid—the product is not a linear macromolecule but a highly branched species that, after sufficient removal of water, is converted into a three-dimensional network. Unlike the thermoplastic linear polyesters, systems of this kind are thermosetting—that is, the original material softens on heating, but once set in a new shape it is extremely resistant to further heating. The thermoset polyesters are insoluble, hard, and tough materials of extremely high molecular weight. They are widely used as coatings, laminating agents, and in making molded objects. They are usually applied as thermosetting compositions that are cured into the thermoset state.

Polyamides. When a bifunctional amine (two $-NH_2$ groups) reacts with a bifunctional acid (two $-CO_2H$ groups) with elimination of water, a linear macromolecule is formed that contains in the chain a regular repetition of amide groups (–CO–NH–) and is therefore called a polyamide. More commonly, these materials are known as nylons. The same nomenclature is used for polyester as for polyamides. Thus, the nylon made of hexamethylenediamine, $H_2N(CH_2)_6NH_2$, and adipic acid, $HOOC(CH_2)_4COOH$, is nylon-6,6; the polymer made from the same diamine and sebacic acid, $HOOC(CH_2)_8COOH$, is nylon-6,10; and the material made by the self-polymerization of aminocaproic acid, $H_2N(CH_2)_5COOH$, is nylon-6. All polyamides consisting of linear, bifunctional components are hard, tough, colourless, crystalline, thermoplastic materials. Their softening and solubility behaviour is controlled by the magnitude of the intermolecular forces between the individual chains; the major contribution to these forces comes from the amide bonds (–CO–NH–), which are polar themselves and also capable of establishing hydrogen bridges ($>C=O\cdots H-N<$) between adjacent macromolecular chains. The closer the amide groups are in the chain, the stronger is the cohesion between chains, and the stronger the cohesion, the higher the melting point, and the greater the insolubility. For example, the polyamide of ethylenediamine and oxalic acid, nylon-2,2, melts above 300° C (with decomposition), whereas nylon-6,6 melts around 260° C and nylon-6,10 at 215° C. An interesting contrast exists between the polyester series and the polyamide series: corresponding members—such as the polyester-6,6 and the polyamide-6,6—differ substantially in softening temperature and solubility, the nylons having much higher melting points and much lower solubilities than the esters. This difference is usually attributed to the fact that the polyamides can develop hydrogen bridges whereas the polyesters cannot.

Polyamides and polyesters compared

Although nylons find many important applications in the field of extrusion and molding, their outstanding use is in the man-made fibre industry. Nylons dominate all other synthetic fibres in quantity and quality and find successful use in fabrics, tire cords, and many industrial fibres.

Polyurethanes. Another important family of synthetic organic polymers is obtained by the reaction of bifunctional isocyanates (containing two –NCO groups) with bifunctional alcohols (two –OH groups). Again, one obtains linear macromolecules that, in this case, are characterized by the regular repetition of the urethane groups (–O–CO–NH–). The presence of the urethane groups permits the buildup of a three-dimensional network because of the reaction of the hydrogen of the –NH– group with the terminal isocyanate group of a chain. These networks are either elastomeric or rigid; they are used with great success in the field of elastomeric fibres (spandex) and in making both soft and rigid foams.

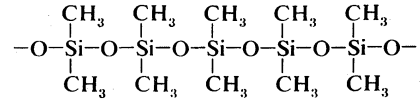
Inorganic. Although the synthetic organic polymers are a very large family of useful materials, are the product of a very large industry, and although world production in 1970 exceeded 50,000,000 tons, synthetic inorganic polymers have not found any large-scale industrial applications. Synthetic polysilicates (in the form of colloidal particles) are used as fillers and reinforcing agents for plastics used as dentures as well as in synthetic rubbers. Certain polyphosphates with medium molecular weight find some use as flameproofing agents.

Synthetic polysilicates

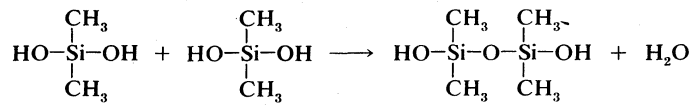
The only linear inorganic polymer of known structure and molecular weight is made by the interaction of phosphorus pentachloride (PCl_5) and ammonia (NH_3), with the elimination of hydrogen chloride (HCl). The product is known as polyphosphonitrile chloride; the molecular chains, which have a molecular weight of about 100,000, are composed of repeating $PNCl_2$ units. The substance is a yellow, amorphous, soft, rubbery material that dissolves in many organic solvents and, if stretched to a few hundred percent, shows a distinct fibre X-ray diagram, an indication of a regular pattern in the structure of the fibres. Because the material is easily hydrolyzed (broken apart with water), it has found little use as a structural elastomer, but it does have potential use as a polymeric flameproofing agent.

Mixed organic–inorganic. Considerable progress was made during the 1950s and 1960s in the synthesis, characterization, and use of linear polymers containing substantial amounts of inorganic elements in the backbone of the chain, together with organic groups attached to the backbone. The most important representatives of this relatively young family of polymers are the polysiloxanes—or silicones, as they are generally called—which have a backbone chain of alternating silicon and oxygen atoms. Also attached to each of the silicon atoms are two monovalent organic groups, such as the methyl ($-CH_3$), or phenyl ($-C_6H_5$) groups. The structure of polydimethyl siloxane is as follows:

Silicones



Polymers of this type are prepared from monomeric dialkylsilanediols by removal of water between molecules. Preparation of the dimer of dimethylsilanediol demonstrates how the process occurs.



Polysiloxanes of low molecular weight, which sometimes have a closed-ring instead of an open-chain structure, are

oils and greases. Higher-molecular-weight types are versatile elastomeric materials; they remain soft and rubbery even at temperatures as low as −80° C and, at the same time, are more stable at high temperatures (3,000° C and above) than most conventional plastics and rubbers. Similar polymer chains, in which oxygen atoms alternate with boron, phosphorus, or aluminum atoms, also have been studied, but generally their molecular weights are too low (not higher than 2,000) to provide useful mechanical strength.

STRUCTURE OF POLYMERS

The fundamental consideration in the chemical study of polymers, as of any molecular substance, is the composition of the material in terms of its content of the various kinds of atoms and its structure or arrangement of the atoms in the molecules. Determining the composition is substantially a matter of elemental analysis and does not differ from the procedures used in the study of ordinary inorganic and organic compounds. Furthermore, the composition of the polymer follows simply from the compositions of the monomer, or monomers, used to build it up—with due consideration for such additives as catalysts, activators, modifiers, and stabilizers, some of which may become incorporated in the polymer molecules.

Molecular structure

It is much more complicated, however, to define and, consequently, to establish what is generally called the structure of a polymeric or macromolecular system. The term structure refers, first, to all of the information that can be collected concerning a representative, isolated macromolecule of the material, and second, to the manner in which the individual polymer molecules are geometrically arranged and physically bonded to one another in a macroscopic sample of the material. The former aspect of the structure may be called the molecular, or intramolecular, structure; whereas the latter is the supermolecular, or intermolecular, structure—also referred to as the texture, or morphology.

Molecular weights. In considering the molecular structure, it must always be borne in mind that, with the exception of proteins and certain nucleic acids, all natural and synthetic polymers are heteromolecular, or polymolecular, in character—*i.e.*, they do not consist of a single species of large molecules, but rather of a mixture of many species, which have different molecular weights but substantially the same chemical composition. The best characterization of a given polymer is, therefore, its molecular-weight distribution, which is a representation of the percentages of each molecular weight species in the sample. Determining such a distribution requires a delicate and, sometimes, difficult study of the system; this can always be achieved in a satisfactory manner, however, as long as it is possible to prepare a dilute solution (0.1–2.0 percent) of the polymer in an appropriate solvent and to reprecipitate the polymer from the solution without chemical changes. The procedures for establishing molecular weight distributions can be divided into analytical and preparative methods, depending upon whether one obtains information only about the distribution or whether one actually obtains the individual fractions with different molecular weights.

Methods of analysis

The oldest of the analytical methods is ultracentrifugation, a high-speed centrifugation process that separates molecules in solution according to their molecular weights. Analysis of the distribution of the polymer molecules after centrifugation makes it possible to arrive at a complete description of the spread of molecular weights in the original polymer material. Refinements of this method have increased its applicability to complicated polymeric systems and may make it possible to accomplish a physical separation of the different molecular species in this way.

Gel-permeation chromatography—*i.e.*, separation of molecular mixtures by filtration of a solution through a jellylike material—is a more recently discovered practical way to arrive at the molecular-weight distribution of soluble polymers and may, with further improvements, make it possible to obtain the individual fractions.

Extraction and precipitation are the standard preparation methods, which actually furnish more or less uniform fractions of a polymer, each of which is composed of a narrower molecular-weight distribution than is the original material. With the aid of repeated, careful extraction or precipitation steps or both, it is possible to obtain relatively large amounts of materials that are substantially more uniform than the raw polymer.

The systematic application of these methods has shown that all natural and synthetic polymers, such as cellulose, rubber, polyethylene, nylon, and polyvinyl chloride, have rather broad distribution functions; *i.e.*, they represent mixtures of molecules that differ considerably in their molecular weight. In nylon, for example, the 10 percent of the molecules with largest molecular weight have on the average about 10 times the molecular weight of the 10 percent of lowest molecular weight. For many purposes, it is satisfactory to approximate the molecular-weight distribution by a numerical average: doing this means that a mathematical function is replaced by a single number.

Number and weight averages

In statistical approaches to any subject it is customary to use different averages, which are characterized by the type of quantity averaged. If a polymer is divided into a number of fractions on the basis of molecular weight, two kinds of average molecular weight can be obtained: number-average and weight-average molecular weights. The number-average molecular weight is obtained by adding together the molecular weights of each fraction (*i.e.*, the number of molecules in each fraction times the molecular weight of that fraction) and dividing the total by the number of all molecules in the system. The number-average molecular weight can be measured experimentally by end-group analysis (detection of the ends of the polymer chains) or by measurements of the vapour pressure or osmotic pressure of solutions of the polymer, both effects depending on the numbers of molecules in the solution.

On the other hand, the weight-average molecular weight is obtained by adding together the actual weights of all molecular weight fractions (*i.e.*, the weight of each fraction times its molecular weight) and dividing this sum by the total weight of all molecules in the system. The weight-average molecular weight can be experimentally determined by light scattering, ultracentrifugation, or viscosity measurements of very dilute solutions, all of which are influenced by the weight of material present in the sample.

Separate determinations of the number- and weight-average molecular weights give an idea of the relative uniformity, or nonuniformity, of a given polymer sample. In an ideal case, in which all molecules have a single molecular weight, the number and weight averages are the same. When the polymer sample contains many molecular classes, however, each with a different molecular weight, the two averages are likely to be quite different. When experimental studies show that the number- and weight-average molecular weights are very similar, the polymer itself is relatively uniform—that is, its molecules all have about the same molecular weights. On the other hand, when the two averages are found to be quite different, then the polymer sample has a broad distribution function and the material is polymolecular and nonuniform.

Arrangement of monomer units. The second step in the description of the molecular structure of a polymeric material is the determination of the overall shape of the polymer molecules, or the arrangement of the individual units in the polymer. Clearly, the polymers with the simplest structures are the linear polymers. In such macromolecules the individual structural units, or monomers, are joined to one another in linear sequence. Each terminal unit is connected to only one other unit; each chain is connected to two others. At no point in the chain is any unit connected with three or more other units. Linear macromolecules invariably have only two end groups. Systems composed solely of polymers of this kind show reversible melting and solubility behaviour.

Nonlinear polymeric systems, such as branched mac-

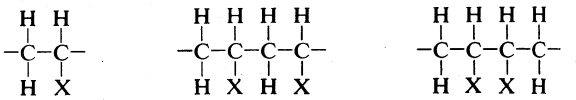
romolecules in which certain of the monomer units are connected with three or four neighbours, are also known.

Linear and nonlinear systems

If the number of monomers joined to multiple neighbours is small, the shape of the molecule is essentially that of a linear backbone carrying a number of relatively short branches. In a system of this kind, the individual branched molecules can still be separated from one another without breaking chemical bonds and, as a consequence, materials consisting of branched macromolecules in most cases melt and dissolve reversibly. If, however, the number of multifunctional units in the individual chains becomes relatively large, each develops into a branch and the branches of different backbone chains join together to create a three-dimensional network in which backbone chains are no longer distinguishable from branches and in which all existing units are joined to a number of others by chemical bonds. Systems of this type are called polymeric networks, or cross-linked polymers. Materials in which this type of molecular bonding is preponderant are nonmelting and insoluble, although they may soften and swell to a certain extent when heated or exposed to certain solvents.

The third step in the description of the molecular structure of a polymeric material is the elucidation of the chain structure or the method of joining of the individual units, or monomers, within the chain.

Depending on the chemical character of the repeating unit, many different types of chain formation are possible even in the case of a linear chain.



(1) (2) (3)

A repeating unit of structure (1) (in which X is any constituent) may, for instance, be joined in two ways to give the structures (2) or (3). Structure (2) represents a "head-to-tail" juncture of the two units, whereas (3) is a "head-to-head" union. Most of the better known vinyl-type polymers have molecular chains composed substantially of units joined in head-to-tail fashion. It also is possible, however, for polymer chains to be composed exclusively or in part of units joined head-to-head.

Spatial arrangement in single carbon chains

In a single chain of carbon atoms joined by single bonds, each carbon atom that carries a substituent X is asymmetric—that is, it can exist in either of two forms, or configurations, termed *d* and *l*, which are mirror images of one another (like right and left hand gloves). The various forms of polymer chains that exist depend on the relative configurations of the asymmetric carbons in the chains. If all carbons in a head-to-tail chain, as in structure (2), have the same configuration, the polymer is called isotactic; if the configuration alternates regularly along the chain from *d* to *l*, the molecule is called syndiotactic; and finally, if there is random alternation between the two possible configurations, it is called atactic. (For a discussion of tacticity see below *Classification and nomenclature*.)

Although the structural unit (1) is derived from a particularly simple type of monomer, $H_2C=CHX$, it leads to the various possible structures just described. Even more possibilities arise from such monomers as $H_2C=CX-CH=CH_2$ or $XHC=CHY$. Polymer stereochemistry, the discipline that deals with these subjects, is of considerable practical importance, because polymers of different chain structures have distinctly different properties, even when they are derived from the same monomer.

Polymer morphology. Experience also has shown that the ultimate, practical properties of all polymeric materials depend not only on their molecular design—size, shape, and chain structure—but also to a considerable extent on the supermolecular character of the system. This quality, also known as texture, or morphology, of the polymer, is influenced chiefly by the fact that all macromolecules can assume a large number of conformations or spatial shapes, which result from free rotation about single bonds in the chains. These conformations include various helical (spiral) forms, as well as random tangles or coils.

To analyze the supermolecular structure of polymeric systems, it is necessary to focus attention not on any individual macromolecule but on a unit of volume within the investigated object—fibre, film, or molded structure—which contains many macromolecules. In general, a cube with 200-angstrom unit (1 angstrom $= 10^{-8}$ centimetre) edges contains many hundreds of polymer molecules and, as a consequence, many thousands of molecular segments. Experience shows that these segments can have several characteristic mutual relationships inside a volume element: (1) they can be randomly distributed in regard to their spatial orientation regardless of whether they belong to the same molecular chain or to different chains; (2) they can possess a degree of lateral (side-by-side) order through the folding of the individual chains; and (3) they can possess lateral order through the parallel arrangement of extended chains.

If all volume elements in the material belong to group (1), the material is an amorphous solid, or glass, which is homogeneous and isotropic (structurally uniform in all directions). It also is noncrystalline and unoriented. Systems of this type usually are obtained when a polymer melt is solidified rapidly by quenching (cooling by sudden immersion).

Orientation and crystallinity

If all volume elements in a sample belong to group (2), the polymer is composed of single crystals, which can be either randomly distributed in space or in a preferred arrangement with respect to a fixed axis or plane. In the first case, the system is unoriented; in the second case, it has a definite orientation. Systems of this type are obtained when a polymer is precipitated slowly from a dilute solution or solidified slowly from a melt.

If all volume elements belong to group (3), the sample consists of extended chain crystallites—that is, crystalline areas or domains—which, again, may be unoriented or may possess particular spatial orientations. Systems of this type are obtained when concentrated solutions or melts are solidified while under the influence of mechanical forces, such as shear or stress.

In practice, any actual polymeric material—in the form of a fibre, film, sheet, or tube—represents a mixture of the three types of volume elements, and its macroscopic character is defined empirically by such expressions as "degree of orientation," "degree of crystallinity," or "later order." It is often difficult to decide whether any given sample has its segments laterally ordered in an intramolecular fashion by means of chain folding, or intermolecularly by way of extended chain crystallization. Extensive experimentation with different methods often is required to discriminate between the two. The degree of orientation is determined quantitatively by optical birefringence (splitting of light waves) or by the measurement of Young's modulus (a measure of elasticity), both properties being dependent on the relative degree of orientation of the chain. The degree of crystallinity, or lateral order, is determined qualitatively by such methods as X-ray and electron diffraction (interference effects observed with X-rays and with beams of electrons) and density, swelling volume, and isotope exchange, all properties influenced by the arrangement of the molecular chains.

Polymeric materials produced by spinning and related processes, with subsequent mechanical drawing or rolling of the product, consist of more or less oriented mixtures (50–90 percent orientation) of both amorphous and extended-chain crystalline domains (25–85 percent crystallinity). Larger pieces, made by extrusion, injection, or compression, have a typical spherulitic (composed of small spheres) texture containing all three possible segment organizations; the overall degree of orientation is usually low (less than 20 percent), the content of laterally ordered domains is high (up to 95 percent), and chain-folding crystallization is the predominant mode.

CLASSIFICATION AND NOMENCLATURE

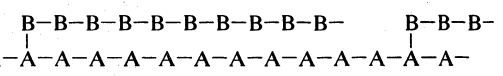
A homopolymer consists of macromolecules formed either by a single type of unit or by two or more chemi-

cally different types in regular sequence. Homopolymers may, and frequently do, contain irregularities in minor amounts, especially at the ends of the chains and at junctions leading to branches.

Copolymers are macromolecules containing two or more chemically different monomeric units in a more-or-less irregular sequence. Normally, the relative numbers of the different types of units are not the same in different molecules, nor even in different portions of a single molecule. A mixture of different macromolecules is called a polyblend. A block copolymer is a type of copolymer in which distinct groups of monomeric units of each type are linked together by chemical bonds and then interspersed along the chain. If, for example, the monomer units are designated A and B, a block copolymer of them might have the following structure:

$$-A-A-A-A-A-B-B-B-B-A-A-A-A-$$

A graft copolymer, on the other hand, contains branches of varying length composed of monomer units of one type joined to a backbone chain made up of monomer units of a second type. Such a copolymer might be composed as follows:



In their structure and in their properties, graft and block copolymers represent an intermediate form partway between normal copolymers and polyblends.

Tacticity

A tactic polymer consists of macromolecules whose monomeric units follow one another along the chain with their steric configurations (spatial orientations) ordered according to some pattern. The pattern, or tacticity, of the polymer, may consist of one or more simple rules. A eutactic polymer is a tactic polymer whose monomeric units are completely ordered, so that it is devoid of any element of structural disorder. An atactic polymer, on the other hand, is one in whose molecules the spatial configuration of the units is purely random. By way of example, a vinyl polymer prepared by conventional processes is generally atactic; it may possess a regular head-to-tail arrangement, but it is disordered with regard to the steric configurations of the tertiary carbon atoms that follow each other along chains. In isotactic polymers the asymmetric units are ordered in such a way that there is a continual, exact repetition of the unit configuration along the chain. A syndiotactic polymer consists of macromolecules whose monomeric units are asymmetric and ordered in such a way that there is a continual inversion of steric configuration from one monomeric unit to the next down the chain.

Polyelectrolytes are substances that dissociate in water (or other ionizing solvents) to give a polymeric material bearing a great many ionic, or electrically charged, groups, together with an equivalent number of counterions (ions of small charge and opposite sign). Polyelectrolytes can be polyacids (if they release hydrogen ions), polybases (if they release hydroxide ions), polysalts (if they release small cations or anions), or polyampholytes (if they release both hydrogen and hydroxide ions).

Under certain conditions chemical bonds may form between growing polymer chains. Such bonds, called cross-links, form whenever the monomeric units contain functional groups capable of reacting with one another, over and above those consumed in the initial polymerization reaction. The initial result of cross-linking is the formation of small domains, within which the macromolecules consist of random three-dimensional networks of chemically linked chains; such domains are called microgels. If a cross-linking reaction extends over macroscopic dimensions, and if the polymerizing system is taken up by a liquid solvent (or liquid monomer), the result is a gel, or jelly-like material. Drying converts such a gel into an insoluble and nonmelting material, composed entirely of a three-dimensional network at the molecular level. Such thermosetting resins as phenol–formaldehyde and urea–formaldehyde resins and copolymers of styrene and divinylbenzene fall into this category. Similar insoluble and infusible systems are obtained under certain conditions when a vinyl derivative is polymerized in the presence of a small amount of a divinyl compound. The materials thus obtained are called "popcorn" polymers; their exact structure has not been completely elucidated.

POLYMERIZATION PROCESSES

There are many processes by which small molecules can be bonded together to form large ones. Classification of these processes usually is based either on the chemical mechanisms by which the monomeric units are joined together or on the experimental conditions that exist during the reaction. Classifications by reaction mechanism distinguishes between polymerization by addition and polymerization by condensation.

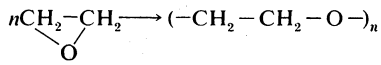
Polymerization mechanisms. *Addition.* Typical cases of addition polymerization are:

1. The addition of small molecules of one type to one another, accomplished by the opening of a multiple bond and without elimination of any part of the molecule, also called vinyl-type polymerization (in the following equations n is any number):

Vinyl-type polymerization

$$n\mathrm{H_2C{=}CHX} \longrightarrow (-\mathrm{CHX{-}CH_2}-)_n$$

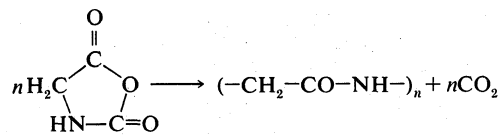
2. The addition of small molecules of one type to one another by the opening of a ring and without elimination of any part of the molecule (epoxide-type polymerization):



3. The addition of small molecules, all of one type, to one another by opening of a multiple bond with elimination of part of the molecule (aliphatic diazo-type polymerization):

$$n\mathrm{RCH{=}N_2} \longrightarrow (-\mathrm{CHR}-)_n + n\mathrm{N_2}$$

4. The addition of small molecules of one type to one another by opening of a ring with elimination of part of the molecule (α-aminocarboxyanhydride-type polymerization):



5. The addition of small molecules of one type to those of another type by opening of a multiple bond (polyurethane, polyurea, and polyamide formation):

$$n\mathrm{HOCH_2CH_2OH} + n\mathrm{OCN(CH_2)_6NCO}$$

$$\longrightarrow [-\mathrm{OCH_2CH_2OCONH(CH_2)_6NHCO}-]_n$$

6. The addition to one another of biradicals of a single type formed by dehydrogenation (poly-*p*-xylylene polymerization):

$$n(\cdot\mathrm{H_2C{-}C_6H_4{-}CH_2}\cdot) \longrightarrow (-\mathrm{CH_2{-}C_6H_4{-}CH_2}-)_n$$

Condensation. A number of possible forms of condensation polymerization are also observed. Typical cases are as follows:

1. The formation of polyesters, polyamides (nylon), and polyethers by elimination of water, or alcohols, from single bifunctional molecules, such as ω-hydroxy- or ω-aminocarboxylic acids or from such combinations as glycols or diamines with diesters or dicarboxylic acids. The reaction of a glycol with a dicarboxylic acid, for example, proceeds as follows:

Polyester and polyamide formation

$$n\mathrm{HO(CH_2)}_x\mathrm{OH} + n\mathrm{HOOC(CH_2)}_y\mathrm{COOH}$$

$$\longrightarrow [-\mathrm{O(CH_2)}_x\mathrm{OOC(CH_2)}_y\mathrm{CO}-]_n + n\mathrm{H_2O}$$

2. The formation of polyhydrocarbons by the elimination of halogen or hydrogen halides with the aid of metal and metal halide catalysts:

$$n\mathrm{ClH_2C(CH_2)_xCH_2Cl} \longrightarrow [-\mathrm{H_2C(CH_2)_xCH_2}-]_n + n\mathrm{Cl_2}$$

3. The formation of polysulfides by the elimination of sodium chloride from bifunctional alkyl or aryl halides with alkali sulfides or by the oxidation of dimercaptans (a type of organic sulfur compound). The former proceeds as follows:

$$n\mathrm{ClH_2C(CH_2)_xCH_2Cl} + n\mathrm{Na_2S_4}$$

$$\longrightarrow [-\mathrm{H_2C(CH_2)_xCH_2S_4}-]_n + 2n\mathrm{NaCl}$$

Experimental conditions. In actual practice the experimental conditions under which polymerization reactions occur are more commonly used to classify the reactions. Commonly, then, polymerization processes are classified according to whether they occur in the gaseous phase, in the pure liquid phase, or as dispersions, solutions, or emulsions.

Gaseous phase polymerization

1. Polymerization in the gaseous phase may occur at normal, reduced, or elevated pressures. The most important process of this type is the polymerization of ethylene, which is usually carried out at pressures of about 1,500 atmospheres in the temperature range between 190 and 210° C. Poly-*p*-xylylene is prepared at reduced pressure by cooling the pyrolyzed monomer from 900° C to room temperature.

2. Polymerization of single or of several monomer types frequently takes place in the pure liquid phase (bulk polymerization). Many vinyl-type polymerizations are carried out in this manner, particularly when large, transparent pieces (such as plates, rods, lenses) of the final product are desired. Polyester and polyamide formation are carried out almost exclusively in the bulk phase, as is the polycondensation of alkylsilanediols and the formation of phenol–, urea–, and melamine–formaldehyde resins.

3. Polymerization of one or more monomers frequently is carried out by dispersion in the form of small droplets in a liquid that does not dissolve the monomer. (This process also is referred to as suspension, bead, or pearl polymerization.) Styrene, methyl methacrylate, vinyl chloride, and other monomers are polymerized in this manner to give beads of very uniform size and quality for use in certain molding techniques.

4. Polymerization of one or more monomers may take place in solution. If the polymer is soluble in its own monomer and in the solvent selected for the reaction, the system thickens increasingly as the reaction proceeds. If the polymer is insoluble in the liquid constituents, a slurry forms as the reaction proceeds, and the polymer settles as a more or less swollen powdery mass. Vinyl chloride, acrylonitrile, vinyl acetate, and other monomers usually are polymerized and copolymerized by this technique.

Emulsion polymerization

5. Polymerization of one or more monomers often is conducted in emulsion. It has been found that polymerization frequently takes place in a particularly favourable manner if the monomers are brought together in the form of an aqueous emulsion with the aid of emulsifying agents such as soaps. The polymer or copolymer is then obtained in the form of a latex, or milky liquid, which either can be used as such or can be coagulated by the addition of acids to form agglomerates of the product polymer. Studies of the mechanism of emulsion polymerization indicate that several important steps in the process occur in the micelles (little particles) of the emulsifying agent. The final growth of the macromolecules takes place in polymer particles swollen by monomers transferred by diffusion from the monomer droplets. Most rubbery polymers and copolymers, and some plastics, are produced on a very large scale by emulsion polymerization.

Several other polymerization methods, such as polymerization in the solid state, in a fine powder, or in a foam of small bubble size, have been studied in the laboratory but have not been employed on a large scale for commercial use.

Most polymerization techniques can be carried out in separate batches or as continuous processes. Batch processing still prevails in most large-scale operations, but there is a tendency toward the development of continuous methods for all polymerization processes. Continuous processes have been developed for production of acrylonitrile, nylon, polyester, and certain synthetic rubbers.

Catalysis. Catalysts, substances that promote chemical reactions without being consumed in the reactions, play an important role in most polymerization reactions. Addition polymerization can be accelerated by three different types of catalysts: free-radical, cationic, and anionic.

Free radical catalysis. Initiation of a polymerization reaction may take place through the initial formation of a free radical, which either is produced by decomposition of a labile (*i.e.*, easily reactive) molecule or is formed in the course of a chemical reaction. Typical cases of free-radical-forming catalysts are peroxides, hydroperoxides, and azo compounds. Various ways of forming free radicals are indicated by equations 1–5:

$$\mathrm{C_6H_5-\overset{O}{\overset{\|}{C}}-O-O-\overset{O}{\overset{\|}{C}}-C_6H_5} + \text{heat} \longrightarrow 2\mathrm{C_6H_5-\overset{O}{\overset{\|}{C}}-O\cdot} \qquad (1)$$

$$\mathrm{C_6H_5-\overset{O}{\overset{\|}{C}}-O-O-\overset{O}{\overset{\|}{C}}-C_6H_5} + \text{heat} \longrightarrow 2\mathrm{C_6H_5\cdot} + 2\mathrm{CO_2} \qquad (2)$$

$$\mathrm{C_6H_5-CH=CH_2} + \text{radiation} \longrightarrow \mathrm{C_6H_5-\dot{C}H-CH_2\cdot} \qquad (3)$$

$$\mathrm{C_6H_5-CH=CH_2} + \text{heat} \longrightarrow \mathrm{C_6H_5-\dot{C}H-CH_2\cdot} \qquad (4)$$

$$\mathrm{\begin{matrix}H_3C\\ \diagdown\\ H_3C-C\\ \diagup\\ NC\end{matrix}-N=N-\begin{matrix}CH_3\\ \diagup\\ C-CH_3\\ \diagdown\\ CN\end{matrix}} + \text{heat} \longrightarrow \mathrm{N_2} + 2\mathrm{\begin{matrix}H_3C\\ \diagdown\\ C\cdot\\ \diagup\\ NC\end{matrix}} \qquad (5)$$

The initiator decomposes either by thermal action at a sufficiently high temperature or at low temperatures under the influence of a photon (packet of light energy) or an electron. The activation energy required to break the critical bond in the molecule of the initiator, usually a O–O, C–N, or C–O bond, amounts to only about 15–30 kilocalories. Some initiators, such as certain hydroperoxides and aliphatic azodinitriles, act at relatively low temperatures (40–50° C); whereas others, such as peroxides or aliphatic azodiesters, require temperatures as high as 150–175° C. The free-radical initiator (produced in reactions such as those shown in equations 1–5) attacks the double bond of the monomer, attaches itself to it, and reproduces the free radical at the other end of the adduct:

$$\mathrm{C_6H_5-\overset{O}{\overset{\|}{C}}-O\cdot} + \mathrm{H_2C=CHX} \longrightarrow \mathrm{C_6H_5-\overset{O}{\overset{\|}{C}}-O-CH_2CHX\cdot}$$

The propagation reaction

This process is called the propagation reaction; as it is repeated again and again, it consumes the molecules of the monomer to form macromolecules:

$$\mathrm{C_6H_5-\overset{O}{\overset{\|}{C}}-O-CH_2-CHX\cdot} + n\mathrm{CH_2=CHX}$$

$$\longrightarrow \mathrm{C_6H_5-\overset{O}{\overset{\|}{C}}-O(-CH_2-CHX-)_nCH_2-CHX\cdot}$$

The activation energies of typical propagation reactions have been found to be 5–10 kilocalories.

A free radical at the end of a growing chain can also react with a monomer molecule in such a way that it abstracts a hydrogen atom from the latter and transfers the free radical character from the growing chain to the monomer:

$$\mathrm{R-CH_2-CHX\cdot} + \mathrm{CH_2=CHX}$$

$$\longrightarrow \mathrm{R-CH_2-CH_2X} + \mathrm{CH_2=C\dot{X}}$$

Such processes are called chain-transfer reactions; they occur between the growing chains and molecules of the monomer or of the already formed polymer. If two free-

radical chain ends react with each other, there results a pairing of the lone electrons either by recombination or by disproportionation. Either process leads to the termination of the propagation step and finishes the polymerization reaction by the formation of stable macromolecules. Examples of both are:

$$\ldots CH_2{-}CHX\cdot + \cdot XCH{-}CH_2\ldots$$

$$\xrightarrow{\text{recombination}} \ldots CH_2{-}CHX{-}CHX{-}CH_2\ldots$$

$$\xrightarrow{\text{disproportionation}} \ldots CH_2CH_2X + XCH{=}CH\ldots$$

The interplay of initiation, propagation, and termination reactions leads to average lifetimes for the free radical-type chain ends of 0.1 to 10 seconds and to steady-state free-radical concentrations of 10^{-6}–10^{-8} mole per litre.

To produce free radicals at lower temperatures, such as room temperature or below, the interaction of electron donors, such as iron (II) ions, Fe^{2+}, with electron acceptors, such as hydrogen peroxide, H_2O_2, is utilized. In this way, an oxidation–reduction system is set up, which acts as a permanent source of free radicals:

$$H_2O_2 + Fe^{2+} \longrightarrow Fe^{3+} + HO^- + HO\cdot$$

Such systems are capable of initiating polymerization reactions at temperatures as low as −20° C. This reaction activation is of great practical importance in emulsion polymerization, where many undesirable complications can be avoided, or at least suppressed, by working at low temperatures.

Free-radical initiation is employed in the bulk polymerization of styrene and methyl methacrylate, in the solution copolymerization of vinyl chloride–vinyl acetate, in the suspension polymerization of acrylonitrile and vinyl chloride, in the emulsion copolymerization of butadiene–styrene and butadiene–acrylonitrile, and in many other vinyl-type addition polymerizations.

Cationic catalysis. Certain vinyl-type monomers, such as propylene, isobutylene, and most vinyl ethers, do not respond to free-radical initiation but can be initiated by catalysts that produce protons by interaction with a cocatalyst. The protons attack the monomer to form a carbonium ion (a molecular fragment with a positive charge on a carbon atom), which then acts as the carrier for the propagation. A typical example of cationic polymerization is the formation of polyisobutylene from isobutylene under the influence of boron trifluoride as catalyst and water as cocatalyst. The process can be represented by a series of equations:

Boron trifluoride as catalyst

$$BF_3 + H_2O \rightleftharpoons H_2OBF_3 \rightleftharpoons H^+ + HOBF_3^- \qquad (1)$$

$$H^+ + CH_2{=}C(CH_3)_2 \longrightarrow CH_3 - \overset{+}{C}(CH_3)_2 \qquad (2)$$

$$CH_3 - \overset{+}{C}(CH_3)_2 + nCH_2{=}C(CH_3)_2 \longrightarrow$$
$$CH_3 - C(CH_3)_2[-CH_2-C(CH_3)_2-]_{n-1}CH_2 - \overset{+}{C}(CH_3)_2 \qquad (3)$$

$$CH_3 - C(CH_3)_2[-CH_2-C(CH_3)_2-]_{n-1}CH_2 - \overset{+}{C}(CH_3)_2 + HOBF_3^- \longrightarrow$$
$$CH_3 - C(CH_3)_2[-CH_2-C(CH_3)_2-]_{n-1}CH_2 - \underset{\displaystyle CH_3}{\overset{\displaystyle CH_2}{\overset{\|}{C}}} + H_2OBF_3 \qquad (4)$$

The interaction of catalyst and cocatalyst produces protons (1), the concentration of which is controlled by an equilibrium that depends upon the concentration of the reactants and on the temperature. Other catalysts of this type are aluminum chloride and stannic chloride; other cocatalysts are ethers and amines. The proton adds to the monomer to form a carbonium ion (2), which propagates the chain by subsequent addition to other monomers (3). No formation of free radicals by splitting of an electron pair is necessary; instead, the displacement of an electron pair within the interacting particles suffices to bring about the chemical changes required to combine the monomeric units into macromolecules. As a consequence, the activation energies for initiation and propagation by cationic initiation are much lower than by the free-radical mechanism, and cationic polymerization leads to rapid formation of large molecules even at temperatures as low as −80° C.

The termination of cationic polymerization takes place by the reaction of the negatively charged complex ion with the carbonium-ion chain end, a reaction that results in the formation of a terminal hydroxyl group or a terminal double bond (4). Essentially this reaction restores the molecule of the catalyst to its original state so that it can attack another monomer and begin another reaction cycle.

Anionic catalysis. Anions, or negative ions, such as hydroxide (OH^-), amide ion (NH_2^-), and various carbanions designated R^-, can also initiate vinyl-type addition polymerization under certain conditions. The reaction proceeds by conversion of the monomer to a carbanion, which then undergoes propagation reactions with further monomers. This mechanism is of great importance, and it has been established that most polymerizations initiated by alkali metals or alkali-metal alkyls are of this type. In certain cases the lifetime of the carbanions is very long, so that there is essentially no termination reaction. Polymers formed under such conditions are called living polymers.

Living polymers

Most addition polymerizations involving ring opening are subject to catalytic influence by positively or negatively charged ions. Thus the polymerization of ethylene oxide and other epoxides is strongly accelerated by bases, and the polymerization of α-aminocarboxyanhydrides by acids. Both types are initiated by water, alcohols, or amines.

Coordination-complex catalysis. A very important group of ionic initiators is formed when compounds of transition metals—such as titanium, zirconium, and vanadium—are mixed with metal alkyls—butyllithium, $LiCH_2CH_2CH_2CH_3$, diethylzinc, $Zn(CH_3CH_2)_2$, and triethylaluminum $Al(CH_3CH_2)_3$, for example. In such cases the result is a solid precipitate, that is able to polymerize ethylenic compounds by a suspension type process. Polymerizations of this type take place very rapidly under mild conditions and lead to many products of great importance; *e.g.*, linear polyethylene, isotactic polypropylene, *cis*-1,4-polybutadiene, and ethylene–propylene copolymers.

Condensation catalysis. Most polycondensation reactions also are influenced by acids and bases. In these cases it is important to distinguish between equilibrium and rate effects. Certain materials, such as catalysts, influence the rate, or speed, of chemical reactions without affecting the relative amount of product formed; other substances affect primarily the amount of product present at equilibrium. With many condensation reactions, water-binding agents, such as phosphoric anhydride, sulfuric acid, or trifluoroacetic anhydride, very markedly increase the overall formation of polymer, but they do so by shifting the equilibrium in the direction of polycondensation rather than by a catalytic influence on the elementary process of chain propagation. As most polycondensation reactions are carried out at rather elevated temperatures (200° C or higher), it is necessary to employ acids and bases of very low volatility, such as antimony (III) oxide, Sb_2O_3, lead dioxide, PbO_2, calcium hydride, CaH_2, or potassium hydroxide, KOH.

PROPERTIES AND USES OF POLYMERS

Mechanical strength of polymers. The most important property of organic polymers, as compared with ordinary organic substances, is that they exhibit considerable mechanical strength in the solid state. Hence polymers are used to make fibres, films, plastics, rubbers, coatings, and adhesives of commercially valuable properties. The degree to which favourable mechanical properties are exhibited depends upon several factors, the interaction of which determines the most appropriate use of a given polymer and characterizes it as a typical material to form a fibre, plastic, or coating.

Degree of polymerization. One decisive factor for mechanical resistance is the degree of polymerization, or the

molecular weight, of the material under consideration. Extensive experiments with many polymers have shown that for each polymer there exists a certain critical molecular weight below which the material shows only poor mechanical strength. Above this critical value, mechanical strength is rapidly developed with increasing molecular weight up to a second level, above which the increase of mechanical resistance with increasing molecular weight becomes less pronounced. An industrially useful material must have, therefore, a certain molecular-weight value above which only a diminishing return of strength is gained. As a matter of safety, it is advisable to choose a molecular weight well above this point, but it has been found that materials with too high a degree of polymerization exhibit exorbitant viscosities in solution or in the molten state, which render such important processes as filtration, spinning, casting, and molding unduly difficult.

Optimum molecular weight

A proper compromise between easy processing ability and desirable ultimate properties must therefore be established by choosing an optimum molecular-weight range in each individual case.

Intermolecular forces. Another important factor for the development of characteristic mechanical properties is the magnitude of the intermolecular forces between individual macromolecules in a given material, as well as the structural regularity of the macromolecules themselves. If the intermolecular attraction is small and the chains are difficult to fit laterally into a lattice-like arrangement, then at a particular temperature and stress material that has been stretched shows a distinct tendency to return to a relaxed state in which the macromolecules are again curled rather than laterally ordered. Such behaviour is typical of rubbers (elastomers). On the other hand, if the molecular cohesion is strong and the fine structure of the chains provides for easy lateral fitting into lattice-like aggregates, a state of substantial crystallinity is favoured and the material is a typical fibre. In intermediate cases, in which the forces are moderate and the geometry of the chains is not very favourable for crystallization, the behaviour of the material will depend greatly upon external conditions such as temperature and pressure. Such systems are typical plastics.

These qualitative considerations show that rubbers, plastics, and fibres are not intrinsically different materials. Their differences are rather a matter of degree and result from the combined effect of the intermolecular forces between the long chains and the general tendency of the chains themselves to curl and fold, which leads to a mixture of crystallized and disordered portions in a given sample of the material. In general, it can be said that under given conditions, such as temperature and external forces applied to the sample, the tendency to crystallize is brought about by two different factors, which to a certain extent conflict with each other:

Intermolecular forces versus molecular geometry

1. Forces between the individual macromolecules: if these forces are strong, then at ordinary temperatures they chiefly determine the behaviour of the material, and it can be expected to behave like a hard plastic or like a typical fibre. This condition appears to be true for cellulose, certain proteins, nylon, polyesters, and polyacrylics.

2. The microgeometry of the molecular chains themselves: If the chains fit easily into a crystal-like lattice, as polyethylene, nylon, and certain other substances do, the material crystallizes under the influence of comparatively weak forces and hence, in general, has a tendency to appear more fibre-like than would be expected if the intermolecular forces alone were considered. On the other hand, materials with bulky and irregular macromolecular chains do not crystallize even if the intermolecular forces are quite strong. These are soft plastics and elastomers, and they behave more like rubber than would be expected on the basis of the intermolecular forces alone.

Commercially important polymers. From the practical point of view, classification and application of organic polymers is dominated by the raw materials (monomers) from which they are derived, because the monomeric units are responsible not only for the intermolecular forces and the structure of the individual polymer molecules but, in most cases, also for the cost of the final product.

Polyhydrocarbons. Chemically the simplest class of commercial polymers is the polyhydrocarbons. These are very widely used and cover the entire range from very soft and resilient rubbers, through plastics of all kinds, to strong, tough, and high-melting fibres. Considering all varieties and all applications, it can be estimated that in 1970 about 10,000,000 tons of polyhydrocarbons were produced and used commercially worldwide.

The most important commercial hydrocarbon polymer is polyethylene, which exists in two types: a lower density material, which is made by high-pressure polymerization, and a high-density type, which is made with the aid of special catalysts at normal pressure. Both are processed by various molding, extrusion, and casting techniques and are widely used to produce such products as trash cans, pails, containers for milk and most household liquids, shopping bags, and wrapping films for many dry goods. Polypropylene is somewhat stiffer and has a higher melting point and a more brilliant and transparent appearance. It is used for the same purposes as polyethylene but also, because of its higher melting point, as a fibre for carpets, upholstery, drapings, and, mixed with wool or acrylic fibres, for blankets. Polystyrene is brilliantly transparent and excellently moldable, but it has a somewhat lower softening point and it is more likely to swell in organic solvents than polyethylene or polypropylene. It is widely used for many household goods, as well as toys, pens, pencils, and decorative molded items. These three polyhydrocarbons all have relatively regular, easily orientable molecular chains, which can be readily brought into laterally ordered bundles resembling a crystal-like structure. They are used, therefore, as plastics and fibres, in spite of the fact that the intermolecular forces between the hydrocarbon chains are low. Other important polyhydrocarbons, however, with more bulky chains, which show more resistance to lateral order, have the properties of a typical rubber. Natural rubber—*cis*-polyisoprene—is the classical representative of these substances. Others are the principle synthetic elastomers, such as polybutadiene and copolymers of butadiene or isoprene with styrene and with each other. It is interesting that random copolymers of ethylene and propylene are also typical elastomers, because the irregular sequence of the two monomers along the chain reduces their capacity for the formation of laterally ordered semicrystalline domains.

Uses of polyethylene, polypropylene, and polystyrene

Chlorinated and cyanated polyhydrocarbons. Commercially important classes of polymers are the chlorinated and cyanated polyhydrocarbons. In these, the presence of a chlorine atom or cyano-group increases intermolecular attraction and has the consequence that regularly built chains of these materials behave like fibres or plastics. Outstanding examples are the successful synthetic fibres made of polyvinylidene chloride, polyvinyl chloride, and polyacrylonitrile. If irregularity is introduced into the molecular chains by copolymerization or by the use of bulky monomers, elastomeric materials are obtained. Examples are neoprene—a polychloroprene rubber—and the copolymers of butadiene with acrylonitrile, which form a large and important class of synthetic elastomers.

Uses of polyvinyl chlorides and polyacrylonitrile

Polyesters. Intermolecular forces between adjacent macromolecular chains can also be increased by the introduction of an ester bond into the backbone of the chain; this addition leads to the large and important family of polyesters. If the chains are regularly built and easy to align, typical fibre-forming polymers are obtained, such as polyethylene glycol terephthalate, which is today one of the most successful synthetic fibres. Polyesters with less regular chains form a large variety of plastics that are reversibly softening and behave like typical thermoplastics; such polyesters can be cross-linked and thereby made into hard, insoluble, and non-melting thermoset materials.

Polyamides. Still larger intermolecular forces are obtained when an amide bond is introduced into the back-

Uses of nylons

bone of linear macromolecules, producing the large and important family of polyamides, or nylons. These have the distinction of being superior fibre-forming materials, and in the late 1930s they initiated the introduction of synthetic fibres into the textile industry. Polyamides with reduced chain regularity produce excellent plastics, which are used in molding and extrusion processes. Polyacrylic compositions excel by their weather resistance, scratch resistance, high gloss, and capacity to accommodate pigments and other fillers. As a consequence, they represent an extremely successful class of materials for synthetic paints and coatings, widely used on products ranging from furniture to automobiles and airplanes.

Other polymers. Other polymers of some practical importance are: the polyurethanes as rubbers, foams, and adhesives; the polyfluorocarbons and the silicones because of their exceptional chemical inertness; the epoxy resins as coatings and adhesives; and several linear macromolecules composed of aromatic rings, because of their useful behaviour at temperatures as high as 600° C.

BIBLIOGRAPHY. Detailed information on this subject may be found in the multivolume *Encyclopedia of Polymer Science and Technology*, ed. by H.F. MARK, N.G. GAYLORD, and N.M. BIKALES (1964-). Textbooks on polymer science include: P.J. FLORY, *Principles of Polymer Chemistry* (1953); C.E.H. BAWN, *The Chemistry of High Polymers* (1948); F.W. BILLMEYER, *Textbook of Polymer Science* (1962); and J.K. STILLE, *Introduction to Polymer Chemistry* (1962). Important early technical works are K.H. MEYER and H.F. MARK, *Der Aufbau der hochpolymeren organischen Naturstoffe* (1930), and H. STAUDINGER, *Die hochmolekularen organischen Verbindungen* (1932), two works on high-molecular-weight compounds; H.F. MARK and G.S. WHITBY (eds.), *Collected Papers of Wallace Hume Carothers on High Polymeric Substances* (1940); and K.H. MEYER and H.F. MARK, *Hochpolymere chemie*, 2 vol. (1940; Eng. trans. of vol. 2, *Natural and Synthetic High Polymers*, 2nd ed., 1950). For current data, see the *Journal of Polymer Science* (irreg.); *Journal of Applied Polymer Science* (monthly); *Makromolekulare Chemie* (10-12/year); and *Macromolecules* (bimonthly).

(H.F.M.)

Polymorphism, Biological

Variation occurs in all species of living things and may be manifested in any character that the organism displays. It may be classified as genetic or nongenetic; *i.e.*, variation may be determined by genetic differences or by differences in the circumstances under which individuals live and are able to adapt. In a few characters, such as the blood groups, the differences are wholly genetic in origin, and no external influence can change them. Most characters, however, show both genetically and environmentally determined variation simultaneously. Furthermore, variation arising from genetic causes cannot be distinguished by inspection from that determined by environment. Thus an animal may be small because of the genes it has received or because it is undernourished.

Distinction between discontinuous and continuous variation

Genetic variation can be classified as discontinuous or continuous. In principle, discontinuous and continuous variation have the same genetic basis; discontinuous variation is usually controlled by unit genetic determiners (alleles) of one gene or a cluster of genes. The many genetic effects displayed by continuous variation, on the other hand, result from the cumulative effects of many genes, each exerting small but similar effects.

A discontinuous variation divides the individuals of a population into two or more sharply distinct forms. If the frequency of both forms is too high to be explained by mutation, the population showing the variation is said to be polymorphic. The most obvious example of a discontinuous variation is the separation of most higher organisms into males and females; another is the human blood types. In continuous variation, on the other hand, the individuals do not fall into sharp classes but are almost imperceptibly graded between wide extremes. This type of variation is found in virtually all characters of all populations and, if it is determined chiefly by heredity, can be altered relatively rapidly by selection. A familiar example is the smooth graduation of height among individuals of human populations.

A character may show both discontinuous and continuous variation within a population. Thus a human population, while showing continuous variation in height among its normal members, may simultaneously include sharply distinct dwarf or giant individuals.

General features. *Nongenetic variation.* Although the adaptive modifications in the effects of genes (phenotype) within a population that are commonly known as nongenetic variation do not involve genetic changes, the ability of a given species to produce several phenotypes is under genetic control.

Several types of nongenetic variation occur. One type occurs in adults of animal species that undergo seasonal changes in phenotype; *e.g.*, the plumage of certain birds and the fur of some mammals change colour with the season. The effects of physical environment, a type of nongenetic variation, are particularly pronounced in plants; *e.g.*, soil, amount of sunlight, and other factors can affect a plant's appearance. The phenotype of some freshwater fishes is dependent upon certain physical and chemical properties of the water they inhabit—e.g., temperature, acidity. The formation of genetically identical but morphologically different castes among social insects such as ants and bees is regulated by the nutrition of the larvae and other environmental conditions. Similarly, among a few species of Lepidoptera—*e.g.*, the European butterfly *Araschnia levana* and the African butterfly *Precis sesamus*—temperature or moisture differences may give rise to highly dissimilar spring and summer forms or dry- and wet-season forms.

Genetic variation. Because no two individuals are genetically identical in sexually reproducing species, genetic variation is important with regard to variations of the phenotype. Thus, almost any feature, be it morphological, physiological, or cytological, may vary. When the differences among individuals within a population can be divided into several classes based on the presence of certain characters, the variation, called polymorphism, is frequently controlled by one gene; however, the genetic requirements for polymorphism may be met by the segregation of a gene, a super-gene (a section of a chromosome containing more than one gene and acting as a unit), or a specialized pair of chromosomes.

A definition of genetic polymorphism

The nature of polymorphism. Genetic polymorphism, found throughout the animal and plant kingdoms, has been defined by the British zoologist E.B. Ford as

> the occurrence together in the same locality of two or more discontinuous forms of a species in such proportions that the rarest of them cannot be maintained merely by recurrent mutation.

Polymorphism provides variation within a species and may better its chances for survival. Characteristics within a species that may be polymorphic include colour, sexual features, blood groups, and chromosomal variations.

Two conditions that can lead to an increase in the relative frequency of a gene cannot lead to a permanent, or balanced, polymorphism. In one condition, the appearance of the gene (*A*, for example) increases through mutation at the expense of its alternative state, or allele (*a*), which is its less dominant counterpart. If the new gene has a harmful effect, it will be eliminated each time it appears. Should the new gene possess some advantage, however, it may eventually become dominant and may spread in the population, so reducing the alternative allele to the status of a rare mutant and thus extinguishing the variation. In the other condition, which occurs only rarely, a new gene of neutral survival value appears; such a gene is able to spread through a population only in a random way, and the process occurs so slowly that the delicate balance required for neutrality is upset by genetic and environmental changes.

A balanced, or permanent, polymorphism can arise in two ways. One is by frequency-dependent selection; the rarer gene of a pair is at a disadvantage when present in a high frequency, but subject to positive selection when at a low frequency. The other is heterozygous advantage, which is described below.

The significance of the heterozygous condition. The usual basis for ensuring a permanent polymorphism is one in which the best-adapted individuals are heterozygous—*i.e.*, with respect to one or more gene pairs they contain different genes, in contrast to homozygotes, who contain similar genes of a pair. The colour polymorphism of the marine crustacean *Tisbe reticulata* and of the fly *Drosophila polymorpha* both involve heterozygotes whose viability is superior to that of the homozygotes.

Origins of heterozygous advantage

A heterozygous advantage, which tends to arise automatically as a gene becomes an asset and spreads through a population, can originate in either of two ways. In one, a commonly occurring type of mutant gene, the so-called recessive semi-lethal gene, is responsible for a trait that is fatal or highly unfavourable to mature development of an individual. Since the harmful effect is recessive, it is not manifest unless the gene is homozygous for the trait. Because the other homozygote is less advantageous than the heterozygote, the latter becomes the superior genotype; *i.e.*, the presence of both the recessive gene and the dominant gene is superior to that of the two dominant or the two recessive genes. The second way by which heterozygous advantage can arise involves the multiple effects of genes; some of these effects are an asset to the organism, and others are harmful. In addition, advantageous qualities tend to become dominant and disadvantageous ones recessive. The heterozygote, as a consequence, eventually possesses advantages only; the two homozygotes, on the other hand, will have some advantages and some disadvantages. Apparently, heterozygous advantage tends to evolve as a gene spreads through a population, and, as a result, the polymorphism to which it gives rise becomes common.

Although polymorphism is generally associated with heterozygous advantage, this advantage is not a necessary outcome of the heterozygous state but evolves as indicated above. In studies of chromosome inversion (reversal of gene order) in *Drosophila pseudoobscura*, it has been found that the heterozygotes have an advantage over both homozygous states in the same population unless the parents come from localities far removed from each other. That is, the advantage of the heterozygote is not intrinsic but exists where it can evolve. The evolution of heterozygous advantage has been demonstrated in studies of a black form of the moth *Biston betularia* in London. Before 1906 there was no evidence of heterozygous superiority of the black form. In the succeeding 50 years, however, heterozygous advantage gradually evolved.

The maintenance of polymorphism. Genetic polymorphism may be controlled by either a gene or a super-gene. A specifically adapted chromosome or chromosomal segment may function as a control for polymorphism, as do the sex chromosomes.

The concept of a super-gene may be explained by using mimicry as an example of the numerous and very different characteristics often involved in adaptations of organisms. In this case, all of the genes controlling features such as colour pattern, shape, habit, and viability segregate as a group. The once-separate genes have been brought together by advantageous linkage and crossing-over (mutual exchange of equivalent loci between chromosomes) to produce a tightly packed unit of several genes sufficiently near each other on the chromosome that they segregate together. The controlling element in polymorphism is commonly such a super-gene.

In addition to segregation, a super-gene may be maintained as an inversion; that is, a region of a chromosome in which the order of the genes is reversed. In this case, crossing-over cannot occur if the region of the inversion is heterozygous because the alleles are no longer opposite their counterparts. All of the genes in such a region thus continue to be held together as a super-gene.

Types of polymorphism. *Colour.* In a snail species, *Cepaea nemoralis*, common to Britain and Europe, the shell colour may be yellow (greenish when the animal is within), pink, or brown. Shell-colour variations are controlled by genes at one locus on a chromosome. In addition, the shell may either be bandless or marked with one to five dark bands. The presence and number of bands is determined by multiple alleles—*i.e.*, two or more alternative factors with the same chromosomal locus and with an effect on the same genetic character. The two loci, for shell colour and banding, form one super-gene. Although the loci normally are so close as to appear inseparable, crossing-over has been shown to occur between them. The greenish or yellow shells predominate in grasses of lowland areas; brown shells are most common on the leafless floor of beech-tree groves. The unbanded shell is more common on a uniform background such as short turf; banded types predominate in the diversified conditions provided by mixed herbage. The difference in the relative proportions of such forms varies from 20 to 40 percent between populations of different habitats.

The snails are preyed upon by various birds, especially the song thrush, and by other animals. The song thrush carries the snails to stones and breaks open the shells. Evidence provided by accumulated shells reveals that predation is not random; *i.e.*, the snails with better natural camouflage are less likely to be eaten. In each locality, therefore, relatively larger numbers of the types least appropriate to the background are eliminated; the populations, however, do not become uniform, because the genes that influence colour pattern in *Cepaea* also affect its physiology, and they favour the heterozygote, thus maintaining the polymorphism; the relative numbers of the different colours, however, are determined by predation. Similar phenomena demonstrating the selective control of polymorphism in other snails with variously coloured and banded shells have been observed in Africa, islands of the South Pacific, and other regions.

Instances of colour polymorphism

Another striking instance of colour polymorphism has been mentioned previously, the industrial melanism, or "blackness," of certain Lepidoptera. Over 100 species of moths have acquired black forms in the industrial areas of Britain since the mid-19th century. Similar changes are also occurring in continental Europe and in North America. This colour difference is genetically controlled; the black form is almost always dominant and in some areas comprises 95 percent of the population of certain species. In all populations affected, however, a small proportion of the pale form persists as a result of heterozygous advantage, which tends to increase the proportion; a polymorphism is thus established.

All of the species involved in industrial melanism rest upon such surfaces as tree trunks, rocks, and fences with no apparent attempt at self-concealment. These forms are selectively eliminated by insectivorous (insect-eating) birds, which tend to destroy the insects that are the least camouflaged—*e.g.*, black forms on pale lichen-covered bark in rural districts and pale forms on soot-begrimed trees in industrial regions. Two conditions thus affect the survival of the moths—a heterozygous advantage that ensures polymorphism and selective elimination by birds.

The appearance of black moths in presumably rural areas may in fact be the harbinger of environmental pollution there. The effectiveness of man's efforts to minimize atmospheric pollution is indicated by the fact that a significant reduction in the proportion of melanic moths has gradually taken place where anti-smoke regulations have been introduced.

Polymorphism in Batesian mimicry

Mimetic. In mimicry of the Batesian type, harmless or palatable species mimic formidable or nauseous ones. In Müllerian mimicry, in which formidable or nauseous species resemble one another, uniformity tends to develop both within and between distinctly marked and warningly coloured species. One encounter with an unpalatable prey should thus have a widespread influence on a predator. Thus polymorphism is generally restricted to mimicry of the Batesian type.

Batesian mimicry may involve general features other than colour pattern. In butterflies, for example, shape, scent, or motion may be mimicked. The "deceiver" is often only the female. Striking similarities may be involved, with disguises so convincing that the insects themselves may be deceived—a male butterfly sometimes chasing a female of the wrong species. Such mimetic forms, controlled genetically by the segregation of genes

or super-genes, have evolved gradually, through selection; *i.e.*, the gene complex is favoured that evokes the appropriate deception most efficiently. The sudden appearance of one gene by mutation may give rise to a more convincing copy in districts in which the species serving as a model is present than in those in which the model is absent because the effects of a gene can be modified by selection acting on the gene-complex.

Polymorphic mimicry appears to provide the essentials for a condition of diversity; for, as a mimic becomes more common relative to its model—by means of the advantage of similarity—the advantage eventually is converted into a disadvantage because a colour pattern, which originally was a warning, comes to be associated by predators with something edible rather than inedible, thus favouring diversity of the mimic. Inherent in the situation is another condition capable of producing polymorphism, for the heterozygotes are superior in viability to the homozygotes. Such a situation illustrates the almost requisite evolution of heterozygous advantage in polymorphism (see also MIMICRY).

Behavioral. The brood parasitism of cuckoos is illustrative of polymorphism in behaviour. Different female cuckoos can lay eggs of differing appearance. These they introduce into the nests of appropriate species. There is an advantage, indeed a necessity, for a good copy since most host species will eject from the nest an egg that differs noticeably from their own. In fact, different females of one cuckoo species (*Hierococcyx sparveroides* of India) can lay eggs that differ in both colour and size in order to match those of two distinct host species.

The cuckoo, it seems, is able to choose the nests of correct species by means of behavioral patterns implanted at an early age; *i.e.*, the adult cuckoo is apparently attracted to a host species by sights and sounds encountered very early in life. Each female lays one type of egg, but she shows no tendency to mate with a male that is genetically appropriate in enabling her to lay a specific type of egg, nor does she show a tendency to have only one mate. The genes for egg colour pattern in cuckoos have apparently become transferred from the autosomal chromosomes (chromosomes other than sex chromosomes) to the Y-chromosome (the chromosome for female sex), which in birds, unlike mammals, is transmitted through the female line.

Polymorphism in behaviour, then, can be genetically controlled, as is the mimicry of butterflies and the early conditioning of cuckoos.

Chromosomal inversions in *Drosophila*

Chromosomal. Another aspect of polymorphism involves chromosomal inversions acting as super-genes. Such rearrangements of genetic material, which are known in the fly *Drosophila*, have been examined in the third chromosome of *D. pseudoobscura* and *D. persimilis*. All such inversions observed in *Drosophila* give rise only to physiological effects (*i.e.*, cryptic variation); the individuals thus look alike, and the polymorphism is detectable only by examining the giant chromosomes in the salivary glands of the immature flies (larvae). The occurrence of the polymorphism is controlled by selection, the heterozygotes having an advantage over either homozygote. The advantages and disadvantages of the different chromosomal inversions ensure that the frequencies of their occurrence correspond to environmental changes.

The frequency of occurrence of so-called standard chromosomes undergoes a seasonal cycle in *Drosophila pseudoobscura*. In one California locality, for example, the frequency of the standard chromosomes regularly decreases from about 53 percent in March to about 28 percent in June. The process is reversed in the hot summer period up to late August until the proportion characteristic of early spring is again reached. Some powerful selection mechanism is able to bring about these seasonal adjustments in a few generations, since *D. pseudoobscura* passes through only six to eight generations per year at the locality studied.

The chromosomal polymorphism of *D. pseudoobscura* provides an instance of recent evolution. The so-called Pikes Peak gene arrangement in this insect has always been common in Texas but was rare in California before 1946. Until then, only four instances of it had been encountered in some 20,000 chromosomes. By 1970, 8 percent of *D. pseudoobscura* in California carried the Pikes Peak gene arrangement. The change is perhaps caused by the evolution of some new gene complex favouring this gene arrangement.

Blood group. Cryptic variation in human blood groups provide further examples of polymorphism. The variation is maintained by a balance of selective advantages and disadvantages.

Several blood groups are known to be controlled by genes, and it may be assumed that the groups have evolved heterozygous advantage. In the O, A, B, AB series for which individuals are commonly typed, heterozygous advantage has not yet been fully demonstrated for the A and B types because the heterozygotes are not yet distinguishable. In matings between individuals who are both of the AB type, however, A, B, and AB genotypes can be distinguished among the offspring. In such cases, an excess of the AB group has been clearly established; it is achieved because prenatal mortality eliminates the homozygotes.

In view of the selective balance involved in the human blood groups, it has been possible to predict susceptibility differences between individuals of different groups to various medical disorders. For example, persons with blood type O are about 35 percent more likely to develop duodenal ulcers than are those of other blood types.

The study of protein radiation by means of electrophoresis (the migration of substances in an electric field) makes possible the observation of differences in the nature of the amino acid components of the proteins. The genetic loci that control amino acid variation of the proteins can also be distinguished. About 40 percent or more of the diversity observed is polymorphic. Studies indicate that many genetic loci are polymorphic in animals and in plants.

This type of protein variation generally produces only cryptic effects; but, since the genes have multiple effects, it may be expected that protein variation is sometimes associated with overt characteristics. These may indeed contribute to its persistence; it has been shown, for instance, that there is a correlation between protein variation and habit in a colony of eider ducks, *Somateria mollissima*, in Scotland. One of the egg-albumen protein types is significantly more common among the migrant ducks than the nonmigrant ducks in the colony.

Development of polymorphism. Particular genes may become an asset and spread through the population—sometimes promoting a balance of selective agents and leading to polymorphism—before the evolution of heterozygous advantage. This may be seen in the colour and banding of snails and in mimicry, both of which are themselves capable of maintaining diversity. A further example is polymorphism for colour pattern in the moth *Panaxis dominula* in a marsh near Oxford. The polymorphism is controlled by a pair of genes in which the two homozygotes and the heterozygote are visibly distinguishable. The female moths of each state prefer to mate with males of one of the other types, although the males show no such inclination. Because competition exists among the males, who are able to pair with a number of females, the rare forms are favoured, thus leading to diversity and establishing polymorphism. No trace of heterozygous advantage has thus far evolved in this case—probably because the polymorphism has developed relatively recently.

Polymorphism for colour pattern in *Panaxis dominula*

Control of the breeding system is an important factor in the development of polymorphism. Such control is exemplified in plants by heterostyly, a condition common in the Primulaceae (primrose family). In heterostyly, now recognized as a super-gene involving as many as seven loci, the respective positions of the male and female parts of the flower are transferred. Either the anthers (the pollen-bearing tips of the male part of the flower) are high and the stigma (the elevated tip of the female part of the flower) low—hence difficult for pollen-transporting insects to reach—or the positions are reversed. These types of primrose, the "thrum" and "pin," respectively, are

polymorphic and appear in approximately equal numbers in populations of the species affected. As pointed out by Charles Darwin, heterostyly makes easier the transfer of pollen by insects between plants with unlike types of flowers, thus favouring outbreeding, or the mating of unlike individuals, while a physiological mechanism ensures sterility when crossing between similar types does takes place. Conversely, a crossover within the supergene ensures that the male and female parts are formed at the same height (homostyly) and favours self-fertilization. Plants with the heterostyle-homostyle mechanism can thus vary between outbreeding with relative variability and inbreeding with relative uniformity to suit the ecological situation.

The relationship between polymorphic and multifactorial variation is important. In the latter, numbers of genes having small, similar, and cumulative effects operate, as in the genetic component of human height. This may allow a change to develop more easily than in the case of polymorphism, since most of the genes concerned in multifactorial variation are generally scattered in the normal population. This distinction occurs in the development of resistance to warfarin, a rodent poison. Such resistance is multifactorial in the mouse (*Mus musculus*); in the rat (*Rattus norvegicus*) it is caused by a single gene.

BIBLIOGRAPHY

General references: THEODOSIUS DOBZHANSKY, *Genetics and the Origin of Species*, 3rd ed. rev. (1951) and *Genetics of the Evolutionary Process* (1970); E.B. FORD, *Genetics for Medical Students*, 6th ed. rev. (1968) and *Ecological Genetics*, 3rd ed. (1971); R.A. FISHER, *The Genetical Theory of Natural Selection* (1930); MIRIAM ROTHSCHILD and THERESA CLAY, *Fleas, Flukes and Cuckoos* (1952).

Technical articles and reviews: B. BATTAGLIA, "Balanced Polymorphism in *Tisbe Reticulata*, a Marine Copepod," *Evolution*, 12:358–364 (1958); A.B. DA CUNHA, "Genetic Analysis of the Polymorphism of Color Pattern in *Drosophila Polymorpha*," *Evolution*, 3:239–251 (1949); THEODOSIUS DOBZHANSKY "Genetics of Natural Populations. XIX. Origin of Heterosis Through Natural Selection in Populations of *Drosophila pseudoobscura*," *Genetics*, 35:288–302 (1950); R.A. FISHER, "The Distribution of Gene Ratios for Rare Mutations," *Proc. R. Soc. Edinb.*, 50:204–220 (1930); E.B. FORD, "Polymorphism and Taxonomy," in JULIAN S. HUXLEY (ed.), *The New Systematics*, pp. 493–513 (1940), and "Polymorphism," *Biol. Rev.*, 20:73–87 (1945); H.B.D. KETTLEWELL, "Problems in Industrial Melanism," *Entomologist*, 90:98–105 (1957); H.J. MULLER, "Our Load of Mutations," *Am. J. Hum. Genet.*, 2:111–176 (1950), and "The Mutation Theory Re-Examined," *Proc. 10th Int. Congr. Genet.*, vol. 1, pp. 306–317 (1959).

Polynesian Cultures

Polynesia is the name used to cover a huge triangular area of the east-central Pacific. A triangle, each side of which is approximately 4,500 miles in length, it has its apex at the Hawaiian Islands and its western and eastern base angles at New Zealand and Easter Island, respectively. (See OCEANIC PEOPLES AND CULTURES for a map of the area.)

The basis for grouping the islands together is not their form and structure from a geographical point of view, for there is a good deal of physical diversity among the islands of the Polynesian group. They are grouped together because they are (in a few cases, *were*) inhabited by a people sharing a common basic cultural, linguistic, and racial heritage—the Polynesians. Although this heritage shows a measure of variation from one island culture to another, the underlying similarities are so marked and evident that the cultural relationship between the peoples of these far-flung islands is clear even to observers who have no training in anthropology.

This article will deal for the most part with traditional aspects of Polynesian society before the coming of the Europeans. The effects of this development are discussed in the concluding section.

THE PHYSICAL ENVIRONMENT

The physical environment of the islands of the Polynesian triangle is not as favourable for human habitation as it might at first seem and certainly presented difficulties when the ancestors of the Polynesians entered the area from the west, well before the turn of the Christian Era. The islands were apparently devoid of much that was needed for human habitation. Most of the food and useful plants and all of the domestic fowl and animals had to be transported by the Polynesian settlers themselves. Since that time, the physical environment has continued to exert a marked influence on the nature and extent of the evolution of Polynesian cultures by virtue of the constraints and opportunities that it provides.

Three Polynesian island types

There are essentially three main types of islands in the Polynesian triangle, each of which offers a different set of potentials and limitations to the inhabitants: coral atolls, "high," or volcanic, islands, and makatea islands.

The islands of the Tuamotu Archipelago are examples of coral atolls, each of which consists of a string of irregularly shaped and low-lying islets fringed by a reef, surrounding a central lagoon that may be as much as 20 to 30 miles (30 to 50 kilometres) in diameter. Endowed with a rich sea and shore fauna, atolls are generally impoverished in terms of land vegetation, lacking soil for planting and water.

High, or volcanic, islands (the Hawaiian Islands, the Samoan Islands, and Tahiti, for example) are volcanic in origin, produced by outflows of lava along faults and other lines of weakness in the sea floor in relatively recent geologic times. These islands display rugged terrain with remnants of extinct volcanic features, such as craters and volcanic plugs, and some live volcanoes. There is ample depth of soil for agriculture and abundant groundwater. The mountains of many of the volcanic islands are high enough to trap and hold rain clouds, resulting in good precipitation and lush vegetation. Usually, a series of vegetation zones can be identified on such islands as one proceeds from the beach uphill and inland to the island's central plateau.

Makatea islands (Makatea Island; Aitutaki in the Cooks, for example) are coral atolls that have been raised, wholly or in part, exposing the coral rock of the atoll high above the sea. Such islands generally have rugged surfaces and steep edges due to the differential erosion of the coral rock by wave, wind, and water. Soil may have developed in pits and hollows upon such surfaces, giving makatea islands a generally richer and more profuse vegetation than coral atolls.

New Zealand is unique in this grouping, for it is the only island in Polynesia of continental origin. It is actually an outpost island of the ancient continent of Australasia, which has been in formation since the Cambrian Period (*i.e.*, since at least 500,000,000 years ago). It has a wide range of environments, including high mountain ranges and glaciers, hot springs, lakes, dense forests, great rivers, and rolling grasslands. New Zealand also had the only indigenous terrestrial food animal of any real significance in Polynesia: the large, ostrich-like birds known to the New Zealand Maori as moa. This bird, a major resource for the early Polynesian settlers of New Zealand, was ultimately rendered extinct by the growing Maori population. New Zealand's uniqueness in Polynesia in terms of its geological history is enhanced by virtue of its position well outside the southernmost limit of the tropical climate zone. The differing climate, flora, and fauna of the New Zealand islands led to the development of the very distinct Maori version of Polynesian culture.

Climatic variations

The Polynesian triangle, lying astride the Equator, includes both tropical and nontropical climates. In the tropical section, the apex of the triangle lies in the northwest trade winds, while the southeast trade winds blow through the base of the triangle approximately 70 percent of the year, bringing rainfall to the southern and eastern coasts of the high islands particularly. For approximately 30 percent of the year northwest winds take over in the Southern Hemisphere. These winds, appearing at roughly the same time each year, played an extremely important part in Polynesian voyaging, for they enabled Polynesian mariners to sail eastward more easily. Much of tropical Polynesia lies in the area of heaviest rainfall in the Pacific, a zone having from 80 to 160 inches of rain per year.

This helped the early settlers considerably in establishing their native agricultural and useful plants. The doldrum region, with its high temperature and humidity and low wind speeds, cuts across the triangle in an east–west direction, presenting what must have been a somewhat imposing barrier to north–south travel with aboriginal vessels.

CULTURE PATTERNS

General characteristics. The popular image of polynesian culture is one of an almost blissfully simple and easy way of life, devoid of harsh extremes of any type, played out on islands of great beauty and natural abundance—in short, the exact opposite of much of Western culture. But the image, a product of popular literature and the motion-picture industry, is wide of the mark. Polynesian culture is both complex and highly specialized, and it abounds with extremes of all types.

The principal characteristic of Polynesian culture is its effective adaptation to the ocean environment in which it has flourished: no other culture of which there is record has ever so successfully mastered the sea. This mastery did not extend merely to the technology involved in ship building, to the techniques of navigation, and to other obvious nautical aspects of the culture but permeated social organization, religion, sexual behaviour, food production, and every other facet of the culture. The Polynesians could not only sail the seas, but socially they could cope with the human problems of shipwreck, split families, and the sudden loss of large portions of the social group. The food plants upon which they existed were all capable of withstanding long sea voyages, and their domestic animals were also easily portable. In short, the culture was well equipped to handle the numerous hazards that it had to face in the beautiful but actually quite hostile Pacific environment.

Polynesian conservatism

Another important characteristic of Polynesian culture is its high degree of conservatism and traditional orientation and the rather detailed structuring that places every man in a well-defined relationship to his society and to the universe.

The conservatism of Polynesian cultures is apparent even in a rapid examination of Polynesian materials from islands separated by hundreds or thousands of miles, whose populations were separated from the common Polynesian stock two or three millennia in the past. The similarity between such items as stone adzes and fish hooks is most remarkable. The same is true for kinship terms, plant names, and much of the rest of the technical vocabulary of the cultures, as well as for art motifs and medical preparations. A profound reverence for the past as embodied in the ornate and voluminous genealogies, chants, legends, songs, and spells that are passed down and elaborated through the generations provides a further illustration of this conservative tendency. The reason for this retentive nature of the culture is likely to be found in its successful adaptation to the Pacific island environment. There was no need to alter a technology or a social system so capable of coping with problems likely to be encountered.

The traditions of Polynesian culture provided a definition of a man's relation to his society and to all of nature. The creation myths told of the origin of the world, setting forth the order of precedence of Earth, sky, and sea and all the phenomena thereof, including man and woman. The genealogies fixed a man tightly into the typically aristocratic Polynesian social order. Other legends interpreted natural phenomena that were beyond his power to explain, while historical accounts often described, with varying amounts of mythological embroidery, the migrations of his people before they arrived at the island on which they were located, their adventures on the way, and the development of the culture following settlement.

The Polynesian was unusually well oriented, as would be expected in a nautical culture, to the entirety of nature. Their languages abound with terminology for stars, currents, winds, land forms, and directions; they are also unusual in their possession of a large number of grammatical elements indicating, for example, direction of motion implied by verbs, including movement toward or away from the speaker, relative positions of objects with reference to the speaker, and direction of movement along a seashore–inland axis.

It is one of the interesting contrasts of Polynesian culture that a way of life that displayed such a thoroughly realistic and practical exploitation of its environment at the same time exhibited a profound interest in the supernatural. This interest was directly related to the general interest in systematization and structure, the supernatural being for the Polynesian another portion of reality, albeit less manageable, that had to be structured and coped with.

Violence and cruelty were ever-present elements of Polynesian culture before and after contact with the Europeans. This is reflected strongly in Polynesian oral literature and in all aspects of life. Custom controlled and repressed direct physical expression of aggression within the kin group and the tribe up to a point, but there were definite limits beyond which only violence could restore status or assuage injured pride. Punishments for transgressing rules of behaviour toward chiefs and high priests or for violating various rituals incorporated ritual sacrifice and cannibalism as major features. Intertribal warfare was extremely common, particularly when population pressure had built up and land and resources were limited.

Perhaps the most publicized, and accordingly overrated, aspect of Polynesian culture is its sensuality: as in many other aspects of life, the Polynesians generally took a very direct, realistic, and quite physical approach to gratification of the senses. But there was no abnormal focus or concentration on any aspect of sensual gratification such as is seen in many other cultures where, for example, eating, drinking, or sex have become culturally important out of all proportion to their real significance and have consequently been the object of great cultural elaboration. This lack of elaboration in the areas of sensual gratification may be due in part to limitations imposed by the environment. The flora of Polynesia did not, for example, provide a very good basis for elaboration of cuisine, as the number of edible plants was rather restricted, and there was a dearth of spices. In general, however, this balanced approach to sensual gratification seems to be just another reflection of the Polynesian straightforward approach to the world. Polynesian culture, for example, is generally thought to have been characterized by extreme sexual promiscuity, but such is not the case. Definite restrictions were placed on sexual behaviour, though the limits of acceptable behaviour were indeed wider than in many other cultures of the world. Children were permitted freedom of sexual expression; from puberty on, sexual activity was strongly encouraged, limited only by incest restrictions (which were not stringent) and considerations of social prestige. Parents would become concerned if daughters were not sought after for sexual favours, and premarital pregnancy only enhanced a girl's attractiveness. Public intercourse was common during religious festivals and crisis rite ceremonies. There was, however, no pathological elaboration of sex for its own sake. Polynesian art, for example, is strikingly devoid of sexual content. There are occasional references made to sexual activity in native Polynesian oral literature, but many of these have to do with creative acts of the Polynesian gods, for the Polynesians saw the process of the creation of the universe as having taken place in a manner similar to human procreation. There is no corpus of Polynesian literature analogous to pornography. Sexual activity was an important part of certain religious ceremonies, but this was performed as a magical technique to promote general fertility rather than as an end in itself.

Sexual mores

Economic systems. *Settlement patterns and housing.* Two major settlement patterns were present in Polynesia prior to European contact: the hamlet and the village. Their origin and development may depend to a large extent on geographical factors, such as the distribution of food-crop resources, on defense considerations, and on social organization.

The hamlet settlement pattern was often found in the

larger volcanic islands, where food resources were diversified and scattered over the range of environmental zones that developed on such islands. A typical hamlet settlement pattern was often found in the Marquesas Islands. There, in prehistoric times as at present, the population spread over the floors and up the sides of the deep and narrow Marquesan valleys in clusters of perhaps four to five houses, often with gardens, taro patches, coconut trees, and breadfruit clustered in the immediate vicinity. The Marquesan houses were built on rectangular platforms, the height and composition of which depended on the prestige of the owner. Individuals of lower status might have a simple paved rectangle no more than a few inches high, while warriors, priest, or chiefs might live in houses perched on platforms seven to eight feet (about two metres) high and containing stones of several tons in weight. The house itself was built on a kind of a step running across the rear of the platform. Composed of a lashed and fitted wooden framework and with a thatched roof, the typical house was open all the way across the front and had square ends. The roof sloped from the high ridgepole directly to the platform floor in the rear. Inside, a polished coconut log often ran the length of the house, serving as a kind of community pillow. The floors were covered with mats, shredded leaves, or bark. Belongings were suspended in bundles from the rafters. Most of the household activity took place on the "veranda," or unroofed front portion of the platform, which was paved with smooth basalt stones from stream beds. Houses of chiefs and other high status individuals often made use of cut stone slabs for decorating the platform. Many also had rectangular pits in the platforms for storing fermenting breadfruit paste (an important delicacy) and small caches in which were interred the carefully cleaned and packaged bones of important family members.

Among the clusters of houses in each major valley section was a large rectangular ceremonial plaza, or *tohua*, which served as the focal point for all tribal or subtribal ceremonies, as well as a point for daily social gatherings. This complex of structures, often 600 feet (200 metres) or more in length, was generally built upon a huge artificial terrace carved out of a slope. Surrounding the plaza atop the terrace were the houses of the tribal chiefs and priests, temples and other sacred structures, and long sheds for spectators, At one end was usually found a temple with a sacred banyan tree in which were suspended the packaged bones of the illustrious dead of the tribe.

In Samoa, on the other hand, the village settlement pattern was followed after AD 1000 or so. There, fortified villages, consisting of 30 or more houses arranged in a network of paths, grew up along the coast. The houses at the time of European contact were built on oval mounds faced with rough stone slabs, although earlier houses had been built on rectangular platforms much like those of the Marquesas. The typical Samoan house was and is today a large shedlike structure without walls, oval in floor plan, with a beehive-shaped, thatched roof supported by a series of stout wooden pillars. Rolled mats along the eaves were let down to protect the inhabitants from sun, rain, or the night air. These houses, arranged in orderly fashion along the paths and sunken roadways that crosscut the villages, were surrounded by a fortification wall of stone or by wooden palisades.

Maori fortified villages

The Maori of New Zealand constructed large and impressive fortified villages (*pa*), surpassing those of all other Polynesian cultures. These forts were protected by ditches, palisades, trenches, and terraces. The interiors were partitioned off by additional defensive works so as to facilitate battle even after the outer defenses had been penetrated by an enemy assault. The houses within were of timber, rectangular in plan and generally dug a foot or so into the earth surface.

Other, unfortified Maori villages consisted of loose collections of houses arranged around a level area where ceremonies and meetings might take place. Forts were usually constructed in the vicinity of such villages for the use of inhabitants in time of war.

Production and technology. The Polynesians were above all mariners, but among the many seafaring peoples of the world they were quite unusual, for, far from having a disdain for the land and all that it involves, they were capable and devoted to horticulture and arboriculture.

The sea provided most of the protein in the Polynesian diet. Fishing was carried on not only on an individual basis, with spear, line, or net, but on a group basis with large numbers of men spreading and drawing in huge nets in bays or lagoons or driving fish toward shore, where they could be captured in nets held in shallow water. In some Polynesian societies (Marquesas and Samoa, for example) there were specialists who directed the mass fishing efforts and the elaborate religious rituals that went along with them. Sea mammals (porpoises and whales) were also taken, the former often being driven into shore by means of large, well-organized drives: when a school of porpoises was sighted in a bay, a fleet of canoes would seal off the bay entrance, and crew members would leap into the water carrying stones in their hands, which they beat together as they swam slowly toward shore. The sounds terrified the porpoises, and they beached themselves in a frenzy.

The Polynesian fisherman's exploits were not, however, confined to coastal waters, for he was just as much at home on the high seas. The waters around their island homes were explored for miles in all directions. The Easter Islanders, for example, used to travel to Sala y Gomez reef to fish—a journey of some 300 miles (500 kilometres). Good fishing places were located by reference to land bodies if visible or just by dead reckoning. Line fishing to depths of 90 feet was not uncommon to take the grouper. School tuna were taken in large quantities on the high seas by means of shiny pearl shell lures with bone points. The flesh of sharks and rays were delicacies.

Mollusks and crustaceans were important as food, and mollusks also provided the shells from which a wide variety of tools and ornaments were made. Clams, cowries, and various snails and conches were collected, generally by women and children, along the reefs or shorelines and in the shallows. Shrimps were netted with fine nets, while lobsters were collected by men who dove and pried their spiny prey loose from underwater crannies. The octopus was lured from its hiding place by an ingenious lure made of leopard cowrie shells. Various types of seaweeds and algae were also collected and were highly prized for their salty taste. There was, in short, nothing edible in the sea that was not food for the Polynesians.

Horticultural practices

Gardens and groves produced the staples of the Polynesian diet and most of the condiments. The major native Polynesian crops were yams, sweet potatoes, taro, breadfruit, bananas, sugarcane, coconuts, and chestnuts. These crops achieved different levels of importance in various Polynesian societies, depending on cultural factors and environmental conditions. The Hawaiians, for example, relied heavily on taro, building extensive irrigation systems to grow the variety that requires muddy soil and planting the "dry" variety in the uplands. Breadfruit was not of great importance in Hawaii, however. In the Marquesas and Tahiti, on the other hand, breadfruit was the major staple, although taro was by no means neglected. In these islands breadfruit was allowed to become overripe and was then beaten into a pulp and stored in large, well-drained pits in the ground, in neat hibiscus-leaf bundles. The breadfruit paste thus stored would ferment but remain edible and nutritious for years, its sour taste being highly prized for imparting flavour to the rather bland fresh breadfruit paste.

In the subtropical climate of Easter Island and in temperate New Zealand, the most important Polynesian food plants could not survive. On Easter Island, for example, the coconut did not survive, although legends maintain that germinating coconuts were brought by the earliest settlers. This deprived the Easter Islanders not only of a variety of condiments and a rich vitamin source but also of the leaves that were used elsewhere in Polynesia for thatch and baskets and of string, for most fishing line and netting in tropical Polynesia is made of plaited coconut-husk fibres, called *nape*. In New Zealand it was not only the coconut but also breadfruit, yams, and bananas that

did not survive. The radically different flora and fauna of New Zealand, however, offered new foods to replace, at least partially, those that it would not support.

Not only the coconut but most food plants in Polynesia also provided materials for other purposes. The breadfruit furnished wood for the hulls of dugout canoes, and the milky sap of the fruit itself was used in caulking the gaps between the planks of larger canoes. The black mud of the taro patches was used to stain wood and stone carvings. The pandanus fruit was highly prized as a kind of chewing gum, as well as a major component of floral necklaces and headpieces, to which it contributed its strong and pleasant colour. Pandanus leaves, when bleached, became the raw material for fine mats. The candlenut was used for torches; its oil was a cathartic, and its wood furnished certain canoe parts.

Polynesian technology can be circumscribed by five words: wood, stone, vegetable fibre, shell, and bone. The canoes, houses, domestic utensils, weapons, religious sculpture, and a host of other incidental tools were fashioned from wood with stone or shell adzes, stone-flake knives, files made of coral, sea urchin, or rough stone, and drills of bone, stone, or shells, Fine carving was done with stone, shell, or animal teeth, particularly those of rats or sharks. Stone, generally fine-grained basalt, was the hardest material available to the Polynesians and was used to produce a variety of adzes, the production of which was a demanding task in itself, for many broke at some point short of completion.

The absence of metal for use in attaching or clamping wooden parts posed no obstacle to Polynesian construction. Wooden components were skillfully fitted together and lashed with cordage made of various types of vegetable fibre, such as hibiscus bark, pandanus-leaf fibre, coconut fibre, or banyan bark. Huge double canoes 100 to 150 feet (30 to 45 metres) in length, for example, were built of numerous small wooden components all held together only by fitting and lashing, yet they were able to withstand the pounding of wind and wave for thousands of miles.

Vegetable material also furnished a major source of clothing in the form of the beaten bark (*tapa*) of the paper mulberry tree or the banyan. This material was beaten out into small sheets that could then be assembled to produce loincloths, capes, skirts, and headdresses. Bark cloth was decorated by painting and water marking and by attaching feathers, shells, teeth, and other ornaments. Fine mats were also used as apparel in western Polynesia.

Home furnishings also consisted mainly of vegetable material in the form of mats for sleeping and sitting, as well as for protection from the weather, and of baskets for holding personal belongings and food. Coconut shells and bottle gourds provided handy, durable containers for various items.

Pottery

Pottery was in use in western Polynesia at the time of European contacts. Its presence was originally thought to be a result of Fijian influence, but archaeological evidence now shows that pottery was in use before the Christian Era in the Marquesas Islands, at the extreme eastern edge of Polynesia, but subsequently it disappeared from the technology. This disappearance is easy to account for if one reflects on the lengthy and involved process required to produce a pot that was actually quite fragile, when there were durable natural containers (coconut shells and gourds) around in abundance and when wooden bowls of great durability could be produced for only a little more effort than that involved in making a pot.

Craft organizations are an interesting aspect of Polynesian culture. These organizations provided for the training of specialists in tasks requiring special knowledge and higher levels of manual skills—canoe construction, wood carving, and adz making, for example, as well as complicated operations, such as fishing and tattooing. When specialist groups existed, they were usually designated by name and would ply their craft for pay in the form of goods and services, as well as support for the period of their service.

Property and exchange systems. Personal property was a well-developed concept in primitive Polynesia. Each individual, regardless of rank, had his own tools, small canoes, clothing, ornaments, trophies, weapons, and other items and was generally free to dispose of them as he wished. Other types of property, however, were owned by kinship or residence groups in common and used for the common good. These included items too large to be produced or managed by a single man alone, such as a large double canoe or a fishing net several hundred feet in length, as well as facilities and land intended directly for community use, such as a tribal ceremonial ground, a refuge fortification, or a large breadfruit paste storage pit.

Land-holding practices

The matter of ownership of land and the means of production is complicated and varies among Polynesian cultures depending on the type of social organization that they developed. In Polynesian societies of the more traditional type, land was vested in a corporate kinship group. Among the Maori this corporate kin group, known as the *hapu*, usually inhabited one or a few villages located in close proximity to each other. The land of the *hapu* was viewed as being symbolic of the *hapu* chief's sanctity and power and as almost a literal extension of his body. Yet, the land was apportioned out to the various family groups within the *hapu* and worked by them. The chief, regardless of his prestige (stemming from a close family tie to the tribe's mythological ancestors), could not evict a family of nonproducers or in any way cut them off from their source of livelihood. He could invoke supernatural powers in the form of a curse of a ritual prohibition, but expropriation was quite uncommon. In return for this essentially inalienable right to his land, the good Maori was expected to provide certain quantities of produce for the chief on various occasions and also for the chief of the next highest element in Maori social organization, the *iwi*, which was composed of several *hapu*. If his offerings were not of sufficient size and quality, the Maori risked punishment and possibly even death in extreme cases but not expropriation. If the food offered was acceptable, however, the offeror would be assured of an appropriate amount of food of other types in return, demonstrating the chief's desire to live up to his own obligations.

A number of Polynesian societies were essentially similar to the Maori in their land-holding practices, including the Tikopia, Pukapuka, Futuna, and Uvea; but in others, evolutionary changes in social organization began to exert pressure on the basic land-holding kinship group (as exemplified by the Maori *hapu*), which seems to have been the basic building block of Polynesian society. As a result, the kinship group ultimately lost control of all rights to land, surrendering it to increasingly powerful and autocratic chiefs. Thus, in Hawaii, perhaps the most sociopolitically evolved of all the Polynesian societies, there was a large mass of completely landless commoners dwelling in territorial districts, each of which ran from the seashore to the island's centre, thus including a segment of each of the island's ecological zones. Each of these districts was ruled by a chief, somewhat like a lord in medieval Europe, who was responsible to the paramount chief of the island for supervising, among other things, the production of that district. The district chief or the paramount chief was free to shift at will the population of commoners within the district, regardless of their kin affiliation, although the chief and nobles themselves still relied heavily on their own kinship affiliations and prerogative. The district chief saw to the production of each household in his district and could evict the unlucky farmer who failed to meet his quota. Both he and the paramount chief could also make a variety of arbitrary demands from which the commoners had little or no protection, except to leave the district and submit to the authority of another district chief. The commoners were responsible for producing certain specified quantities of food at certain times, which was collected and passed up to the district chiefs and ultimately to the paramount chief. At that level, certain types of food reserved for chiefs only would be set aside, and a redistribution of the food would take place, with portions flowing back down the same channels that had been used to collect the

Economic redistribution in Hawaii

food in the first place. Kinship groupings, consisting of all individuals related through both male and female lines, still existed to be sure, but their main functions seem to have been to serve as a channel for low-level exchange of food and other commodities and to assist in the organization of the massive food collections held periodically throughout the year for the nobility.

The Polynesian system of exchange of goods and services may be summarized by two terms: redistribution and reciprocity, both already mentioned in connection with the discussion of the land-holding systems. The redistributive system was essentially a vertical system with goods moving up from the lower strata of the society to the chiefs and other high-ranking persons and then being apportioned and redistributed, so that all would share in more equal fashion in the productivity of every kinship group or region. This system involved a complex, shifting web of reciprocal obligations, which is very much at the heart of Polynesian culture, even today. Goods and services rendered, even if not requested, necessarily give rise to an obligation for a return in kind. There were no markets in Polynesian cultures, nor was there any standard medium of exchange. All exchange was in the form of barter, often under the general supervision of some senior family or kin-group member. Thus, for example, a portion of the fish catches made on a minor fishing expedition by coastal residents would be passed inland to residents at central villages, who might return dry taro for the fish at some fixed ratio of units of taro to units of fish. The services of any of the numerous specialists in Polynesian cultures (tattooists, fishermen, wood-carvers, for example) was also paid for by barter, usually over and above the cost of the specialist's keep during his period of service. Early European visitors to Polynesia who were able to analyze the importance of reciprocal exchange and put it to their own use generally fared quite well among the Polynesians, although they sometimes found themselves overwhelmed by the increasing size of the obligations they had undertaken. Capt. David Porter, United States Navy, for example, in 1813 won the friendship of the chief of Taiohae, Nuku Hiva, with a gift of sperm whales' teeth, but within a few months found himself fighting a tribal war essentially on behalf of the same chief, in order to live up to his obligations.

Social structures. *Local and territorial organization.* The two major types of settlement in Polynesian cultures —hamlet and village—were bound together by a web of kinship and territorial ties of varying degrees of importance.

There were two main and probably related types of kinship-based descent groups prevalent in Polynesian society. By far the most common and perhaps the most like the ancestral form of Polynesian social organization was the "ramage" type, in which the whole society might be represented in the form of a multibranched tree. There was a senior line of descent, in which descent passed from firstborn son to firstborn son and was traceable back to the mythological past. Branches off this main line were founded by junior sons, and these branches in turn produced further branches. Because the main line was the direct line to the gods and therefore carried the maximum traditional prestige and sanctity, it followed that each of the branches was ranked in terms of its proximity to the main line, and each individual in each branch could presumably place himself in a prestige-ranking scale relative to other members of his household settlement and kin group.

The various levels of the branches off of the main line assumed different levels of importance in various Polynesian societies. In New Zealand, for example, four levels were recognized: (1) the *waka* were a group of people descended on both maternal and paternal sides (male preferred) from the crew of one of the original canoes in the traditional Maori migration to New Zealand. They occupied a given territory and had a chief who functioned only in times of crisis and members who recognized their common tradition and interest. (2) The *iwi* were component parts of the *waka* but were made up of individuals tracing descent from a single occupant of one of the canoes, through either maternal or paternal line (again, the male line was the prestige line). The *iwi* had a chief and was a more tightly knit kinship group occupying a defined section of land within the *waka* territory. (3) The *hapu*, a component unit of the *iwi*, was the principal Maori descent group, consisting of all descendents of a given ancestor, usually sited in a particular village, and controlling the farming land around the village. The *hapu* also had a chief who was responsible for all economic and civil affairs. (4) The *whanau*, or households of the *hapu*, formed the lowest unit within Maori society and usually consisted of three to four generations.

Each of the hierarchical kinship segments just mentioned could be ranked relative to the others, with the exception of the *iwi*. The *hapu* could all be ranked within any given *iwi* on the basis of their chiefs' ancestry. Within the *waka*, each *iwi* could also be ranked in terms of the distance of its founding ancestor from the chief of the canoe.

This form of hierarchical branching-descent-group organization with territorial overtones was found in most Polynesian societies, with appropriate variations for local environmental conditions and cultural history.

Descent-line kinship groups

The other major form of kinship-based descent group found in Polynesia is that known as the "descent line." Descent-line organization appears to be the result of a breakdown in genealogical ties between the lower levels of a ramage organization. The descent line in Samoa, for example, consists of a group of people tracing descent in the male line from a commom mythical ancestor. This group was known as a *sa*. There was no concern for genealogical relationship of one descent line to another, nor was there any concern for ranking based on distance from or proximity to any particular male line of descent. What were passed down through the descent line were titles, each of which had rank and prestige attached to it. Each descent line held a number of these titles, which enabled it to participate in certain ways in the village council (*fono*). A number of descent lines were represented in each village council, and each descent line was spread through a number of villages in a given area. Within a given village, the senior title, determined by mythological connections, gave its holder the position of chief.

A characteristic of the descent-line system is its flexibility. Depending so heavily on myth and tradition for validation of status and title, it was only natural that ambitious titleholders could advance the prestige of their titles at the expense of others by various kinds of prestige behaviour, such as displays of wealth and power. The traditions governing title seniority could then be judiciously tampered with to produce the realignment that would allow the advance in status to occur. New descent lines might also be founded with similar "rewriting" of oral tradition.

Family and kinship patterns. The typical Polynesian family was an "extended" patrilocal patrilineal family—*i.e.*, a family consisting of more members than just parents and their children, living close to the senior male's people, and tracing descent through the male line. It consisted of a senior male and two to three generations of his male descendants and their children. Adoption, a very common custom in Polynesia, introduced additional children into the extended family, and these children behaved toward their adopted parents and siblings as they would toward their own. The flexibility of Polynesian social customs was such, however, that there were many variations on this patrilocal theme within and among Polynesian societies, depending on local conditions. There were also variations in lineality. In Tahiti and elsewhere, for example, inheritance could pass through females, if no male heirs were available. In other societies, more distant, collateral relatives and even slaves might be included in a household.

The basic Polynesian kinship pattern is essentially a very simple one: it distinguishes among generations, as might be expected in a society so strongly oriented on tradition and genealogy. There are sets of terms for: grandparents generation, distinguishing by sex only; par-

ents generation, in which parents are distinguished from various categories of aunt and uncles; members of a person's own generation, in which the terms permit identification by sex, relative age, and sometimes marital potential (marriage of certain cousin categories is preferred in some Polynesian societies); and, finally, for children, in which age and sex are again distinguished.

In general, the kinship terminologies in use in the various societies in eastern and marginal Polynesia are relatively simpler than those in the west.

Stratification. The tendency toward social stratification is perhaps inherent in Polynesian society, with its great interest in and emphasis on status and the indicators and prerogatives thereof. There seems to have been a process of social evolution occurring in Polynesia that tended to be somewhat unidirectional, leading toward more sharply defined differences in status and development of social classes that were clearly defined in terms of rights, duties, behaviour, and general life style.

The chief in ancient Polynesian society was the person of highest status, yet he was regarded by his people and generally conducted himself as merely "first among equals." In most of the more traditional of the societies, the chief could not appropriate the land of his followers, nor did he appear to be too interested in increasing his own group's holdings at the expense of neighbouring groups. In terms of his clothing and behaviour, there was little to distinguish him from other males. Nevertheless, he was the repository of sacred power for the group, a symbol of its tie with the past, and the vehicle whereby this tie would be perpetuated for coming generations. Even in the traditional Polynesian societies, however, it was possible for a man to rise in prestige by various achievements; by giving gifts, holding feasts, or displaying military prowess, for example. These activities seem to have contained the seed of change in the traditional social order.

Breakdown of traditional social order in certain Polynesian societies

There was a group of Polynesian societies in which the more traditional social order when breaking down gave way to one in which chiefs and other high status individuals no longer held office on the traditional genealogical basis but by virtue of having taken it, often by force, or as a result of some act that raised their prestige higher than that of the actual chief. In these societies the chiefs behaved more autocratically and set themselves off from the rest of the population by various kinds of privileges and behaviour. Changes in leadership were frequent, and kinship groups were split internally and reorganized in the course of internecine wars. Finally, out of this dynamic social situation emerged the most socially and politically advanced Polynesian societies (Hawaii, Tahiti, Tonga), with clear-cut social stratification and concentration of all power at the apex of the social pyramid—the paramount chief—below whom were several levels of nobility usually selected on genealogical relationship to the chiefly line. The paramount chiefs and upper classes had absolute powers over the commoners, inflicting cruel physical punishments, such as human sacrifice, for all manner of affronts to their dignities (touching a chief's shadow, for example). Warfare was frequent, well organized, and marked by acts of great cruelty. Religious and political powers were closely linked, with religious concepts and the gods themselves mirroring the cruelty and viciousness of the societies that had produced them.

Socialization and education. The Polynesian child was and is today generally born into a large and warm family environment. Even before the child can walk, it is turned over for care to the other children of the household (siblings, cousins), who generally associate in a kind of amorphous play group with children of other families. It is in this context that the Polynesian child receives a great deal of its socialization, first carried piggyback by an older sister and then running with a crowd. A particularly warm relationship exists between a child and its grandparents, often taking the form of rather gross bantering but providing the vehicle for inculcation of traditional love, technical training, and even sexual advice.

There was little by way of formal education in Polynesian society, except for the training in special crafts and skills, such as canoe making or tattooing, and the sacred academies for training the priest-bards who were the repositories of the society's traditions, mythology, and genealogies.

Rites of passage

Rites of passage varied in type and importance from society to society, but there were several that were generally held throughout Polynesia. The birth of a child was a matter of a great significance, particularly if the child happened to be a firstborn son of a high-status lineage. Various procedures were called for to announce the birth to the community, to the ancestors, and to the gods, and to care for the welfare (both physical and supernatural) of the infant and mother by application of medical and magical techniques.

Circumcision was a major gateway in the male life cycle and was appropriately marked by elaborate rituals, which increased in importance with the status of the male. Polynesians generally have an unusual abhorrence of the uncircumcised, to the extent that in some societies it was believed such people would automatically go to a very unpleasant netherworld after death. Although no such rite involving mutilation of any sort is reported for girls, some societies, such as the Marquesas, had orgiastic ceremonies in which adolescent girls made a more or less public debut, in a sexual sense.

There were other milestones in life that were marked with ceremonies in Polynesian cultures, such as the formal presentation of a royal heir, the completion of a tattooing operation or ear piercing in a high-status child, and the formal investiture of a priest or chief. The observances of these milestones were marked with a variety of rituals, quite often including human sacrifice, wanton killing of members of the population who might happen into the path of the ceremonial party, or members of the party who might make mistakes in performance of the rituals.

Death, the terminal milestone, was universally celebrated with extravagant behaviour, increasing in extravagance in direct proportion to the status of the deceased. In many societies these ceremonies were marked by violence, with mourners mutilating themselves and others, and the seemingly inevitable human sacrifices, obtained from within the social group or without. Feasts were also common, as was orgiastic behaviour. The extravagance of funeral rites was surpassed, in some societies, by ceremonies to deify a departed chief or priest. These went on for prolonged periods, involving prodigious feasting and drinking, violence, and sexuality.

Belief and aesthetic systems. *Religion and magic.* Religion and magic pervaded all aspects of Polynesian culture. All things, animate and inanimate, were believed to be endowed to a greater or lesser degree with *mana*, or sacred supernatural power. This power could be nullified by various human actions. Polynesian chiefs had great *mana*, so great, in fact, that in some islands if a commoner touched the chief's shadow the injury to the chief's *mana* could only be compensated by the death of the commoner. In many Polynesian cultures even today, it is still considered to be in very poor taste to step over a person's legs, pass one's hand over a person's head, or stand with one's head higher than that of a person of high rank. These actions are believed to sap a person's *mana*, rendering him profane.

It was not only people who had *mana*, however, but buildings, stones, tools, canoes, and all things. Polynesian life was loaded down with an often-bewildering variety of complicated rules designed to prevent damage to the *mana* inherent in various things. Groves, trees, temples, or tracts of land were considered sacred and could not be entered by ordinary people because they were pervaded by the *mana* of a high-status person or god. If anyone inadvertently stepped over a tool left on the ground, it was thus rendered profane and would often be discarded. Women were not permitted in canoes under normal conditions in the Marquesas because their presence defiled the canoe. In many societies, men preparing for war or for any other hazardous or demanding undertaking had to avoid the company of women, eat certain foods only, and often go into seclusion so as to protect their powers from

defilement. Some chants and songs were so sacred that every syllable had to be pronounced correctly. The penalty for major violations of these *tapu*, or prohibitions, was often death, swiftly executed and gruesome. Violations of lesser *tapu*, such as trespassing in a sacred grove or disturbing the bones of the dead, were visited with supernatural punishment in the form of illness, bad luck, or debilitation.

Mana was not the only problem that the Polynesian had to live with, however. The universe, in their beliefs, was peopled with spiritual beings of various types, many of whom were malevolent. In addition, there was a host of gods of various degrees of importance, ranging from the great gods of the Polynesian pantheon, such as Tangaroa, Tu, and Lono, to strictly local gods who were deified priests or chiefs of great renown. All of these gods had to be worshipped in their own way, the most important by full-time priests, of highest status, those of lesser importance by part-time priests. Worship of the gods involved sacrifices (including humans), chants and recitations, feasting (often with great prodigality), sexual orgies (to promote fertility), and elaborate rituals, often preceded by long fasting and abstinence.

Magic flourished in Polynesian society, with everyone practicing various rituals for success in love, war, planting, fishing, or just to bring misfortune to rivals. There were also magical specialists who could be resorted to when the problem at hand was too big and complicated to be solvable by ordinary magic. Such magical practices have survived in Polynesian cultures right down to the present time.

Abolition of *tapu* restrictions in Hawaii, 1819

The numerous restrictions and demands placed upon the average Polynesian by religion and the supernatural were in many cases seen as oppressive by the Polynesians themselves, and, in one case, the whole system was simply abolished. This occurred in Hawaii in 1819, when the young king, Kamehameha II, assisted by the high priest, his adviser, and other leaders of the court, publicly and intentionally broke the main food and eating *tapu* at a festival, after which the high priest desecrated the images of the gods and burned the temple. While many factors played a part in the prelude to this dramatic act (not the least of which was the presence of Europeans and the desires of the royal women to be free from the burden of *tapu*, which hit them the hardest), the principal factor motivating Kamehameha II seems to have been a real impatience and frustration at an increasingly constricting system that provided little obvious benefit in return for the great inconvenience. Although no other islands in Polynesia followed the Hawaiian example, there is evidence that the religious system was viewed as extremely frustrating in many other Polynesian cultures.

Artistic endeavour. Polynesian art was highly developed and brilliant, but little of what is known to have existed has survived down to the present day. This is mainly due to the perishability of much of the material (bark cloth, basketry, featherwork) and the dispersion of such items that took place during the era of European contact, when such items were bartered with European ship captains, soldiers, and collectors for firearms, liquor, iron tools, or various trinkets.

Each Polynesian society developed its own particular area of artistic endeavour—monumental stone sculpture on Easter Island and the Marquesas; highly decorated bark cloth in western Polynesia; fine mats in Tonga. In this development, a vocabulary of art motifs, styles, and artistic principles was elaborated, which also differ somewhat from culture to culture. There are, nevertheless, certain types of motifs that are widely distributed in Polynesia. There are, for example, a number of small geometric decorative elements, such as a toothed pattern or units of diagonally sloping lines, that are found in most cultures in many media. Stylized floral and animal elements are also widely distributed. In eastern and marginal Polynesia, an anthropomorphic figure and an often grotesque anthropomorphic face with bulging eyes and protruding tongue are prominent in all types of plastic art, as well as in tattooing.

Architecture as an art form was developed everywhere in Polynesia, but only the ruined stone structures of the eastern Polynesian islands remain to bear witness to the Polynesians' very considerable architectural skill. The stone temples of the Society Islands, Easter Island, and the Marquesas demonstrate the skill of the architects and masons. Even in their ruined state, many of these structures display striking proportions and design, clever use of stones of different colours, shapes, and textures and the evidence of sound combination of practical engineering with artistic objectives.

Architecture

EUROPEAN CONTACT AND ITS CONSEQUENCES

Integrating and disintegrating factors. The history of Polynesian culture since European contact is in many respects a sad repetition of what happened in all other parts of the world when a primitive society, well-adjusted to its own environment, was suddenly confronted with a vastly superior technology.

A number of factors internal to Polynesian culture had, in a sense, rendered it quite vulnerable to outside influences of a certain type. The political instability of the more evolved Polynesian societies, with social mobility gained by breaking down kinship ties and pushing aside traditional leadership, was a major weakness. The constant warfare that had developed as a concomitant of this political instability or as a result of overpopulation on some of the smaller islands in many cases crippled the economics of whole sections of a society, destroying cultural goods that had been produced at great cost, as well as constituting a serious chronic drain on the population. The emphasis on conspicuous consumption in lavish feasts and ceremonies caused a further drain on the economy, with huge quantities of food being consumed or wasted without heed for long-term needs.

Another series of vulnerabilities were in the area of health. Certainly, the Polynesians were not free from disease, but they had long been isolated from continental diseases by miles of cleansing sea and had thus developed no immunity to many types of disease then prevalent in other parts of the world. Certain of their customs, such as their relative sexual permissiveness, the practice of having the kava beverage prepared by mastication, and their affinity for this intoxicant, also predisposed them to difficulties in dealing with the health problems of advanced Western cultures.

On the other hand, some Polynesian societies were quite large and well-established systems with productive economies. The very flexibility and productivity that permitted the quest for prestige also enabled them to survive severe shocks, and their almost Oriental reverence of cultural tradition was a strong cohesive element.

The external factor that destroyed so many Polynesian societies and wrought such drastic changes in those that have survived was very simply Europeans and European culture. Their presence was a catalyst that set in action a chain of events that is still continuing to this day. Before colonization ever took place, whalers and other mariners were calling at many of the Polynesian islands to provision and refit. In the course of these calls, venereal and communicable childhood diseases were introduced, with often appalling results. Desire for the simplest products of European technology, such as nails, led many Polynesian societies to strip themselves of needed supplies in order to purchase the articles. Firearms, once introduced in quantity, only exacerbated the social and political instability, as did the colonies of European deserters and beachcombers who quickly formed on many islands.

Superior knowledge of firearms and tactics often made the Europeans indispensable allies of hard-driving warrior chiefs. Active missionary work began quite early, leading to systematic and almost total destruction of several cultures (in Magareva and the Marquesas, for example) and resulting in major alterations in most. Many missionary-introduced measures may have been beneficial; some, such as the puritanical insistence on full European style clothing for both sexes, were quite deleterious for health, and the stringent measures taken to segregate the sexes at all times led to an unhealthy psychological atmosphere regarding sexual activity. Full European

clothing was unsuitable for people accustomed to near or complete nudity: the heavy fabrics impeded respiration, raised body temperature, trapped bacteria, and caused chills when soaked with rain or perspiration.

Benefits of European contact

The European presence was not totally negative in its impact, however, for when the initial shocks of contact and colonization were past, European influence could often result in a stabilization of the social and political situation, with a concomitant diversion of resources and energies to more socially useful ends. (An exception was the Marquesas, where the French used Christianized Marquesans in combat against their still-pagan compatriots, further inflaming old tribal animosities.) Further, European presence often brought with it excellent medical assistance to stamp out health problems and correct certain customs and practices not conducive to health, such as the disdain for latrines displayed in some Polynesian societies.

Current cultural conditions. In the early 1970s, after nearly two centuries of heavy European contact, Polynesia is a patchwork quilt of cultural evolution. Southeastern Polynesia, particularly that part under French rule, is still very much between two worlds. The old ways are definitely gone, and so far, no good substitute has been provided by the ruling power. Conditions throughout much of the area are very spartan, and life is quite bleak. Tahiti, and particularly its sole town, Papeete, has been the beneficiary of a recent boom as a result of a rapid but rather badly handled tourist expansion and the use of the island as a staging area for the French nuclear test range, located some distance to the south, at Mururoa. This influx of foreigners has led to the development of a variety of "native culture," seasoned and packaged for tourist consumption. It can be seen everywhere in Papeete, in souvenir stores, art shops, and nightclub floor shows. In Tahiti, European conveniences such as motor bikes, pressure stoves and lanterns, radios, and television sets are common at all levels of society, and automobile use has increased dramatically in the last 15 years. European clothing, jewelry, and cosmetics are almost universal. One must go far from the town and off the main road to find anything representative of the indigenous culture.

Hawaii today represents what Tahiti will probably look like some years in the future, after land developers, the tourist industry, and industrialization have run their course. The 50th state of the U.S. still has a strong native Polynesian substratum in its culture, evident in language, the arts, dress, and customs. This substratum has been influenced, however, by a succession of waves of immigrants to the islands—including several European nationalities as well as Japanese, Chinese, and Filipinos—and, most recently, by organized crime, based apparently in the U.S. The result has been a strong "Orientalization" of much that was basically Polynesian. Definite attempts are being made to preserve and foster the Polynesian cultural heritage, including courses in universities and public schools in the Hawaiian language and the study of Polynesian cultures.

New Zealand, a member of the British Commonwealth, was far better endowed than its Polynesian compatriots with natural resources and has, in consequence, attained a reasonable degree of self-sufficiency as a political unit. The Maori play an important role in New Zealand culture as individuals, but the majority of them are in the lower income brackets, working in agricultural or livestock work in the country or in manual labour in the cities. Urbanization has caused social problems for the Maori, resembling those encountered in any area when the original proprietors of the land suddenly find themselves a minority, but social welfare programs of various types are being undertaken to better equip the Maori for his role in society.

The recent drive for independence consuming small colonies throughout the world has also hit Polynesia; Western Samoa (formerly a New Zealand protectorate), and Tonga (an independent kingdom under British advisorship) having decided to become totally independent in recent years. These societies have managed to retain the largest share of their native culture intact, including, in Tonga, the whole traditional political system. While cars, aircraft, and television have also invaded, the majority of the population live in traditional houses, wear modified traditional dress, and carry on many aspects of their subsistence activities, as did their forebears centuries ago.

Polynesia is thus being torn in many directions, one part standing on the threshold of the 21st century with the rest of the Western world, another still trying to recover its identity and purpose after the long period of colonization, which was largely destructive of its traditional ways. Meanwhile, still another segment is attempting to find a different path by maintaining much of its traditional culture and at the same time accepting some of the very new ways that are available to it.

BIBLIOGRAPHY. There is a wealth of published information on Polynesia, but much of it is to be found in the monograph series of the Bernice D. Bishop Museum, Honolulu, Hawaii, and in the publications of the *Journal of the Polynesian Society*, Wellington, New Zealand. Neither series is widely available. Many major works dealing with the area were published long ago, in very limited editions, and are not available in many libraries. PETER H. BUCK, *Vikings of the Sunrise* (1938), is quite dated but still provides an excellent introduction to Polynesian culture. No better survey presently exists. A fascinating account of Polynesian cultures as seen through the eyes of one of the greatest European explorers of the 18th century may be found in the Hakluyt Society's edition of the *Journals of Captain James Cook*, 4 vol. (1955–67). DAVID MALO, a courtier in the Court of Kamehameha I, published an interesting account of pre-European-contact customs in *Hawaiian* (1898; Eng. trans., *Hawaiian Antiquities*, 2nd ed., 1951). RAYMOND W. FIRTH has provided impressively detailed accounts of life on Tikopi, a Polynesian "outlier," in *We, the Tikopia* (1936) and in *Tikopia Ritual and Belief* (1967). The same author has also published on Maori social organization and economics in *Primitive Polynesian Economy* (1939). MARGARET MEAD, *Coming of Age in Samoa* (1928, reprinted many times), discusses the maturation process in the context of that Western Polynesian culture. BENGT DANIELSSON in *Den lyckliga ön* (1951; Eng. trans., *The Happy Island*, 1956), describes life, both past and present, on a Tuamotu atoll in French Polynesia. ALFRED METRAUX'S *classic L'Ile de Pâqueo* (1941; Eng. trans., *Easter Island*, rev. ed., 1957), presents a highly readable and nonsensational account of the aboriginal culture of that island based on research conducted there in the 1930s. Two recent summaries of Polynesian social organization from radically differing viewpoints are M.D. SAHLINS, *Social Stratification in Polynesia* (1958); and I. GOLDMAN, *Ancient Polynesian Society* (1970). The latter is encyclopaedic, summarizing data on archaeology, linguistics, and physical anthropology, and presenting summary portraits of 18 Polynesian societies representing the entire spectrum of social organization. It also has a very lengthy bibliography to guide further reading. R.C. SUGGS, *Marquesan Sexual Behavior* (1966), is the only book-length treatment of sex in a Polynesian society. Marshall discusses Mangaian sexual behaviour based on previously unpublished data in *Human Sexual Behavior*, ed. by D.S. MARSHALL and R.C. SUGGS (1970). Discussions of Polynesian art in the context of Oceanian art as a whole are found in A. BUHLERS, *Art of the South Sea Islands* (1960); J. GUIART, *Arts of the South Pacific* (1963); and P.J. WINGERT, *Art of the South Pacific Islands* (1953).

(R.C.Su.)

Polytheism

Polytheism, the belief in many gods, characterizes virtually all religions other than Judaism, Christianity, and Islām, which share a common tradition of monotheism, the belief in one God. Sometimes above the many gods a religion will have a supreme creator and focus of devotion, as in certain phases of Hinduism (there is also the tendency to identify the many gods as so many aspects of the Supreme Being); sometimes the gods are considered as less important than some higher goal, state, or saviour, as in Buddhism; sometimes one god will prove more dominant than the others without attaining overall supremacy, as Zeus in Greek religion. Typically, polytheistic cultures include belief in many demonic and ghostly forces in addition to the gods, and some supernatural beings will be malevolent; even in monotheistic religions there can be belief in many demons, as in New Testament Christianity.

Nature of polytheism. In the course of analyzing and recording various beliefs connected with the gods, historians of religions have used certain categories to identify different attitudes toward the gods. Thus, in the latter part of the 19th century, the terms henotheism and kathenotheism were used to refer to the exalting of a particular god as exclusively the highest within the framework of a particular hymn or ritual; *e.g.*, in the Vedic hymns (the ancient sacred texts of India). This process often consisted in loading other gods' attributes on the selected focus of worship. Within the framework of another part of the same ritual tradition, another god may be selected as supreme focus. Kathenotheism literally means belief in one god at a time. The term monolatry has a connected but different sense; it refers to the worship of one god as supreme and sole object of the worship of a group while not denying the existence of deities belonging to other groups. The term henotheism is also used to cover this case, or more generally to mean belief in the supremacy of a single god without denying others. This seems to have been the situation for a period in ancient Israel in regard to the cult of Yahweh.

Categories of attitudes toward the gods

The term animism has been applied to a belief in many *animae* ("spirits") and is often used rather crudely to characterize so-called primitive religions. In evolutionary hypotheses about the development of religion that were particularly fashionable among Western scholars in the latter half of the 19th century, animism was regarded as a stage in which the forces around man were less personalized than in the polytheistic stage. In actual instances of religious belief, however, no such scheme is possible: personal and impersonal aspects of divine forces are interwoven; *e.g.*, Agni, the fire god of the Ṛgveda (the foremost collection of Vedic hymns), is not only personified as an object of worship but is also the mysterious force within the sacrificial fire.

Belief in many divine beings, who typically have to be worshipped or, if malevolent, warded off with appropriate rituals, has been widespread in human cultures. Though a single evolutionary process cannot be postulated, there has been a drift in various traditions toward the unification of sacred forces under a single head, which, in a number of nonliterate "primal" societies, has become embedded in a supreme being. Sometimes this being is a *deus otiosus* (an "indifferent god"), regarded as having withdrawn from immediate concern with men and thought of sometimes as too exalted for men to petition. This observation led Wilhelm Schmidt, an Austrian anthropologist, to postulate in the early 20th century an *Urmonotheismus*, or "original monotheism," which later became overlaid by polytheism. Like all other theories of religious origins, this theory is speculative and unverifiable. More promising are attempts by sociologists and social anthropologists to penetrate to the uses and significance of the gods in particular societies.

Besides the drift toward some unification, there have been other tendencies in human culture that entail a rather sophisticated approach to mythological material—*e.g.*, giving the gods psychological significance, as in the works of the Greek dramatists Aeschylus and Euripides and similarly, but from a diverse angle, in Buddhism. At the popular level there has been, for instance, the reinterpretation of the gods as Christian saints, as in Mexican Catholicism. A fully articulate theory, however, of the ways in which polytheism serves symbolic, social, and other functions in human culture requires clarification of the role of myth, a much debated topic in contemporary anthropology and comparative religion.

Forms of polytheistic powers, gods, and demons. *Natural forces and objects.* A widespread phenomenon in religions is the identification of natural forces and objects as divinities. It is convenient to classify them as celestial, atmospheric, and earthly. This classification itself is explicitly recognized in Indo-Aryan religion: Sūrya, the sun god, is celestial; Indra, associated with storms, rain, and battles, is atmospheric; and Agni, the fire god, operates primarily at the earthly level. Sky gods, however, tend to take on atmospheric roles; *e.g.*, Zeus's use of lightning as his thunderbolt.

Celestial, atmospheric, and earthly gods

In the earliest cultural levels, in which hunting and then pastoralism and agriculture are clearly vital, religion exhibits these identifications in rites connected with fertility. The sun's vitality is seen in the cyclical effects of causing things to grow and wither. Moreover, because of its dominance of the world, the sun is often seen as all-knowing, and thus sky gods of various cultures tend to be highly powerful and knowledgeable, if also sometimes rather remote. The sky is also often associated with creation. By contrast the moon is rarely of the same importance (though in Ur, a city of ancient southern Babylonia, the moon god Sin was supreme). The role of the sky god in ensuring food and in providing light and warmth, over against the chaotic effects of darkness, was a theme of various myths of the cosmic drama and was one main reason for the connection in mythic thought between creation and light.

Heavenly divinities have also been influential in the development of astrology, which assigns a special significance to stars and planets. In the Middle East astrology was important but was weakened by monotheism; and in Indian culture it came to be deeply woven into the fabric of both Hinduism and Buddhism. Astrology was influential in the Greco-Roman world and in the astral religion attached to Gnosticism (dualistic sects that emphasized salvation through esoteric knowledge) and other cults of the early Christian Era. Astrology was also elaborated in Central America, for instance in Aztec religion.

Gods of the sky become especially powerful when they take on an atmospheric guise. The association of gods such as Zeus and Indra with storm, as well as with fertility-bearing rain, makes their connection with warfare fairly natural; thus, Indra is the most perfect example of an Indo-Aryan warrior. Many societies, however, have had separate gods of war. The ambivalence of atmospheric deities is paralleled in female counterparts who are both creative and destructive. The combination of sky and earth and the joining of differing cosmic forces are sometimes represented in the *hieros gamos* ("sacred marriage"); *e.g.*, between Apsu and Tiamat in Mesopotamia, Śiva and Śakti in India, and Gaea and Uranus in Greece. The forces of water and fire are particularly significant in bridging the gap between the earthly and heavenly realms. Fire is manifested not only in the hearth but also in lightning and the sun; water is sometimes connected with the moon. Thus, earthly fire and water can also be seen at work higher in the cosmos.

Important in the development of fertility religion were the "dying and rising" gods, such as Adonis, Attis, Osiris, and Tammuz. Their cults had a new life in the mystery cults of the Greco-Roman world, where the original agricultural significance of the rites was transformed into more personal and psychological terms.

On earth, besides the divine mother out of whose womb plant life has its birth, there are a host of divinities connected with agricultural and pastoral life. In addition, sacred significance is often attached to features of the particular environment in which a given group finds itself. Thus, sacred mountains, such as Olympus in Greece, have their resident deities; a river, such as the Ganges (Ganga), may be divinized. Underground rivers have special significance in connecting with the underworld, or nether regions, which can be important as the place of repose of the dead but also as the matrix for the re-creation of life. Geographical locations can also have cosmic significance; *e.g.*, Delphi, Greece, was known as the navel of the earth. Further, many cultures have gods and goddesses associated with the sea.

Vegetation. In a number of cultures trees are seen as a primordial form of vegetation and have a symbolic connection both with heaven and earth; sometimes they are held to contain spirits, as the *yakṣa*s of Indian tradition. Particular sorts of trees, such as the *aśvattha*, or pipal (sacred fig), are held in special veneration. Among plant deities, however, probably the most important are those connected with cultivated plants, such as maize in Central America and the vine in the Mediterranean world. Notable is the cult of Dionysus, the ecstatic wine

Plants and animals as divine forces

god who became one of the most influential objects of devotion in the classical period. The vine linked agriculture and ecstasy. The connection between vegetation and dying and rising gods has already been noted; to some extent such motifs were carried over into Christianity in the notion that the cross was the tree both of death and of new life. One of the most obvious modern survivals in the West of vegetation cults is the use at the winter solstice of mistletoe, symbolizing fertility and continued life.

Animal and human forms. Just as plants can be seen as divine forces, so can types or species of animals. For instance, the cult of the snake is widespread and is especially important in the Indian tradition. The serpent is vital in the Old Testament story of Adam and Eve and appears in the Babylonian *Epic of Gilgamesh* as one who knows the secret of rejuvenation. The snake has a fertility aspect because of its possible phallic significance and because it lives in holes in the life-giving earth. The cult of the monkey is important in India, having its essence in the figure of Hanumān, half monkey and half human. It is possible that such theriomorphic cults (in which gods are represented by various animal forms) have been assisted by rituals in which priests wear masks representing the relevant divinities, a practice that may in turn explain the hybrid half-human form. Examples of the wide variety of animal and living forms in which gods appear include Huitzlipochtli (hummingbird; Aztec); Cipactli (alligator; Aztec); Viṣṇu's avatars, or incarnations (fish, tortoise, boar, man-lion; Hindu); the Rainbow Snake (Australian Aboriginal); Cernunnos (stag god with antlers; Celtic religion); Nandi (bull; Hindu). A figure partly in animal guise found in Les Trois-Frères cave at Ariège, France, may represent a complex lord of the beasts analogous to the supposed Śiva (the destroyer and re-creator in Hindu mythology) figure found at sites in the Indus Valley; while a bird-man figure at Lascaux, France, may depict a priestly representation of a divine being. Thus, theriomorphism seems to have a very ancient pattern. In brief, various cultures have taken existing species in their environment and woven them into the pantheon—partly because of their essential dependence on the animals and partly for other reasons, such as similarities between animal forms and other sacred forces (*e.g.*, the analogy of the lion to the force behind kingship).

Gods as human and men as divine

Because man can enter into living relationship with the supernatural beings that surround and dominate his life, it has always been natural to model the gods as human beings. Such anthropomorphism is most evident in the Greek tradition, in which the Homeric gods are brilliantly and unashamedly human in their passions and thoughts. The human model has been assisted by the representation of the gods in art; for a statue is not just a symbolic representation of a god but often his place of presence and influence. Thus, in a number of cultures, the images are treated as replete with divinity.

Just as gods can be human in character, so men can be conceived as divine, either by becoming identified with deities (*e.g.*, through descent) or by displaying appropriate power. Thus, divine kingship was a not uncommon feature of the ancient Middle East; it was also found in the Roman world, when the emperors were divinized, and in Japan and China, where the emperor was son of heaven. Culture heroes and other significant humans could be elevated to semidivine status or more; *e.g.*, Kuan Ti and other heroes in the Chinese tradition, Rāma and Kṛṣṇa (Krishna) in India. Strictly, the succession of sages known as buddhas and tīrthaṅkaras in the Buddhist and Jain traditions, respectively, were not conceived as divine but came to be objects of a cult. In the Mahāyāna (Greater Vehicle), celestial buddhas and *bodhisattvas* (those vowed to become buddhas) came to be profoundly important for devotional religion; from a functional point of view, the Mahāyāna has operated as a polytheistic system, united, however, under an overarching doctrine of emptiness, or the void (*śūnya*), according to which all things are said to be empty of the characteristics assigned to them. The Theravāda (Way of the Elders) accepted the principle that virtuous followers of the Buddha could be translated in the next life to a heavenly existence in which they would have godlike status (an impermanent status, however, for gods share the universal transitoriness of all living beings), but such gods were scarcely the objects of a cult.

Functional deities. In addition to the various forces operating in nature, various social and other functions are divinized. Thus, the god Brahmā in the Vedic tradition, besides being creator, contains and expresses in personal form the power implicit in the Brahmin class. Again, there are gods of healing, such as Asclepius in Greece, and of seafaring, agriculture, and so on. The most elaborate reflection of human concerns is, perhaps, to be found in the later Taoist pantheon, which provided a heavenly counterpart to the Chinese Imperial court. In a number of societies gods of war, such as Mars (ancient Rome) and Skanda (India); gods of learning, such as Sarasvatī (India); and gods of love, such as Aphrodite (Greece) and Kāma (India), have been important. Even such abstractions as the directions (north, south, east, and west) have been divinized. The fact that these varied entities and relationships have been taken as gods is, perhaps, partly the result of the mythic style of thinking, in which distinctions between natural forces and social conventions are not clearly perceived.

Of special importance regarding human affairs are the gods concerned with death and judgment after death, such as Osiris in ancient Egypt, Yama in India, Hades in Greece, and Hel in pre-Christian Scandinavian religion. There are also gods associated with cemeteries and more generally with patterns of the disposal of the dead.

Supernatural forces

The various gods must be seen against the background of a whole host of spirits, demons, and other supernatural forces prevalent in the environment of pastoral and agricultural communities. Among entities hostile to man are the antigods, very often older gods, such as the Titans in Greece, who have been displaced by later deities or gods worshipped by a people conquered by a new dominant folk. The warfare between the old and new can be woven into dramatic myths of the fight between good and evil. This is well brought out in the major myth of the Orphic writings: Zeus's son Dionysus-Zagreus was killed and eaten by the Titans, who in turn were destroyed, burned up by Zeus's lightning flash. Man is made of the ashes, and therefore he is a compound of divinity and Titanic evil. Purification from this evil brings redemption and release from the round of reincarnation. Sometimes, however, the ambivalence of good and evil is built into the same deity, so that creation and destruction and good and evil are seen as complementing one another.

Types of polytheism. *Greco-Roman religion.* By the time of the establishment of the Roman Empire, the Greek tradition was already exerting considerable influence on the Roman, to the extent that once relatively independent traditions became somewhat fused. Equations between gods were freely made: Zeus became Jupiter, Aphrodite became Venus, and so on. Originally Roman *pietas* (sense of duty to the gods) was a good deal less personalized than the relationship to the anthropomorphic gods of the Homeric pantheon and was directed at spirits called *numina*. In addition, the various philosophical systems, such as Epicureanism and Stoicism, provided a more systematic cosmology and sense of man's destiny than traditional polytheism. Influential in the Hellenistic period were mystery cults—such as those of Isis, Cybele, Mithra, and Demeter—which catered more to personal concerns with salvation than did the official and civic cults. Under the mid-4th-century emperor Julian, a last vigorous attempt was made to revive paganism and to restore the cult of the gods over against the widespread grip of Christianity.

Germanic, Scandinavian, Celtic, and Slavic mythologies. The sources for a reconstruction of northern European religion are far better than those for the south Germanic peoples, but there were evidently similarities between the religions. The three main Scandinavian gods were Odin, Thor, and Freyr: Odin (or Wodan) had great magical power and wisdom and was called Allfather; Thor (or Donar) was the warrior god; and Freyr

was the god of fertility. It is possible that these gods are a reflection of the tripartite division of Indo-European society—priest, warrior, and cultivator. Among other deities, Balder, the dying god who was killed by a mistletoe branch, had a poignant charm. Nordic mythology also carries with it a sense of final doom of the gods, looking to the point when the world will be burned up, before its eventual re-creation.

The pattern of Celtic cults is not easy to decipher because of lack of written records; but the stag-headed god Cernunnos was highly significant in iconography. There was also a variety of ancestral gods and goddesses, including a "great mother" of the type found in fertility cults of the ancient Middle East. Celtic religion had a special reverence for water in such forms as pools and rivers.

The Slavic religions of eastern Europe and Russia are likewise imperfectly known, but they involved worship of a high god who is both a creator and an atmospheric force. Another important figure was the war god Svantovit. Finno-Ugrian pre-Christian religion bears some resemblance to the Scandinavian, possibly indicating some mutual influences, while Baltic cults are of Indo-European type.

Political considerations in Egypt

Egypt and the Middle East. The Egyptian pantheon evolved into a complex form; many deities were theriomorphic but were presided over by such great gods as Re, the sun god, and Nut, the sky goddess. Re's transformation as Horus, with a hawk's head, was connected with the Osiris legend. The pharaoh was identified with him as the "living Horus." Despite the attempt of Akhenaton, pharaoh in the 14th century BC, to exalt Aton as the single god, the Egyptian cult remained essentially polytheistic but highly articulated. With the domination of Egypt by the Ptolemies about ten centuries later, the worship of Sarapis, a hybrid Greco-Egyptian deity, was instituted as a means of binding together the two groups.

Though in Egypt the cause of the rise and fall of gods was partially the political struggles between the major city-states, the Sumerian religion was much less affected by such "earthly" considerations. An, the god of heaven, remained supreme, and such deities as the water god Enki and the air god Enlil were prominent. In Babylon, partly the successor state of Sumer, the most vital god was Marduk, creator of the world and of mankind, and victor over the primeval Tiamat, or chaos, who all but absorbed the older surrounding gods. His story is recounted in the epic *Enuma elish* ("When on High"). In Assyrian religion Marduk was in effect replaced by Ashur; and Ishtar, the mother goddess, was also important. In general, it can be said that Middle Eastern religion stemmed from early Sumerian and Egyptian sources and that the latter eventually had some effect on Hellenistic religion.

Early Indo-Iranian religions. For almost a millennium close relations existed between the Vedic and Iranian religions—from before the time of the Iranian prophet Zoroaster, who reformed the ancient religion in the late 7th and early 6th centuries BC, back to the time of the Vedic religion of the Aryans, who invaded India about 1500 BC. Zoroaster, in his reforms, succeeded in excising the many gods, some of whom were subsumed as qualities of the supreme Ahura Mazdā. The rich pantheon of the Vedic hymns developed into the world of classical Hindu mythology, which was fed by streams other than the Aryan.

Classical and modern Hinduism. Certain gods of no great importance in the Vedic tradition came to dominate classical Hinduism, above all Śiva and Viṣṇu. The latter was associated with belief in *avatāra*, or incarnation. Most male gods in the Hindu pantheon also came to be represented with a female consort, symbolizing the *śakti*, or creative power of the deity. The increasing elaboration of Hindu cults as different groups were absorbed into a systematized social fabric has led to the estimate of as many as 33,000,000 Hindu gods. It has been common practice for devotees to select the form under which the divine is worshipped, and such a deity is called the *iṣṭadevatā*. Most Hindus are inclined to interpret the many gods as being symbols of the one divine reality.

Buddhism. Buddhism's tolerance of popular cults, provided that the main essentials of the faith are maintained, means that in most Buddhist cultures several gods are worshipped. In Mahāyāna Buddhism, increased devotion to the Buddha became elaborated as a belief in many celestial beings, notably Amitābha and Avalokiteśvara (Kwannon), who were, however, in essence all unified in the absolute (*śūnya*, the void). In Tibet, a synthesis between the indigenous religion and Buddhism was established. The most notable feature of this form of Buddhism, known as Vajrayāna (Vehicle of the Thunderbolt), was the use of divine forms to symbolize the various factors of existence, such as the different elements making up human personality.

Far Eastern religions. In ancient China the cult of Heaven and ancestor worship were elements woven into the system of Confucianism. Numerous lesser deities were worshipped in popular Chinese practice, and the dividing lines between Confucianism, religious Taoism, and Buddhism were hard to draw. In Taoism, an elaborate pantheon was evolved, modelled in part on the Imperial bureaucracy, and was presided over by the Jade Emperor (Yü-Huang). Other deities included atmospheric gods, gods of locality, and functional gods (of wealth, literature, agriculture, and so on). The Taoist gods were in part a response to the richness of Mahāyāna myth, with its cults of celestial buddhas and *bodhisattvas*.

The Taoist pantheon

The religions practiced in China influenced Japanese culture, which took over some main elements of Confucianism and Buddhism, that interacted with the indigenous polytheistic religion, Shintō (Way of the Gods). The divinities of Shintō tend to be connected with natural forces and localities; the most important deity is Amaterasu, who is the sun goddess and divine ancestress of the emperor.

Religions of ancient Meso-America. The Aztec culture, successor of earlier civilizations, together with the associated Maya culture, laid great emphasis on astronomical observation and on a complex religious calendar. Important were the high god Ometecuhtli, the morning star Quetzalcóatl, and the various legends woven round Tezcatlipoca, patron of warriors, who in the form of Huitzlipochtli was patron of the Aztec nation. Inca religion also possessed a high god, Viracocha; a number of the most important deities were associated with celestial bodies, notably the sun, patron of the Incas. Both in Central and South America, the fertility aspects of deities were also emphasized.

Modern ethnic religions in Africa and elsewhere. In some areas, such as much of Africa and Oceania, the indigenous religions are ethnic or tribal; each group has its own particular tradition. These traditions have been affected considerably by the impact of Christian missions and Western technology. Clearly there is no single pattern of belief, though certain patterns do recur in some of the cultures, such as belief in a high god, totemism (characterized by recognition of a relationship between certain human groups and particular classes of animal, plant, or inanimate object in nature), spirit possession, and so on. In various respects there are matches between myth and social organization that are likewise quite varied. Anthropologists, however, are far from a consensus on the role and origin of the gods.

Polytheism in its wider relations. Polytheism can bear various relationships to other beliefs. It can be incompatible with some forms of theism, as in the Semitic religions; it can coexist with theism, as in Vaiṣṇavism; it can exist at a lower level of understanding, ultimately to be transcended, as in Mahāyāna Buddhism; it can exist as a tolerated adjunct to belief in transcendental liberation, as in Theravāda Buddhism.

BIBLIOGRAPHY. S.G.F. BRANDON (ed.), *A Dictionary of Comparative Religion* (1970), contains articles on the various gods and also on theories, and these have bibliographical sections. A useful compendium is the *Larousse Encyclopedia of Mythology* (1959). One example of the numerous recent anthropological literature is G.E. SWANSON, *The Birth of the Gods: The Origin of Primitive Beliefs* (1960).

(N.Sm.)

Pombal, Marquês de

Virtual ruler of Portugal from 1750 to 1777, remembered for both his ruthless exercise of his despotic powers and his reforms, Sebastião José de Carvalho e Mello (known for most of his life as Sebastião de Carvalho) was born in Lisbon on May 13, 1699.



Pombal, engraving by João S. Carpinetti, 1759.
By courtesy of the Biblioteca Nacional, Madrid

His father, Manuel de Carvalho e Ataíde, had served in the army as a cavalry captain and had been a nobleman of the royal house. He died relatively young, and Sebastião's mother remarried. Sebastião's uncle, Paulo de Carvalho, a professor at the Universidade de Coimbra, enrolled his nephew in that institution. But Sebastião abandoned his studies to enlist in the army, in which he reached the humble rank of corporal. Disillusioned with the army, he quit and dedicated himself to the study of history and law, and was later admitted, at the age of 34, to the Academia Real da História Portuguesa.

In 1733 he married Teresa Maria de Noronha e Almada, a widow, niece of the Conde de Arcos. They moved to the village of Soure, near Coimbra, where he had property. There he dedicated himself to his studies and to agriculture. In 1738 he returned to Lisbon. His uncle, who was professor at the Universidade de Coimbra, archpriest of the patriarchal see, and a person of political influence, recommended him to João da Mota, prime minister for King John V, who appointed him Portuguese ambassador to England. His wife, due to poor health, was unable to accompany him; she died in 1739.

Appointment as ambassador to England

His diplomatic career opened wider political horizons for him. He distinguished himself by the zeal with which he conducted several negotiations. And, for the seven years he stayed in London, Carvalho carefully studied the English in their political, social, and economic aspects.

After returning to Lisbon in 1745, Carvalho was immediately appointed ambassador plenipotentiary to Vienna, with the mission of serving as mediator in the solution of a serious quarrel between the Holy Roman empress Maria Teresa and the Vatican. The probabilities of success were very slight, but he overcame all obstacles, winning the sympathy of the Empress and the love of Eleonora von Daun, daughter of the Graf von Daun, whom he married in December 1745. The Austrian climate was bad for his health, however, and he submitted his resignation and returned to Lisbon at the end of 1749.

As King John V did not like him, Carvalho's progress was temporarily halted. But soon after the King's death on July 31, 1750, he was called by Queen Maria Ana, the widow, of whom he was a favourite, and was appointed to one of the royal councils. The heir to the throne, Prince Joseph, on being crowned king, made him a minister, along with two other favourites. He soon came to dominate Portuguese politics and the new monarch gave him a free hand. Thus began what may be termed the reign of the Marquês de Pombal.

Carvalho instituted domestic administrative reforms and succeeded in elevating Portugal's prestige in external politics. He granted England privileges that entitled it to receive large amounts of gold in exchange for manufactured articles. On the other hand, he stimulated national industry, forbidding the exportation of certain raw materials and developing the manufacture of silk, woolens, ceramics, and glass. Aiming at the development of commerce in the Orient, he founded a company for trade with India, similar to England's but which, however, was unsuccessful. But he was successful in another, similar enterprise—the Companhia do Grão-Pará—aimed at stimulating trade with Brazil.

His reforming activity was interrupted by a catastrophe, the earthquake of November 1, 1755. Two-thirds of Lisbon was reduced to rubble. Carvalho mobilized troops, obtained supplies, and had shelters and hospitals improvised. The day after the catastrophe, he was already outlining ideas for reconstruction. With architect Eugénio dos Santos' plans, old medieval Lisbon was changed into one of the most beautiful European cities.

Carvalho's firm and effective handling of the crisis enhanced his prestige and strengthened even further his position with the King. But his ascendancy had from the beginning created envy and animosity among two very powerful and influential groups: the high nobility and the Society of Jesus. On the night of September 3, 1758, an unsuccessful attempt was made on the King's life. This served as a pretext for Carvalho to rid himself of his enemies among the nobility and Jesuits, whom he accused of conspiracy. The court, influenced by him, attributed the crime to the Duque de Aveiro and to other members of the Távora family. On January 13, 1759, they were tortured to death. Carvalho then began persecuting members of the Society of Jesus. Almost all were deported to Rome, but some were imprisoned, along with many noblemen confined without proof of guilt.

Sebastião de Carvalho's power had become absolute. He was created conde de Oeiras in 1759 and continued to enact reforms, including the reform of university education, the initiation of commercial education, the creation of trading companies, and the reorganization of the army. In September 1769, the King conferred on him the title of Marquês de Pombal.

Fall from power

With the death of King Joseph on February 24, 1777, however, all the Marquês' power vanished. Under the new queen, Maria I, political prisoners were freed, and Pombal was accused of having abused his powers. He was found guilty by a judicial tribunal that subjected him to severe interrogation from October 1779 to January 1780. Queen Maria then banished him from Lisbon, and he retired to Pombal, where he died on May 8, 1782.

BIBLIOGRAPHY. J.C. MURPHY, *Travels in Portugal* (1795), recollections and comments on the policies of Pombal; VISCONT DE SANTAREM, *Quadro elementar das relacoes politicas e diplomaticas de Portugal*, 18 vol. (1842–60), especially valuable on the policies of Pombal relative to the Society of Jesus; J. LUCIO DE AZEVEDO, *O Marquez de Pombal e a sua época* (1909), biography of the Marquês and analysis of his economic policies; MARCUS CHEKE, *Dictator of Portugal: A Life of the Marquis of Pombal, 1699–1782* (1938), a well-documented biography; WILLIAM BECKFORD, *The Travel-Diaries of William Beckford of Fonthill*, ed. by G. CHAPMAN, 2 vol. (1928), narratives and comments about the Marquês and analyses of the Portuguese; CAMILO CASTELLO BRANCO, *Perfil do Marques de Pombal* (1900), a violent attack on the man and the minister, with intent to discredit.

(M.J.D.)

Pompadour, Madame de

Jeanne-Antoinette Poisson, marquise de Pompadour, held for nearly 20 years the important position of *maîtresse en titre* (officially recognized mistress) to the French king Louis XV. The future Madame de Pompadour was born in Paris on December 29, 1721.

Early years

Her parents were on the fringes of a class gaining in importance, speculators in the world of finance. Some of these people made immense fortunes, but many ended in the gutter if not in prison. Her father, François Poisson, involved in a black-market scandal, had to flee the coun-

try in 1725—his beautiful wife and two small children were then looked after by a more fortunate colleague, Le Normant de Tournehem. Both children were clever, and the girl was fascinating; she was educated to be the wife of a rich man. In those days rich men, even if they came from a low class, were interested in art and literature, and they expected their wives to share these interests. By the time Mademoiselle Poisson was of an age to marry, she could hold her own in any society and had made friends with many distinguished men, including Voltaire. Le Normant de Tournehem arranged a match for her with his own nephew, Charles-Guillaume Le Normant d'Étioles, a rising young man; they had a little girl, Alexandrine. Madame d'Étioles became a shining star of Parisian society and was admired by the King himself. In 1744 Louis XV's young mistress, the Duchesse de Châteauroux, died suddenly. She was soon replaced by Madame d'Étioles, who obtained a legal separation from her husband and was created marquise de Pompadour.

Nineteenth-century historians thought that Madame de Pompadour had complete ascendancy over Louis XV. These post-Revolution writers were concerned with portraying the Bourbon monarchs as poor creatures; it is now generally admitted that Louis XV was a much more able man than he has been painted. Shy and introspective, he had difficulty in communicating with people whom he did not know well. Madame de Pompadour acted as his private secretary; but although she gave the orders, the decisions were made by the King.

She began her reign at Versailles modestly. She was lodged in a few rooms under the roof; she set out to make herself agreeable to all those who counted for anything in the palace, beginning with Queen Marie (Maria Leszczyńska). Marie could hardly have been a more unsuitable wife for the handsome, artistic, sensual, and pleasure-loving Louis XV. Older than he, she was dowdy, stupid, and interested only in religion. After giving birth to an heir to the throne, she let the King understand that she had no wish to go on sleeping with him. Louis XV has a bad reputation as a husband, but few men ever had a better excuse for taking mistresses.

After five romantic years in her attic, Madame de Pompadour moved downstairs to a regal apartment. Louis XV now began to take other mistresses, but Madame de Pompadour was more firmly established than ever before; favours, promotions, and privileges could be obtained only through her good offices.

Her artistic and political collaboration with Louis

Her collaboration with the King was twofold, artistic and political. The artistic side was wholly successful. On her suggestion, her brother was appointed director of the King's buildings and created marquis de Marigny; the brother, the sister, and Louis XV, working in perfect harmony, planned and built the École Militaire and the Place Louis XV (now the Place de la Concorde) in Paris, most of the palace of Compiègne, the Petit Trianon Palace at Versailles, a new wing at the palace of Fontainebleau, and the exquisite Château de Bellevue, as well as many pavilions and summer houses. He and his mistress patronized all forms of decorative art: painters, sculptors, cabinetmakers, and craftsmen worked under the royal eye; the famous porcelain factory was built at Sèvres. Madame de Pompadour's 20 years of power marked the very apogee of taste in France. The protector of most of the authors and the editor of the *Encyclopédie*, she would have liked to do for literature what she did for the arts, but the King had no literary interests and disliked the intellectuals whom he knew.

The political collaboration between the King and his mistress was much less successful than the artistic, mainly because the French politicians and generals of the day were of such poor calibre. The Duc de Choiseul, by far the ablest of the ministers, was Madame de Pompadour's protégé. He was brought in to implement the famous Reversal of Alliances, which allied France with its old enemy Austria against the German Protestant principalities. This was a statesmanlike conception, but it was unpopular and led to the Seven Years' War, disastrous to France. Frederick the Great crushed the huge, incompetently led French and Austrian armies, while the English were driving the French out of Canada. All these defeats were laid at the door of Madame de Pompadour. She fell a prey to melancholy, and soon after the end of the war she died, on April 15, 1764, probably of cancer of the lung, in her apartment at Versailles. One of her last actions was to get Louis XV's support for the revision of the Calas case, a gross miscarriage of justice in which Voltaire was interested. Voltaire said of her:

> I mourn her out of gratitude. . . . Born sincere, she loved the King for himself; she had righteousness in her soul and justice in her heart; all this is not to be met with every day.

BIBLIOGRAPHY. The best life is E. and J. DE GONCOURT, *Madame de Pompadour*, new ed. (1888), in French; also very good are P. DE NOLHAC, *Portraits du XVIII^e siècle* (1933), *Madame de Pompadour et la politique, d'après des documents nouveaux* (1928), and *Louis XV et Madame de Pompadour* (1902). In English, see NANCY MITFORD, *Madame de Pompadour* (1964).

(Na.M.)

Pompeii and Herculaneum

The ancient cities of Pompeii and Herculaneum, situated in the fertile Campanian plain of Italy southeast of Naples, 14 miles (23 kilometres) and five miles (eight kilometres) respectively, owe their fame to their sudden destruction by Vesuvius in AD 79. Their origins are obscured by legend, and few events in their history are preserved in the writings of the ancient authors. But the archaeological evidence derived from the excavation of these cities and numerous villas in the adjacent area is an invaluable source of information about a great many aspects of life in the Greco-Roman world.

History. It seems certain that Pompeii, Herculaneum, and Stabiae were first settled by the Oscans, who were descendants of the Neolithic inhabitants of Campania. Archaeological evidence indicates that the Oscan village of Pompeii, strategically located near the mouth of the Sarnus River, soon came under the influence of the cultured Greeks who had settled across the bay in the 8th century BC. Greek influence was challenged, however, when the Etruscans came into Campania in the 7th century. The Etruscans' influence remained strong until their sea power was destroyed by King Hieron I of Syracuse in a naval battle off Cumae, in 474 BC. A second period of Greek hegemony followed.

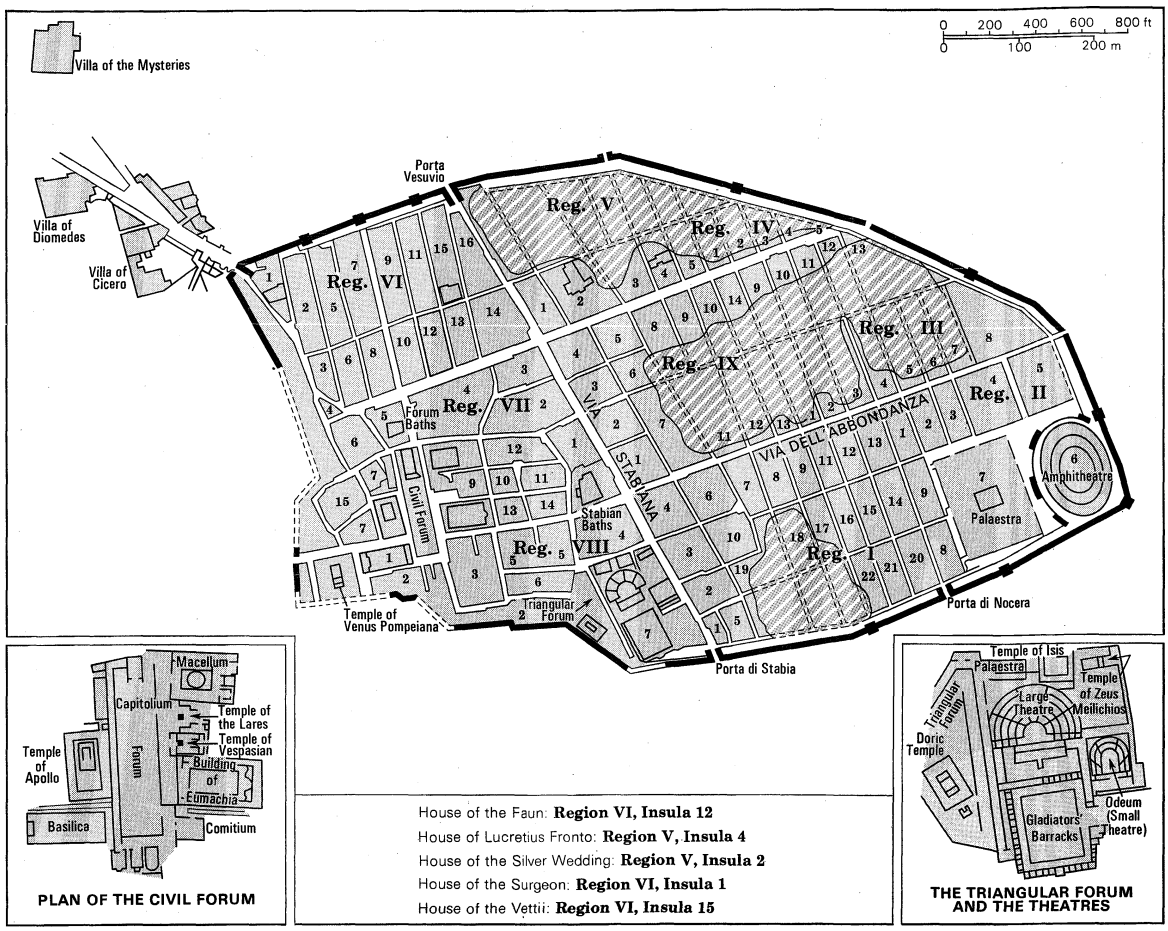
Then, toward the end of the 5th century, the warlike Samnites, an Italic tribe, conquered Campania, and Pompeii, Herculaneum, and Stabiae became Samnite towns. Coinage suggests that a century later the city of Nuceria was the head of a league to which Pompeii, Herculaneum, and Stabiae may have belonged.

Conquest by Rome

Pompeii is first mentioned in history in 310 BC, when, during the Second Samnite War, a Roman fleet landed at the Sarnus port of Pompeii and from there made an unsuccessful attack on Nuceria. At the end of the Samnite Wars, Campania became a part of the Roman confederation, and the cities became "allies" of Rome. But they were not completely subjugated and romanized until the time of the Social War (the war of the *socii*, or "allies"). Pompeii joined the Italians in their revolt against Rome in this war and was besieged by the Roman general Lucius Cornelius Sulla in 89 BC. After the war, Pompeii, along with the rest of Italy south of the Po, received Roman citizenship. As a punishment for Pompeii's part in the war, a colony of Roman veterans was established there under Publius Sulla, the nephew of the Roman general. After initial difficulties, Pompeians and colonists seem to have adjusted with a minumum of friction. Whether Herculaneum had such a colony is not known. Latin now replaced Oscan as the official language of both towns, and they soon became romanized in institutions, architecture, and culture. Stabiae was destroyed and its territory given to Nuceria.

A riot in the amphitheatre at Pompeii between the Pompeians and the Nucerians, in AD 59, is reported by the Roman historian Tacitus. An earthquake in AD 62 did great damage in both Pompeii and Herculaneum. The cities had not yet recovered from this catastrophe when final destruction overcame them 17 years later.

Vesuvius erupted on August 24, AD 79. A vivid eyewit-



Plan of Pompeii.
Adapted from A. Maiuri, *Pompeii*; ed. by Istituto Geografico de Agostini, Novara, Italy

Eruption of Vesuvius

ness report is preserved in two letters written by Pliny the Younger to Tacitus, who had inquired about the death of Pliny the Elder, commander of the Roman fleet at Misenum. Pliny the Elder had died at Stabiae while trying to rescue friends. When the eruption ceased on the second day, Pompeii was covered with lapilli (volcanic debris) and ashes to a depth of 19 to 23 feet (six to seven metres). The villas at Stabiae, at a greater distance from Vesuvius, did not have as deep a covering. At Herculaneum the avalanche of volcanic materials carried along by torrents of water hardened into a kind of mud-lava that penetrated everywhere and covered the city up to a depth of 65 feet (20 metres). Excavation there is thus much more difficult, but the tufa-like covering preserved the city and made it inaccessible to plunderers.

History of excavations. The ruins at Pompeii were first discovered late in the 16th century by the architect Domenico Fontana, who tunnelled under the hill known as La Cività to build a water channel to bring water from the Sarno River to the town of Torre Annunziata.

Excavation of the buried cities began first at Herculaneum, in 1709, during Austrian occupation. When the workmen of an Austrian officer, the Comte d'Elbeuf, sank a shaft that reached the theatre among the finds were three marble statues of young women, which eventually found their way to Augustus III of Poland, who took them to Dresden. There they fired the imagination of his daughter, Maria Amalia, who became the wife of King Charles IV of Naples (later Charles III of Spain). In 1738 Charles began excavations at Herculaneum. Work did not begin at Pompeii until 1748, and in 1763 an inscription (*rei publicae Pompeianorum*) was found that identified the site as Pompeii.

Excavations have continued intermittently until the present. Earlier digging was haphazard and often irresponsible; excavators were primarily treasure seekers, hunting for imposing buildings or museum objects. Haphazard digging was brought to a stop in 1860, when the Italian archaeologist Giuseppe Fiorelli became director of the excavations. Areas lying between excavated sites were cleared and carefully documented. Pompeii was divided into nine Regions; the insulae (blocks) in each Region were numbered, and each door on the street was given a number so that each house could be conveniently located by three numbers (see plan). Fiorelli also developed the technique of making casts of bodies by pouring cement into the hollows formed in the volcanic ash when the bodies disintegrated. This technique is also used to make casts of ancient doors, furniture, and roots.

Systematic excavations

In 1951, after the interruption caused by World War II, intensive excavation was resumed under Amedeo Maiuri, who was in charge of the excavations from 1924 to 1961. Large areas to the south of the Via dell'Abbondanza in Regions I and II were uncovered, and the debris piled outside the city walls was cleared away. This revealed the Porta (Gate) di Nocera and an impressive stretch of cemetery lining each side of the road leading from the gate to Nuceria. By the early 1970s, about three-fourths of the city had been excavated.

At Herculaneum excavation was carried on by means of shafts and underground tunnelling until 1765. In this way the theatre, the so-called Basilica, and the Villa of the Papyri on the outskirts of town (named from a library of Greek papyri found there) were explored and a tentative plan of the city was made. Many magnificent paintings and statues were uncovered. In 1828 excavation in the open was begun, but because of the hardness of the fill and because the modern town of Resina is built over part of the ancient town, the work has been carried on intermittently, and only four insulae have been completely excavated, along with parts of four others. The regular plan of the city, laid out on a grid system, indicates Greek influence. The public buildings that have been uncovered include two baths, an unusually large palaestra (sports



The excavations at Pompeii, looking north from Region I across the Via dell'Abbondanza to Region IX. The farm above the excavations is at ground level.
Edwin Smith

ground), and the recently discovered Shrine of the Augustales, which was dedicated to the emperor cult and is located on the partially excavated Forum. Among the most luxurious homes are the House of the Stags and the House of the Mosaic Atrium, both with terraces overlooking the bay. At the opposite extreme is the Casa a Graticcio, named from its cheap construction. Its small apartments opened on a common courtyard.

In the vicinity of Stabia and Gragnano, excavations initiated by Charles IV discovered 12 villas between 1749 and 1782. Work was not resumed again until the present century, and it is still in progress. The Villa of San Marco, with its two large peristyle gardens and bath, is the best preserved. Other villas have been found near Gragnano, in the vicinity of Pompeii, and at nearby Scafati, Domicella, Torre Annunziata, and on the lower slopes of Vesuvius near Boscoreale and Boscotrecase. Many of these villas were reburied after excavation, but a few can be seen, notably the Villa of the Mysteries.

Description of the remains. The city of Pompeii was shaped irregularly because of the prehistoric lava flow on which it was built. Excavations indicate that the southwestern part of the town was the oldest, but scholars are not agreed as to the stages by which the walls were expanded, or by whom. The walls are two miles (three kilometres) in circumference, and they enclose an area of about 155 acres. They were probably given their final boundaries after the Samnites took over Pompeii, but there is evidence of Greek influence. Seven city gates have been excavated. The chief street running in a southeast–northwest direction was the Via Stabiana; it connected the Porta Vesuvio (144 feet [44 metres] above sea level), in the highest part of the city, with the Porta di Stabia (26 feet [eight metres]), in the lowest part of town. Through this gate came traffic from the Sarnus River and Stabiae. This street was crossed by two other main streets, the Via dell' Abbondanza and the Via di Nola.

The public buildings at Pompeii are for the most part grouped in three areas: the Forum (elevation 110 feet [34 metres]), located in the large level area on the southwest; the Triangular Forum (82 feet [25 metres]), standing on a height at the edge of the south wall overlooking the bay; and the Amphitheatre and Palaestra, in the east.

The Forum was the centre of the religious, economic, and municipal life; it was a large rectangular area surrounded by a two-storied colonnaded portico. Dominating the Forum on the north was the temple dedicated to the Capitoline triad of deities: Jupiter, Juno, and Minerva. To the east was the Macellum, or large provision market; to the south the small sanctuary of the city Lares (guardian deities), built after the earthquake in AD 62, the Temple of Vespasian, and the imposing headquarters of the woollen industry, erected by the wealthy patroness Eumachia. Opposite the Capitolium on the south end of the Forum were the meeting place of the city council and the offices of the magistrates of the city. The large basilica, with its main room surrounded on four sides by a corridor, is the most architecturally significant building in the city and of considerable importance in studying the origin and development of the Christian basilica. It served as a covered exchange and as a place for the administration of justice. To the west was the Temple of Venus Pompeiana, patron deity of Pompeii. Across from the basilica was the Temple of Apollo, one of the oldest in the city. Subsoil excavations here give evidence for Greek and Etruscan influence.

The Triangular Forum is the site of the Doric Temple, the oldest temple in the city. Between the third and the first centuries, a theatre, a palaestra, and a small covered theatre were built to the east of the Triangular Forum. The temples of Zeus Meilichius and of Isis, and the old Samnite palaestra were nearby. In the east corner of Pompeii was the Amphitheatre; the oldest so far known, it was built after the Roman colony was established at Pompeii. To the west a large palaestra was built to replace the old Samnite palaestra. Baths were scattered throughout the town: the Stabian Baths (which date before the Roman period), the Forum Baths, the Central Baths—still under construction at the time of the eruption—and many baths in luxurious private homes.

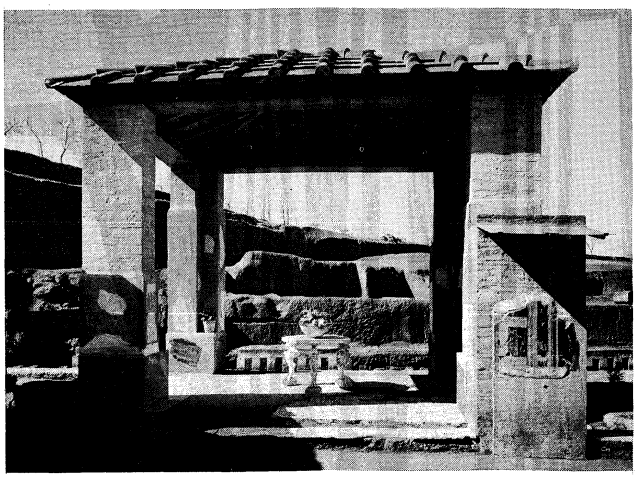
But more significant than the public buildings, which have been excavated at other sites, are the hundreds of private homes. These are unique, for only at Pompeii is it possible to trace the history of domestic architecture for at least four centuries. The earliest houses date from the First Samnite period (4th–3rd centuries BC). The House of the Surgeon is the best known example of the early atrium house built during this period.

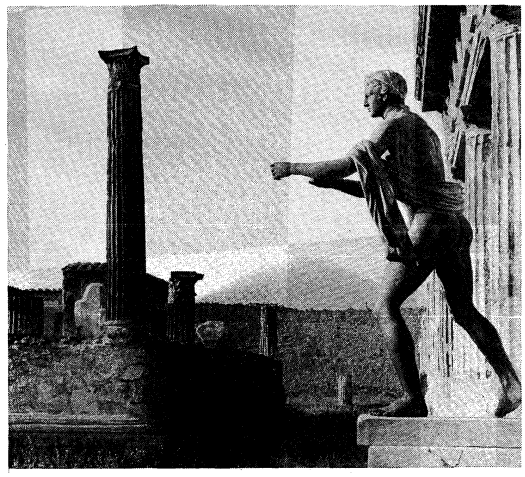
Importance of private homes

The most luxurious houses were built during the Second Samnite Period (200–80 BC), when increased trade and cultural contacts resulted in the introduction of Hellenistic refinements. The House of the Faun occupies an entire city block, has two *atria* (chief rooms), four *triclinia* (dining rooms), and two large peristyle gardens. Its facade is built of fine-grained gray tufa from Nuceria, the chief building material of this period. The walls are decorated in the Incrustation, or First Pompeian style of painting, which imitates marble-veneered walls by means of painted stucco. The famous Alexander the Great mosaic found in the House of the Faun is probably a copy of a lost Hellenistic painting. Many of the houses from this period were decorated with elaborate floor mosaics. The House of the Silver Wedding, with its imposing high-columned atrium, was also built during this period, but it underwent later alterations. The handsome banquet hall and the exedra, which served as a schoolroom for children of the family, were decorated in the Second Pompeian, or Architectural, style, which was popular from 80 BC to AD 14.

The large number of houses built during the Samnite period made it necessary to build fewer houses in the Roman period. Those that were built were usually less imposing, with lower *atria,* but with more elaborate decoration. The House of Lucretius Fronto is a small but elegant house of the Roman Imperial period. The tablinum (master's office) is decorated in especially fine Third Pompeian, or Egyptianizing, style, usually dated from the early empire to the earthquake. The House of the Vettii is typical of the homes of the prosperous merchant class of the Roman period. Some of its rooms are decorated in the Fourth Pompeian, or Ornamental, style.

The atrium-peristyle house, with its handsome paintings, elegant furniture, beautiful gardens with fountains and bronze and marble sculptures, is not as typical as has generally been supposed. There are also numerous small homes throughout the city, many of them shop-houses. Excavators now preserve as completely as possible all aspects of ancient life. The homes of the humble have as much to tell us as those of the wealthy. Roofs, second stories, and balconies can now be restored.





Excavated buildings.
(Left) Pergola in the garden of the House of the Stags, Herculaneum. (Right) Temple of Apollo, Pompeii, with Mt. Vesuvius in the background.
By courtesy of (left) the Italian State Tourist Office, New York; photograph (right), Edwin Smith

Influence on European culture. The discoveries at these sites had a profound influence on European taste. News of the excavations kindled a wave of enthusiasm for antiquity that spread throughout Europe. The laudatory pronouncements of the eminent German classicist Johann Joachim Winckelmann, who made his first trip to Naples in 1755, and the etchings of Giambattista Piranesi did much to popularize the excavations. Naples, Pompeii, and Herculaneum became important stops on the European Grand Tour made by English visitors, who found an enthusiastic guide and a well-informed amateur archaeologist in Sir William Hamilton, ambassador to the Naples court from 1764 to 1800, and a captivating hostess in his second wife, Lady Emma Hamilton.

Artists, architects, potters, and even furniture makers drew much inspiration from Pompeii and Herculaneum. Contemporary painted interiors were inspired by the frescoed walls found in the excavations. The stucco work popularized in England by the 18th-century architects Robert and James Adam used the same motifs. The famed potter Josiah Wedgwood changed the style of his work and named his factory "Etruria," sharing the mistaken belief that Pompeii and Herculaneum were Etruscan. In France, the Louis XVI style incorporated Pompeian motifs, and the apartment of Louis's queen, Marie-Antoinette, at Fontainebleau was decorated in this style, which became popular throughout Europe. Jacques-Louis David and his student Jean-Auguste-Dominique Ingres drew inspiration for their paintings from the excavations. Indeed, the Neoclassic style stimulated by the discoveries at Pompeii and Herculaneum completely replaced the Rococo and became the artistic style of the French Revolution and of the Napoleonic period.

Importance as historical source. The extent of these archaeological sites makes them of the greatest importance, for they provide a unique source of information about so many aspects of social, economic, religious, and political life of the ancient world. The many well-preserved house shrines give a hitherto unexpected picture of the vitality of religion in the family. The bakeries, complete with mills, kneading machines, and ovens, some still containing loaves of bread, show how this staple of everyday life was produced. Numerous fulleries (wool processing and cleaning plants) make it possible to study this important industry. The shops of the sculptor, toolmaker, gem cutter; the *garum* (fish sauce) and the lamp factories; and the many wine and food shops document other aspects of ancient life. Pompeii was a busy port town that exported wine, *garum*, woollen products, and perfume throughout the Mediterranean. Merchants and tradesmen found food and lodging in the taverns, restaurants, and inns located near the city gates and the Forum. Some were quite attractive and served food to guests who reclined in the garden; in the cheaper places the rooms were small and dark, and customers sat on stools.

Everyday life

Inscriptions provide further information. They include monumental inscriptions on public buildings, tombs, and statue bases; the business transactions recorded in the famous wax tablets of the banker Lucius Caecilius Jucundus; announcements of gladiatorial combats; and many election notices, echoes of hotly fought contests. Preserved in the graffiti are accounts, lists of market days, the exchanges of lovers, quotations from Virgil, and even the scratched alphabets of children. Epigraphical and archaeological evidence makes it possible to study the stratification of society and learn more of the freedmen, slaves, small businessmen, and aristocrats of the ancient Roman world.

Further, Pompeii offers the best opportunity for the study of city planning and land use in an ancient city. Recent excavations have revealed an unexpected amount of open land. The large insula across from the Amphitheatre was not the "Foro Boario" (Cattle Market), as has been long supposed, but a vineyard. Many vineyards, fruit trees, and gardens have been found, indicating less intensive land use and a smaller population than has been thought.

Unfortunately, the excavations are constantly endangered by the ravages of weather and destructive vegetation. Of first importance is the preservation, restoration, and study of the valuable evidence already uncovered before it is lost forever.

BIBLIOGRAPHY. AUGUST MAU, *Pompeji in Leben und Kunst* (1900; rev. Eng. trans. by F.W. KELSEY, *Pompeii: Its Life and Art*, 2nd ed., 1902), still the most comprehensive treatment, but now outdated; MATTEO DELLA CORTE, *Case ed abitanti di Pompei*, 3rd ed. (1965), a thorough, scholarly treatment of the houses and places of business, and their owners as identified from the inscriptions and the graffiti; ROBERT ETIENNE, *La Vie quotidienne à Pompéi* (1966), a good discussion of daily life; H.H. TANZER, *The Common People of Pompeii: A Study of the Graffiti* (1939), brief but interesting; WILHEMINA JASHEMSKI, *Letters from Pompeii* (1963), actual letters written from Pompeii, with descriptions of new excavations, coloured photos of excavated sites, and paintings by M. SASEK showing how the sites looked in antiquity; *The Gardens of Pompeii, Herculaneum, and Stabiae* (in prep.), a comprehensive discussion of these sites, for almost every house and many public buildings had gardens; AMEDEO MAIURI, *Pompeii*, 14th ed. (1970), the best brief treatment available; *Ercolano: I nuovi scavi, 1927–1958*, 2 vol. (1959), the definitive work on this site; *Ercolano*, 5th ed. (1959; Eng. trans., *Herculaneum*, 5th ed., 1959), a brief but excellent guidebook; LIBERO D'ORSI, *Gli scavi archeologici di Stabia* (1968), a brief guide to the excavations and to the Antiquarium.

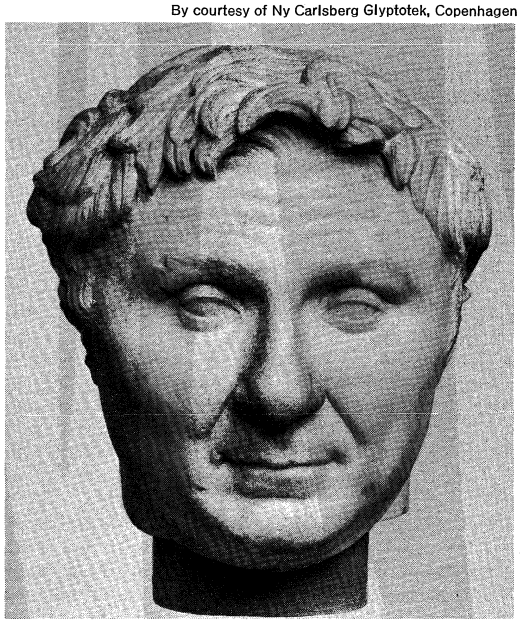
(W.F.J.)

Pompey the Great

In the desperate struggle for power in the late Roman Republic, Gnaeus Pompeius Magnus (anglicized Pompey) stands as one of the greatest statesmen and generals

of the age. His campaigns in Africa, Spain, Asia Minor, and the Mediterranean were conducted with a thoroughness and efficiency that marked his entire career. His reorganization of Rome's eastern frontier regions lasted with few changes for five centuries. He lacked the genius of his great rival Julius Caesar, but to admirers of the republic he represents the bulwark of the conservative nobility against Caesar's towering ambitions.

By courtesy of Ny Carlsberg Glyptotek, Copenhagen



Pompey, bust *c.* 60–50 BC. In the Ny Carlsberg Glyptotek, Copenhagen.

Early career. Pompey was born in Rome on September 29, 106 BC. He belonged to the senatorial nobility, although his family first achieved the office of consulship only in 141. Fluent in Greek and a lifelong and intimate friend of Greek literati, he must have had the normal education of a young Roman nobleman; but his early experience on the staff of his father, Pompeius Strabo (consul in 89), did much to form his character, develop his military capacities, and arouse his political ambition. The family possessed lands in Picenum, in eastern Italy, and a numerous body of clients, which Strabo greatly enlarged in the year of his consulship. In a civil war (88–87) between the rival generals Lucius Sulla and Gaius Marius, Strabo defied Sulla and favoured the Marians and a fellow general, Cinna (see ROME, HISTORY OF).

Support of Sulla

But, after his father's death, Pompey detached himself from the Marians. A report that he was "missing" in Cinna's army, when it was embarking for the Balkans to deal with Sulla, led to the lynching of Cinna by his troops (84). Pompey's part in this mutiny is unclear; he next appears with three legions recruited in Picenum, joining Sulla as an independent ally in the campaign to recover Rome and Italy from the Marians (83). Sulla made ample use of his youthful ally's military abilities. Pompey married Sulla's stepdaughter, Aemilia, pregnant from her former husband, only to see her die soon after in childbed. On Sulla's orders the Senate gave Pompey the job of recovering Sicily and Africa from the Marians—a task he completed in two lightning campaigns (82–81). Pompey ruthlessly executed Marian leaders who had surrendered to him. To his enemies he was Sulla's teenage butcher; to the troops he was "Imperator" and "Magnus." From Africa Pompey demanded that a triumph be given him in Rome; he refused to disband his army and appeared at the gates of Rome, obliging Sulla to yield to his demand. After Sulla's abdication, Pompey supported the renegade Sullan Marcus Lepidus for the consulship of 78. Once in office Lepidus attempted revolution, and Pompey promptly joined the forces of law and order against him. The rising crushed, however, Pompey refused to disband his army, which he used to bring pressure on the Senate to send him with proconsular power to join Metellus Pius in Spain against the Marian leader Sertorius.

Campaigns in Spain

The reconquest of Spain taxed Pompey's military skill and strained his own and the state's resources to the utmost. In the end it was he, not Metellus, who imposed on Spain a settlement reflecting and promoting his own political aims. His policy was one of reconciliation and rehabilitation. His personal authority and patronage now covered Spain, southern Gaul, and northern Italy. Unlike Metellus, Pompey took his army back to Italy with him, ostensibly to assist in putting down a slave revolt led by Spartacus but, in reality, to secure a triumph and election to the consulship for 70. The nobles whom Sulla had restored to power had proved to be more corrupt and incompetent than ever. Pompey promised reforms at home and abroad, especially the restoration of the power of the tribunes, elected officials who protected popular rights by initiating legislation. A bargain was struck with his rival Marcus Licinius Crassus, the two were jointly elected consuls, and Pompey was given another triumph. The uneasy partners cooperated in sweeping away the more important of Sulla's reactionary laws.

Reorganization of the East. Although the nobles were to continue to dominate the consular elections in most years, the real sources of power henceforth lay outside of Italy. Extraordinary commands would have to be created if Rome was to recover control of the sea from the pirates and bring peace to the eastern provinces, threatened by Mithradates VI of Pontus, a kingdom in eastern Anatolia. It was Pompey who benefited most from the restoration of tribunician initiative. After his consulship, he waited in Rome while rival nobles undermined the position of Lucius Licinius Lucullus, who was campaigning against Mithradates in Anatolia, and made halfhearted attempts to deal with the pirates. Finally, in 67, the tribune Aulus Gabinius forced a bill through the popular assembly empowering Pompey to settle the pirate problem, and Pompey cleared the sea in three months (67).

Pompey was still in the East, resettling pirates as peaceful farmers, when in Rome another tribune, Gaius Manilius, carried through, against weakened opposition, a bill appointing Pompey to the command against Mithradates, with full powers to make war and peace and to organize the whole Roman East (66). Pompey displaced Lucullus and lost no time defeating Mithradates in Asia Minor. With the death of Mithradates in 63, Pompey's hands were free to plan the consolidation of the eastern provinces and frontier kingdoms. For 6,000 talents he set up King Tigranes in Armenia as a friend and ally of Rome —and as his own protégé. Pompey rejected the Parthian king's request to recognize the Euphrates as the limit of Roman control and extended the Roman chain of protectorates to include Colchis, on the Black Sea, and the states south of the Caucasus, Iberia and Albania. In Anatolia, he created the new provinces of Bithynia-Pontus and Cilicia. He annexed Syria and left Judaea as a dependent, diminished temple state. The organization of the East remains Pompey's greatest achievement. His sound appreciation of the geographical and political factors involved enabled him to impose an overall settlement that was to form the basis of the defensive frontier system and was to last, with few important changes, for over 500 years.

Defeat of Mithradates

Pompey's power and prestige were at their height in December 62, when he landed at Brundisium (Brindisi) and dismissed the army, whose continued loyalty he had ensured by lavish gifts from his spoils of war. His third triumph (61) trumpeted the grandeur of his achievement. The following decade was the period of his ascendancy in Italy, an ascendancy that was to be eroded through Caesar's growing military power and gradual capture of Pompey's worldwide *clientelae*, from the power base he, in turn, created in north Italy and Gaul. Pompey's inveterate enemies in Rome were the Optimates, the inner ring of nobles, not Crassus or Caesar, who had merely tried to steal the limelight in Pompey's absence and to manoeuvre into a better position for bargaining with their former political ally. The nobles mean-

Pompey's political enemies

while had gradually reasserted their dominance in Rome and hampered attempts to alleviate the condition of Italy and the Roman populace. Once back in Italy, Pompey avoided siding with popular elements against the Optimates. He was no revolutionary. He wanted all classes to recognize him as first citizen, available for further large-scale services to the state. He had divorced his third wife, Mucia, and now proposed to ally himself by marriage to the party of the young senatorial leader Marcus Porcius Cato the Younger. But the nobles were closing their ranks against him, and his offer was rebuffed. Lucullus and others were determined to prevent the *en bloc* ratification of Pompey's eastern settlement and to reject his demand for land for his veterans. Pompey turned to Cicero and sought his help, in vain.

The First Triumvirate. Help only came when Casear returned from his governorship in Spain, prepared to give up a triumph if he could secure election to the consulship for 59. Pompey, Crassus, and Caesar formed the unofficial and at first secret First Triumvirate (mid-61). It was to become more than a mere election compact. It would strain all the resources of the triumvirs to wrest one consulship from the Optimates; their continued solidarity was essential if they were to secure what Caesar gained for them in 59. Caesar, for his part, wanted a long-term command. He was Pompey's pupil in this. Pompey, who now married Caesar's daughter, Julia, saw Caesar as his necessary instrument. Caesar, once consul (January 59), immediately forced through a land bill and, shortly after, another appropriating public lands in Campania. Once he had secured a five-year command in Illyria (on the Adriatic coast) and Gaul he could be relied on to take off a large proportion of Pompey's discharged troops and give them further opportunities for profitable employment. The year saw a continuous battle with the enemies of the triumvirs. Caesar's obstructive consular colleague, Marcus Calpurnius Bibulus, did his best to undermine the validity of Caesar's legislation, which included the ratification of Pompey's eastern settlement. Pompey was forced to give open approval to Caesar's violent, unconstitutional methods. To bolster the triumviral cause in his absence, Caesar persuaded Pompey to support the election to the tribunate of Publius Clodius, a personal enemy of Cicero. Cato was silenced by being sent on an honourable and lucrative mission to annex Cyprus. In his tribunate (58) Clodius first carried out the policy of triumvirs. The consuls were induced to block attempts to rescind Caesar's measures by the bribe of lucrative provincial commands. Clodius, aided by gangs of street thugs, passed a law forcing Cicero into exile, and Pompey, ashamed of his failure to help Cicero, could only confess that his hands were tied. But Clodius saw his tribunate as an opportunity that would never recur. Using the violence of his street gangs and bribing voters with free corn, Clodius used his legislative initiative to acquire power and wealth, but only by cutting into Pompey's monopoly of patronage in the East. Pompey at once began to work for Cicero's return, to have him as a counterweight against Clodius. At first Clodius' gangs hustled Pompey off the streets and besieged him in his own house. Extreme Optimates joined Clodius in baiting Pompey, but moderate senators rallied to him. Caesar consented to the recall of Cicero, who returned late in 57 committed to support Pompey. The Optimates and Clodius continued their opposition but had to consent to the passage of a law through the Senate assigning Pompey the task of organising Rome's grain supply (57). Pompey's enemies were brought to heel by the threat of a more extreme tribunician proposal to give Pompey the supreme military powers they suspected him of desiring.

Opposition to Clodius

Pompey solved the problem of Rome's grain supply with his usual efficiency, but the nobles kept up their opposition. The year 56 was a critical one for the triumvirs. The nobles concocted religious impediments to prevent the dispatch of Pompey on a military mission to Egypt, while Clodius contrived to persuade Pompey that Crassus had designs on his life. An attempt was made to suspend Caesar's law for the distribution of Campanian land.

Alarmed at Pompey's suspicions and truculence, Crassus set off to meet Caesar at Ravenna, and Caesar in turn came to the limit of his province at Luca to meet Pompey. The Luca conference (56) prepared the ground for the next phase of triumviral cooperation: Pompey and Crassus were to secure election to the consulship for 55, for they, too, wanted five-year commands in the provinces, while Caesar's command was to be renewed for another five years. The three secured their ends by violence and corruption after a prolonged struggle in Rome. Early in 55 Pompey and Crassus were at last elected consuls, with most of the lesser magistracies going to their supporters. Caesar obtained the extension of his command, while Pompey and Crassus received commands in Spain and Syria, respectively. Pompey could stay on in Italy and govern his provinces by deputies. His ascendancy was now more pronounced than ever. It was then that he dedicated a great new theatre in Rome. Laws were passed in Caesar's interest, allowing him the expectation of a second consulship after the regular ten-year interval. Clodius was made to come to heel, the power of the political clubs in Rome was broken, and the functioning of the courts improved.

The conference at Luca

But the triumvirate was coming to an end. The death of Julia (54) destroyed the strongest bond between Pompey and Caesar, and Crassus suffered disastrous defeat and death in Mesopotamia. The triumvirate existed no longer; but Pompey as yet showed no inclination to break with Caesar, whose greatest victories in Gaul were yet to come.

Civil war. Meanwhile, from outside the walls of Rome, Pompey watched the anarchy in the city becoming daily more intolerable. He was prepared to wait without committing himself until the Optimates found an alliance with him unavoidable. He refused further offers from Caesar of a marriage alliance. There was talk in Rome as early as 54 of a dictatorship for Pompey. Street violence made it impossible to hold the elections. In January 52 Clodius was killed by armed followers of Titus Annius Milo, whose candidacy for the consulship was being bitterly opposed by both Pompey and Clodius. Now both factions exploded into even greater violence. The senate house was burnt down by the mob. With no senior magistrates in office the Senate had to call on Pompey to restore order. It was the hour he had waited for. He speedily summoned troops from Italy. The nobles would not have him as dictator; they thought it safer to appoint him sole consul. Though Pompey now chose the Optimate alliance, he continued to profess loyalty to the old bond with Caesar. He could now impose his will on the Optimates, defeating their efforts to engineer Milo's acquittal.

Pompey's legislation of 52 reveals his genuine interest in reform and the duplicity of his conduct towards Caesar. He reformed procedure in the courts and produced a panel of respectable jurors. A severe law against bribery at elections was made retrospective to 70 and, for all Pompey's protests, was rightly taken by Caesar's friends as aimed at him. Another useful law enforced a five-year interval between tenure of magistracies in Rome and assumption of provincial commands. But this law and another, which, in one of its clauses, prohibited candidature in absence, effectively destroyed the ground of Caesar's expectation (based on earlier legislation and an understanding with Pompey) that he should become designated consul, and so safe from prosecution, before he had to disband his army in Gaul. Several attempts were made in the years 51 and 50 to recall Caesar before the expiration of his second term in Gaul. They were frustrated by the growing assertiveness of Caesar's faction and agents in Rome. Pompey, for all his growing fear and suspicion of Caesar's power and ambitions, did not come out openly against Caesar until late in 51, when he suddenly made clear his intentions. He declared that he would not consider the suggestion that Caesar should become designated consul while still in command of his army. His proposals for a compromise date for Caesar's recall were unacceptable to Caesar whose sole, but effective, resource now was to use the wealth he had accumulated in Gaul to buy men who could obstruct his enemies

Break with Caesar

in the Senate, in particular Gaius Scribonius Curio, tribune in 50, who eventually induced the Senate to vote for a simultaneous surrender of their provinces and armies by both Pompey and Caesar. When war came, the Senate and even the nobles were evenly divided as between Caesar and Pompey. The consulars—that is, those who had held the rank of consul—were solidly for Pompey, although they saw him simply as the lesser evil. Late in 50 the consul Gaius Marcellus, failing to induce the Senate to declare Caesar a public enemy, visited Pompey with the consuls designate and placed a sword in his hands. Pompey accepted their invitation to raise an army and defend the state. Caesar continued to offer compromise solutions while preparing to strike. On January 7, 49, the Senate finally decreed a state of war. Four days later Caesar crossed the Rubicon.

Pompey's strategic plan, unpopular but sound, was to abandon Rome and Italy to Caesar and rely on his command of the sea and the resources of the East to starve out the Caesarians in Italy and overcome his rival; but he did not have the disciplined loyalty and full cooperation of his Optimate allies, and Caesar's swift advance southward only just failed to prevent his withdrawal from Italy. Across the Adriatic at Dyrrhachium the wisdom of Pompey's strategy became clear, even to Cicero. Caesar, after a hazardous crossing in pursuit, found himself cut off from his base in Italy by sea and facing superior land forces. But Pompey eventually had to abandon his naval blockade of the rest of Caesar's forces in Brundisium and failed to prevent their crossing to join Caesar. Caesar's army was repulsed in an assault on Pompey's camp at Dyrrhachium and, failing a quick decision in the West, he was obliged to move eastwards into Thessaly. Pompey followed and joined forces with the Senate's army there under Scipio, rendering Caesar's position intolerable. But at this juncture Pompey, under pressure from his Optimate allies, decided for battle, a sensible enough decision, if his opponent had not been a commander of genius.

Defeat at Pharsalus

Pompey suffered a disastrous defeat on the plain of Pharsalus (48); it was his first major reverse in battle, and it shattered his morale. He fled from his camp as the enemy stormed it and made his way to the coast. His supporters were to rally and involve Caesar in strenuous fighting in Africa, Spain, and the East for three more years; but Pompey did not live to play a part in this struggle. Hurried on by Caesar's rapid pursuit, hesitating to seek Parthian help, he lost contact with his own fleet. He moved on southward to Cilicia and Cyprus and thence, despairing of raising fresh forces, towards Egypt. He decided to land at Pelusium and seek the assistance of Ptolemy, his former client. The King marched down to the coast, ostensibly to welcome him; but he and his counsellors had chosen not to risk offending the victorious Caesar. Pompey's small squadron lay offshore while Pompey, bidding farewell to his wife Cornelia, complied with an insidious invitation to enter with several companions a small boat sent to bring him to land. As he prepared to step ashore he was treacherously struck down and killed (September 28, 48 BC). The Egyptians cut off his head and left the corpse to be cremated and given funeral rites by a faithful retainer. A simple mound by the sea marked the spot.

Assessment. Pompey's great name cast a lasting shadow. His end inspired some of Lucan's finest verses. In the empire he acquired official respectability, and the greatness of his achievement was fully appreciated by such writers as the elder Pliny and Plutarch. Tacitus, writing in the 1st century AD, echoed harsher judgments of Pompey's near contemporary, Sallust. But there are few clear-headed or unbiased accounts of Pompey by his own contemporaries. Caesar would have his readers believe that he wrote of Pompey more in sorrow than in anger; his propaganda was discreet and subtly damaging to his rival's reputation. Cicero's veering, day-to-day judgments of Pompey reveal his inability to see clearly through the distorting medium of his own vanity. The inflated eulogies of Pompey in Cicero's speeches are punctured by his persistent sniping at him in his letters. Yet he looked up to him for leadership and, in the moment of decision, joined him. Modern scholars have been slow to reach a balanced judgment. To the 19th-century German historians Theodore Mommsen and Wilhelm Drumann, Pompey was merely a blundering foil to Caesar, magnified by circumstances. Eduard Meyer, providing in the early 20th century a corrective to Mommsen's Caesar worship, saw Pompey in his later years as the forerunner of Augustus. The German scholar Matthias Gelzer in the mid-20th century has stressed the early influence of Pompeius Strabo. Sulla's example, too, was an all-important determinant in Pompey's career. But he was neither a revolutionary nor a reactionary, willing to wreck the fabric of the commonwealth for the advantage of self or class. He expected a voluntary acceptance of his primacy but was to discover that the methods he had used to get his commands had permanently alienated the dominant nobility. So year after year he had to play a passive role, covertly intriguing or waiting for successive occasions to arise that would force them to accept his leadership. Some thought his waiting game duplicity, others, sheer political incompetence. He was an ineffective politician, not from any incapacity for intrigue or ruthless action but from lack of candour and consistency in speech and action. He saw the need for basic solutions to the formidable problems presented by the rapid expansion of the empire, but his talents were suited not for the elucidation and determining of ends but for the deployment of means. In domestic affairs his preference was for constructive compromise and reforms of a kind that fitted into and gave additional momentum to the slowly evolving process of adaptation of Roman city-state institutions to the needs of a large empire. The crisis of state and empire demanded more drastic remedies than he was able to envisage or willing to apply.

Comparison with Caesar

As a military leader, he fell short of real greatness, lacking Caesar's genius, his dynamism and panache, and his geniality in personal relationships. He was circumspect and thorough—the perfect administrator. His vision of empire was no narrower than Caesar's. Like many a more recent imperialist, he was satisfied with the ideal of efficient and clean-handed administration and justice, and many of his contemporaries believed that he went far to achieve that aim in his own practice. Pompey, by far the wealthiest man of his age, invested his millions prudently; his landed estates were distributed throughout Italy in manageable units. For all the extravagance of his triumphal shows and the inexcusable heartlessness of the contests in slaughter with which he entertained the populace, he was a plain living man, friend and admirer of the Stoic Panaetius.

His third wife, Mucia, bore him two sons, Gnaeus and Sextus, and a daughter, Pompeia, before he divorced her for infidelity (62). Julia was the wife he loved most dearly; Cornelia outlived him and mourned his death.

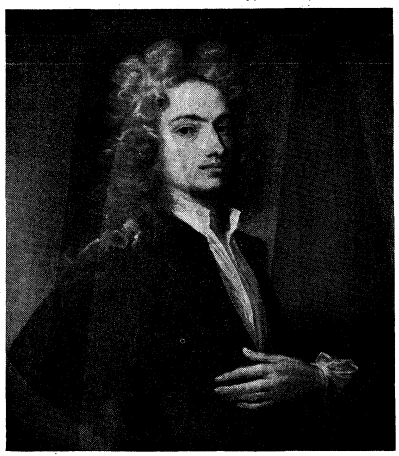
BIBLIOGRAPHY. Of ancient sources PLUTARCH'S *Lives* (of Pompey and his Roman contemporaries) are invaluable; also the *Histories* of CASSIUS DIO (Book 36–42); APPIAN'S *Civil Wars*, I and II; and CAESAR'S *Civil War*; but Cicero's speeches and letters throw most light on Pompey's role in the public life of his day. For background to his career, see *The Cambridge Ancient History*, vol. 9 (1932); or more briefly, RONALD SYME, *The Roman Revolution*, ch. 2–5 (1939), a masterly analysis. MOMMSEN did scant justice to Pompey in his *Roman History*, Book 5, now partly superseded by more recent scholarship. W. DRUMANN and P. GROEBE, *Geschichte Roms*, 2nd ed., vol. 4, pp. 332–561 (1908), is a well-documented historical biography, as is the more recent survey by F. MILTNER in *Pauly-Wissowa Realencyclopädie*, vol. 21, col. 2062–2211 (1952). The title of EDUARD MEYER, *Caesars Monarchie und das Principat des Pompejus* (1918), indicates the thesis of that classic work and its limitations. MATTHIAS GELZER, *Pompeius*, 2nd ed. (1959), is a balanced and realistic political biography. JULES VAN OOTEGHEM, *Pompée le grand* (1954), breaks little new ground, but is useful for its overall coverage, its brief discussion of the iconography of Pompey, and its bibliography of articles on points of detail. See also, for Pompey's operations in the East, DAVID MAGIE, *Roman Rule in Asia Minor*, vol. 1, ch. 15–16 (1950); and, for the Civil War, THOMAS RICE HOLMES, *The Roman Republic and the Founder of the Empire*, vol. 3, ch. 14–19 (1923).

(E.W.G.)

Pope, Alexander

The Augustan age in English literature produced no greater satirical poetry than that of Alexander Pope. He became renowned for his translations of Homer and for his biting rebuttals to the criticism of numerous literary enemies, but is best remembered today for his mock-epic poem *The Rape of the Lock* and *The Dunciad* and for his satires in imitation of Horace. Pope was born in Lombard Street, London, on May 21, 1688. His father, a wholesale linen merchant, retired from business in the year of his son's birth and in 1700 went to live at Binfield in Windsor Forest. The Popes were Roman Catholics, and at Binfield they came to know several neighbouring Catholic families who were to play an important part in the poet's life. Pope's religion procured him some lifelong friends, notably the wealthy squire John Caryll (who persuaded him to write *The Rape of the Lock*, on an incident involving Caryll's relatives) and Martha Blount, to whom Pope addressed some of the most memorable of his poems—the satire on women, entitled "Of the Characters of Women," "To Mrs. M.B. on Her Birthday," and "An Epistle to Miss Blount with the Works of Voiture"—and to whom he bequeathed most of his property. But his religion also precluded him from a formal course of education; he was trained at home by Catholic priests for a short time and attended Catholic schools at Twyford, near Winchester, and at Hyde Park Corner, London, but he was mainly self-educated. He was a precocious boy, eagerly reading Latin, Greek, French, and Italian, which he managed to teach himself, and an incessant scribbler, turning out verse upon verse in imitation of the poets he read. The best of these early writings are the "Ode on Solitude" and a paraphrase of St. Thomas à Kempis, both of which he claimed to have written at the age of 12. But perhaps in later life he believed himself to have been more precocious than he was.

By courtesy of the curators of the Bodleian Library, Oxford, Oxfordshire



Pope, oil painting by Charles Jervas. In the Bodleian Library, Oxford University.

Early works. Windsor Forest was near enough to London to permit Pope's frequent visits there. He early grew acquainted with former members of Dryden's circle, notably William Wycherley, William Walsh, and Henry Cromwell, the recipients of his earliest surviving letters, which are devoted to a somewhat mannered discussion of the art of poetry. By 1705 his "Pastorals" were in draft and were circulating among the best literary judges of the day. In 1706 Jacob Tonson, the leading publisher of poetry, had solicited their publication, and they took the place of honour in his *Poetical Miscellanies* in 1709.

Pope's physical disabilities

This early emergence of a man of letters may have been assisted by Pope's poor physique. As a result of too much study, so he thought, he acquired a curvature of the spine and some tubercular infection, probably Pott's disease, that limited his growth and seriously impaired his health. His full-grown height was four feet six inches; but the grace of his profile and fullness of his eye attracted numerous painters (such as Sir Godfrey Kneller and Charles Jervas) and sculptors (such as John Michael Rysbrack and Louis-François Roubillac), who transmitted his likeness in such variety as to make his iconography more extensive than that of any other English writer perhaps to the present day. He was a lifelong sufferer from headaches, and his deformity made him abnormally sensitive to physical and mental pain; and though he was able to ride a horse and delighted in travel, he was inevitably precluded from much normal physical activity, and his energetic, fastidious mind was largely directed to reading and writing.

When the "Pastorals" were published, Pope was already at work on a poem, in an equally well-established tradition, on the art of writing. This was *An Essay on Criticism*, published in 1711. Its brilliantly polished epigrams (*e.g.*, "A little learning is a dangerous thing," "To err is human, to forgive, divine," and "For fools rush in where angels fear to tread"), which have become part of the proverbial heritage of the language, are readily traced to their sources in Horace, Quintilian, Boileau, and other critics, ancient and modern, in verse and prose; but the charge that the poem is derivative, so often made in the past, takes insufficient account of Pope's success in harmonizing a century of conflict in critical thinking and in showing how nature may best be mirrored in art.

Essay on Criticism

The well-deserved success of the *Essay on Criticism* brought Pope a wider circle of friends, notably Richard Steele and Joseph Addison, who were then collaborating in *The Spectator*. To this journal Pope contributed the most original of his pastorals, "The Messiah" (1712), and perhaps other papers in prose. He was clearly influenced by its policy of correcting public morals by witty admonishment and in this vein wrote the first version of his mock-epic, *The Rape of the Lock* (two cantos, 1712; five cantos, 1714), to reconcile two Catholic families. A young man in one family had stolen a lock of hair from a young lady in the other. Pope treated the dispute that followed as though it were comparable with the mighty quarrel between Greeks and Trojans, which had been Homer's theme. Telling the story with all the pomp and circumstance of epic made not only the participants in the quarrel but also the society in which they lived seem ridiculous. Though it was a society where

Publication of *The Rape of the Lock*

> . . . Britain's statesmen oft the fall foredoom
> Of foreign tyrants, and of nymphs at home;

as if one occupation concerned them as much as the other; and though in such a society a young lady might do equally ill to

> . . . Stain her honour, or her new brocade;
> Forget her pray'rs, or miss a masquerade;

Pope managed also to suggest what genuine attractions existed amid the foppery and glitter. He acknowledged how false the sense of values was that paid so much attention to external appearance but ridicule and rebuke slide imperceptibly into admiration and tender affection as the heroine, Belinda, is conveyed along the Thames to Hampton Court, the scene of the "rape":

> But now secure the painted vessel glides,
> The sunbeams trembling on the floating tides:
> While melting music steals upon the sky,
> And soften'd sounds along the waters die;
> Smooth flow the waves, the zephyrs gently play,
> Belinda smil'd, and all the world was gay.

A comparable blend of seemingly incompatible responses—love and hate, bawdiness and decorum, admiration and ridicule—is to be found in all Pope's later satires.

Pope had also been at work for several years on "Windsor-Forest." In this poem, completed and published in 1713, he proceeded, as Virgil had done, from the pastoral vein to the georgic and celebrated the rule of Queen Anne as the Latin poet had celebrated the rule of Augustus. By choosing for its background a region so rich in historical associations, he was enabled to prophesy the greater glories of a regenerate and peaceful society.

In another early poem, "Eloisa to Abelard," Pope borrowed the form of Ovid's "heroic epistle" (in which an abandoned lady addresses her lover) and showed imaginative skill in conveying the struggle between sexual passion and dedication to a life of celibacy.

Translation of Homer. These poems, with "Verses to the Memory of an Unfortunate Lady" (modelled on the elegies of Ovid and Tibullus), some translations from the classics, and a few shorter pieces, were collected in the first volume of Pope's *Works* in 1717. This is a volume of experiments in various poetical kinds; it might almost be called a volume of juvenilia. When it was published, he was already far advanced with the greatest labour of his life, his verse translation of Homer. He had announced his intentions in October 1713 and had published the first volume, containing the *Iliad*, Books I–IV, in 1715. The *Iliad* was completed in six volumes in 1720. The work of translating the *Odyssey* (vol. i–iii, 1725; vol. iv and v, 1726) was shared with William Broome, who had contributed notes to the *Iliad*, and Elijah Fenton. The labour had been great, but so were the rewards. By the two translations Pope cleared about £10,000 and was able to claim that, thanks to Homer, he could ". . . live and thrive/Indebted to no Prince or Peer alive."

The merits of Pope's Homer lie less in the accuracy of translation and in correct representation of the spirit of the original than in the achievement of a heroic poem as his contemporaries understood it; a poem Virgilian in its dignity, moral purpose, and pictorial splendour, yet one that consistently kept Homer in view and alluded to him throughout. Pope offered his readers the *Iliad* and the *Odyssey* as he felt sure Homer would have written them had he lived in early 18th-century England. In Dr. Johnson's estimate, he achieved "the noblest version"—*i.e.*, translation—"of poetry which the world has ever seen."

The Homer quarrel

Political considerations had affected the success of the translation. As a Roman Catholic his affiliations were Tory rather than Whig; and though he retained the friendship of such Whigs as William Congreve, Nicholas Rowe, and Charles Jervas the painter, his ties with Steele and Addison grew looser owing to the political animosity at the end of Queen Anne's reign, and he found new and lasting friends in Tory circles—Swift, John Gay, John Arbuthnot, Thomas Parnell, the Earl of Oxford, and Viscount Bolingbroke. With the first five he was associated (1713–14) in the Scriblerus Club to write joint satires on pedantry, later to mature as *Peri Bathouse, Or the Art of Sinking in Poetry* (1728) and the "Memoirs of Martinus Scriblerus" (1741); and these were the men who encouraged his translation of Homer. The Whigs, who associated with Addison at Button's Coffee-House, put up a rival translator in Thomas Tickell, who published his version of *Iliad,* Book I, two days after Pope's. Addison preferred Tickell's manifestly inferior version; his praise increased the resentment Pope already felt owing to a series of slights and misunderstandings; and when Pope heard gossip of further malice on Addison's part, he sent him a satirical view of his character, published later as the character of Atticus, the insincere arbiter of literary taste in the "Epistle to Dr. Arbuthnot" (1735), but with the most topical couplet omitted:

> Who when two wits on rival themes contest,
> Approves them both, but likes the worst the best.

The stages in this famous quarrel and the partial reconciliation that followed remain obstinately obscure; but Pope's subsequent praise of Addison in the verse epistle first published in Addison's *Works* (1721) and in the "Epistle to Augustus" (1737) is generous and genuine.

"The Dunciad." Even before the Homer quarrel, Pope had found that the life of a wit was a perpetual warfare. There were few years when either his person or his poems were not objects of attacks from the critic John Dennis, the notorious bookseller Edmund Curll, the historian John Oldmixon, and other writers of lesser fame. The climax was reached over his edition of Shakespeare. He had emended the plays, in the spirit of a literary editor, to accord with contemporary taste (1725); but his practice was exposed by the scholar Lewis Theobald in *Shakespeare Restored* (1726). Though Pope had ignored some of these attacks, he had replied to others with squibs in prose and verse. But he now attempted to make an end of the opposition, and to defend his standards, which he aligned with the standards of civilized society, in the mock-epic, *The Dunciad* (1728). Theobald was represented in it as the Goddess of Dullness' favourite son, a suitable hero for those leaden times; and others who had given offense were preserved like flies in amber. Pope dispatches his victims with such sensuousness of verse and imagery that the reader is forced to admit that if there is petulance here, as has often been claimed, it is, to parody Wordsworth, petulance recollected in tranquillity. Pope reissued the poem in 1729 with an elaborate mock-commentary of prefaces, notes, appendixes, indexes, and errata; this burlesque of pedantry whimsically suggested that *The Dunciad* had fallen a victim to the spirit of the times and been edited by a dunce.

Pope's edition of Shakespeare

His life at Twickenham. Pope and his parents had moved from Binfield to Chiswick in 1716. There his father died (1717), and in 1719 he and his mother rented a villa on the Thames at Twickenham, at that time a small country town where several Londoners had retired to live in rustic seclusion. This was to be Pope's home for the remainder of his life. There he entertained with studied modesty such friends as Swift (who returned from Ireland in 1726 and stayed with Pope while seeing to the publication of *Gulliver's Travels*), Bolingbroke, Oxford, and Jonathan Richardson the painter, regaling them with gudgeons and flounders from the Thames at his door and with figs and walnuts from the little garden that he took such pains to design and cultivate. The famous grotto, which provided a tunnel under the main highway and formed a study, ornamented with rare minerals, is all that now survives of this domain. Pope kept his own boatman to row him downstream on his frequent visits to London, and his own coach to take him on his customary round of summer visits to country houses: the Duke of Argyll's at Adderbury, Lord Bathurst's at Cirencester, Lord Cobham's at Stowe, Ralph Allen's at Prior Park, Bath, and Lord Peterborough's at Southampton. These friends were all enthusiastic gardeners, and it was Pope's pleasure to advise and superintend the laying out of their grounds on the best contemporary principles, formulated in his "Epistle to the Right Honourable Richard Earl of Burlington" (1731). This poem, one of the most characteristic works of his maturity, is a rambling discussion in the manner of Horace on false taste in architecture and design, with some suggestions for the worthier employment of a nobleman's wealth.

Pope's plan for the *Essay on Man*

Pope now began to contemplate a new work on the relations of man, nature, and society. This was evidently to be such another grand organization of human experience and intuition as Milton had attempted in the 17th century, and Wordsworth was to attempt in the 19th; but he was destined never to complete it. He told Joseph Spence that the following lines, later incorporated in the *Essay on Man,* would have conveyed its dominant tone and interest:

> Laugh where we must, be candid where we can;
> But vindicate the ways of God to man.

An Essay on Man (1733–34) was intended as the introductory book discussing the grand design. The poem has often been charged with shallowness and philosophical inconsistency. There is indeed little that is original in the thought, almost all of which can be traced in the work of the great thinkers of Western civilization. It is not strictly an argumentative poem as much as a discourse or extended monologue delivered by the poet to a friendly auditor. As becomes a good conversationalist, he alters his mood from time to time, now grave, now gay, colloquial at one moment, formal at another, now affirmative, now quizzical. He deliberately prefers the middle flight to the sublime; but within that range he uses all the arts of diction and versification, the charm of imagery, and variation of rhythm to effect his persuasion. Subordinate themes were treated in greater detail in "Of the Use of Riches, An Epistle to Bathurst" (1732), "An Epistle to Cobham, Of the knowledge and characters of men" (1733), and "Of The Characters of Women: an Epistle to a Lady" (1735). He was deflected from this "system of ethics in the Horatian way" by the renewed need for self-defense. The character of Timon, an ostentatious nobleman in the "Epistle to Burlington," was reported to represent the

Duke of Chandos, who had generously subscribed to Pope's Homer. Though the report was malicious, it was plausible enough to rouse Pope's apprehensions. This attack and others of a like nature drove him to consider his position as satirist. He chose to adapt for his own defense the first satire of Horace's second book, where the ethics of satire are propounded, and after discussing the question in correspondence with Dr. John Arbuthnot, he addressed to him an epistle in verse (1735), one of the finest of his later poems, in which were incorporated fragments written over several years. His case was the satirist's traditional case, that depravity in public morals had roused him to stigmatize outstanding offenders beyond the reach of the law, concealing the names of some and representing others as types, that he was innocent of personal rancour and habitually forbearing under attack.

Imitations of Horace

The success of his "First Satire Of the Second Book Of Horace, Imitated" (1733) led to the publication (1734–38) of ten more of these paraphrases of Horatian themes adapted to the contemporary social and political scene. Pope's poems followed Horace's satires and epistles sufficiently closely for him to print the Latin on facing pages with the English; but whoever chose to make the comparison would notice a continuous enrichment of the original by parenthetic thrusts and compliments, as well as by the freshness of the imagery. Thus Horace declares that he is a peace-loving fellow, "but he who provokes me (better not touch me, I cry) will weep for it and be well-known in song throughout the city"; while Pope, in following him, not only glances at the King's first minister, Sir Robert Walpole, and his French counterpart, Cardinal Fleury (at that time both anxious for peace), but enlivens the whole passage:

> Peace is my dear delight—not Fleury's more:
> But touch me, and no Minister so sore,
> Whoe'er offends, at some unlucky time
> Slides into verses, and hitches in a rhyme,
> Sacred to ridicule his whole life long,
> And the sad burden of some merry song.

The series was concluded with two dialogues in verse, republished as the "Epilogue to the Satires" (1738), where, as in the "Epistle to Dr. Arbuthnot," Pope ingeniously combined a defense of his own career and character with a restatement of the satirist's traditional apology. These imitations and dialogues show Pope at his easiest and most genial in manner. In them he directed his attack upon the materialistic standards of the commercially minded Whigs in power and upon the corrupting effect of money, while restating and illustrating the old Horatian standards of serene and temperate living. His anxiety about prevailing standards was shown once more in his last completed work, *The New Dunciad* (1742), reprinted as the fourth book of a revised *Dunciad* (1743), in which Theobald was replaced as hero by Colley Cibber, the poet laureate and actor-manager, who had not only given more recent cause of offense but seemed a more appropriate representative of the degenerate standards of the age. In *Dunciad*, Book IV, the Philistine culture of the city of London was seen to overtake the court and seat of government at Westminster, and the poem ends in a magnificent but baleful prophecy of anarchy. He had begun work on *Brutus*, an epic poem in blank verse, and on a revision of his poems for a new edition, but neither was complete on his death at Twickenham on May 30, 1744.

The New Dunciad

Assessment. In the "Epistle to Dr. Arbuthnot" and the "Imitations of Horace," Pope had presented himself to his readers as he would have them see him. He supplemented this defense of his character by editing a selection of his letters. To publish them himself seemed an offense against modesty; he therefore arranged for copies to be placed in the hands of Curll, the bookseller whose taste for piratical publication ensured that a "surreptitious" edition would be produced. The stratagem succeeded (1735) and enabled Pope to complain that he was injured in his reputation and must bring out a correct and authentic edition (1737). Other stratagems were required to secure the publication of his letters to Swift (1741) without his open connivance. Collation of the printed texts with such manuscripts as have survived shows that Pope published a text not substantially different from that which his correspondents read, though he felt himself free to repair his syntax and to readdress a few letters from correspondents who had kept many to correspondents who had kept none. These and his many other surviving letters reveal him at first as a young man labouring successfully to be a wit in the Restoration manner, but show him later in a more sympathetic light as a dutiful son, a kind and thoughtful friend, a well-bred host, a disinterested critic of society, warm in wishes for his country's good, patient under attack, a man of plain living, high thinking, and unimpeachable integrity. Not every critic of Pope has been able to accept this picture; his friendliest critics could wish that he had been less accustomed to equivocation and devious dealings; but when allowances have been made for this and for the too relentless pursuit of some who had injured him—Lady Mary Wortley Montagu, for example, though why Pope came to hate this famous bluestocking for whom he had once expressed a gallant admiration is still obscure—the picture obtained from the letters is convincing, though somewhat conscious in its pose. The picture is paralleled in the later poems, where Pope frequently describes himself as a model of modest, honourable, contented living, a survival of old Roman simplicity in an age of degenerate taste.

Personality and reputation

Pope's favourite, but not invariable, metre was the ten-syllable, iambic pentameter, rhyming ("heroic") couplet. He handled it with increasing skill and adapted it to such varied purposes as the epigrammatic summary of the *Essay on Criticism*, the pathos of "Verses to the Memory of an Unfortunate Lady," the mock-heroic of *The Rape of the Lock*, the discursive tones of the *Essay on Man*, the rapid narrative of the Homer translation, and the Miltonic sublime of the conclusion of *The Dunciad*. But his greatest triumphs of versification are found in the "Epilogue to the Satires," where he moves easily from witty, spirited dialogue to noble and elevated declamation, and in the "Epistle to Dr. Arbuthnot," which opens with a scene of domestic irritation suitably conveyed in broken rhythm:

> Shut, shut the door, good John! fatigu'd, I said:
> Tie up the knocker, say I'm sick, I'm dead.
> The Dog-star rages! nay 'tis past a doubt,
> All Bedlam, or Parnassus, is let out:
> Fire in each eye, and papers in each hand,
> They rave, recite, and madden round the land;

and closes with a deliberately chosen contrast of domestic calm, which the poet may be said to have deserved and won during the course of the poem:

> Me, let the tender office long engage
> To rock the cradle of reposing age,
> With lenient arts extend a mother's breath,
> Make languor smile, and smooth the bed of death,
> Explore the thought, explain the asking eye,
> And keep a while one parent from the sky!

His command of diction is no less happily adapted to his theme and to the type of poem, and the range of his imagery is remarkably wide. He has been thought defective in imaginative power, but this opinion cannot be sustained in view of the invention and organizing ability shown notably in *The Rape of the Lock* and *The Dunciad*. His standing as a poet was recognized not only at home. He was the first English poet to enjoy contemporary fame in France and Italy and throughout the Continent, and to see translations of his poems into modern as well as ancient languages. Soon after his death different standards began to prevail; but even when his stock was at its lowest, in the 19th century, he found such champions for his poetry as Byron, Lamb, and William Hazlitt, and, for his letters, as Thackeray. In the 20th century he has long been restored to the esteem, if not to the pre-eminence, he enjoyed in his lifetime.

MAJOR WORKS

VERSE: *An Essay On Criticism* (1711); *The Rape of the Lock* (1712; 2nd enlarged version, 1714); "Windsor-Forest" (1713); "Ode for Musick (on St. Cecilia's Day)" (1713); "Verses to the Memory of an Unfortunate Lady" and "Eloisa to Abelard" (published in a collected works, 1717); translations from *The Iliad Of Homer*, 6 vol. (1715–20), and

The Odyssey of Homer, 5 vol. (1725–26); *The Dunciad*, 3 books (1728), *The Dunciad Variorum* (1729); *The New Dunciad* (Book iv, 1742); the *Dunciad* (complete, 4 books, 1743); "Moral Essays" (1731–35); "An Epistle to Dr. Arbuthnot" (dated 1735).

PROSE: *The Works of Mr. Alexander Pope, In Prose* (vol. 1, 1737; vol. 2, 1741); prefaces to *The Iliad* (1715), *The Odyssey* (1725), Shakespeare (1725); *Mr. Pope's Literary Correspondence* (1735–37).

BIBLIOGRAPHY. The standard bibliography of Pope's own writings is R.H. GRIFFITH, *Alexander Pope: A Bibliography*, 2 vol. (1922–27). A valuable list of modern studies of the poet and his works is contained in J.E. TOBIN, *Alexander Pope: A List of Critical Studies Published from 1895 to 1944* (1945); and in C.L. LOPEZ, *Alexander Pope: An Annotated Bibliography, 1945–67* (1970). J.V. GUERINOT, *Pamphlet Attacks on Alexander Pope, 1711–1744: A Descriptive Bibliography* (1969), gives a memorable account, with quotations, of the kinds of hostility that the poet's actions and his meteoric success generated among the lesser writers of his time and so has biographical and critical as well as bibliographical value.

Letters and manuscripts: Pope's letters from his boyhood to his death (1704–44) may be consulted in the edition of his *Correspondence* by GEORGE SHERBURN, 5 vol. (1956). Several of his poetical manuscripts have been published in facsimile, with transcript and editorial commentary: *Windsor Forest*, ed. by R.M. SCHMITZ (1952); the *Epistle to Bathurst*, ed. by E.R. WASSERMAN (1960); *An Essay on Criticism*, ed. by R.M. SCHMITZ (1962); *An Essay on Man*, ed. by MAYNARD MACK (1962). A census of the surviving poetical manuscripts is contained in JOHN BUTT, "Pope's Poetical Manuscripts," *Proceedings of the British Academy*, 40:23–39 (1954).

Editions and selections: Pope published the first collected edition of his poems in 1717, adding a second volume of poems in 1735, a volume of letters in 1737, and a volume of further letters and miscellaneous prose in 1741. After 1734 not a year passed until his death without some earlier writings being reissued, so great was the demand. The standard edition of the poems today is the Twickenham Edition, general editor, JOHN BUTT, 11 vol. (1939–67). There is a one-volume version of the complete text of the first six volumes with selected notes, ed. by JOHN BUTT (1963). *The Prose Works of Alexander Pope*, vol. 1, *The Earlier Works, 1711–20*, ed. by NORMAN AULT, was published in 1936 (reprinted 1968); vol. 2 is now in preparation. Two of his most important prose works have been edited and published separately: *The Memoirs of Martinus Scriblerus*, by CHARLES KERBY-MILLER (1950, reprinted 1966); and *The Art of Sinking in Poetry (Peri Bathous)*, by E.L. STEEVES (1952).

Biography: The best of the early biographies was that of DR. JOHNSON, in his *Lives of the English Poets* (1781); he had access to the most important primary document (apart from Pope's letters), JOSEPH SPENCE'S *Anecdotes* of his conversations with Pope and others, which was eventually published, ed. by S.W. SINGER (1820); a much fuller edition from the manuscripts, by J.M. OSBORN, appeared in 1966. New facts were gathered by ROBERT CARRUTHERS in his *Memoir of Pope*, 2 vol. (1853; rev. ed., 1857); by C.W. DILKE, whose papers, mainly on the authenticity of Pope's letters, were published as *The Papers of a Critic*, 2 vol. (1875); and by NORMAN AULT in *New Light on Pope* (1949). The standard biography so far as it goes (1727) is GEORGE SHERBURN, *The Early Career of Alexander Pope* (1934, reprinted 1969), which may be supplemented for Pope's life at Twickenham by MAYNARD MACK, *The Garden and the City: Retirement and Politics in the Later Poetry of Pope, 1731–1743* (1969).

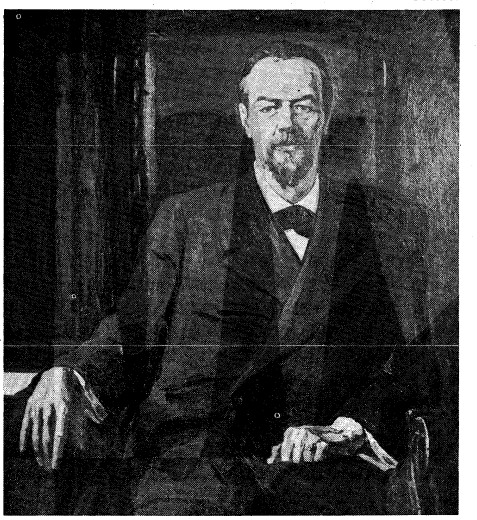
Criticism: The first substantial study of Pope's poetry was by JOSEPH WARTON in *An Essay on the Writings and Genius of Pope*, 2 vol. (1756–82; 4th corrected ed. reprinted 1970). Of 20th-century assessments, the most important are GEOFFREY TILLOTSON, *On the Poetry of Pope*, rev. ed. (1950), and *Pope and Human Nature* (1958); and R.A. BROWER, *Alexander Pope: The Poetry of Allusion* (1959). Of many shorter studies the most rewarding are by GEORGE SHERBURN, "Pope at Work," in *Essays on the Eighteenth Century Presented to David Nichol Smith* (1945); and by MAYNARD MACK, "Wit and Poetry and Pope: Some Observations on His Imagery," in *Pope and His Contemporaries: Essays Presented to George Sherburn* (1949), and "The Muse of Satire," *Yale Review*, 41:80–92 (1951). On specific poems the following studies are helpful: E.N. HOOKER, "Pope on Wit: *The Essay on Criticism*," in *The Seventeenth Century: Studies by R.F. Jones and Others* (1951); D.M. KNIGHT, *Pope and the Heroic Tradition* (1951), on the *Iliad*; A.L. WILLIAMS, *Pope's Dunciad* (1955); ROBERT W. ROGERS, *The Major Satires of Alexander Pope* (1955); T.R. EDWARDS, JR., *This Dark Estate: A Reading of Pope* (1963), mainly on the satires and epistles; PETER DIXON, *The World of Pope's Satires* (1968). There is a useful reprint of critical studies entitled *Essential Articles for the Study of Alexander Pope*, ed. by MAYNARD MACK (1964, rev. and enl. 1968).

(J.E.Bu.)

Popov, Aleksandr Stepanovich

A Russian physicist and electrical engineer who is remembered for his pioneering investigation of electromagnetic waves and acclaimed in the Soviet Union as the inventor of radio, Popov in 1895 devised equipment to receive electromagnetic (wireless or radio) waves sent out by giant electric sparks forming lightning flashes; this detector thus preceded the first patent taken out by the Italian physicist Guglielmo Marconi in June 1896 for the transmission and reception of signals by means of wireless (radio) telegraphy.

Sovfoto



Popov, painting by P.D. Korin (1892–1967).

Early years and education

Aleksandr Stepanovich Popov was born on March 16 (March 4, old style), 1859, at Bogoslavsky, Perm, a mining village in the Ural Mountains of Russia, where his father was a priest. He received his early education in an ecclesiastical seminary school and planned to enter the priesthood. But in 1877 his interests changed to mathematics and he entered St. Petersburg University, from which he was graduated with distinction in 1883. Joining the teaching faculty of the university, he lectured in mathematics and physics in preparation for a professorship.

Popov's main interest soon changed to electrical engineering, however; and because Russia in that period lacked colleges that taught the subject, he became an instructor at the Russian Navy's Torpedo School at Kronstadt (presently Kronshtadt), near St. Petersburg (presently Leningrad) where students were trained to take charge of electrical equipment on warships. Popov took advantage of the school's library, which was stocked with foreign books and periodicals, and also of its well-equipped laboratory to follow scientific developments abroad and carry out experiments. Recognizing the importance of Heinrich Hertz' discovery of electromagnetic waves in Germany, Popov began to work on methods of receiving them over long distances.

Popov constructed an apparatus that could register atmospheric electrical disturbances and, in July 1895, installed it at the meteorological observatory of the Institute of Forestry in St. Petersburg. In a paper published a few months later, Popov suggested that such an apparatus could be used for the reception of signals from a man-made source of oscillations, provided a sufficient power source became available. Later, in March 1896, he appeared before the St. Petersburg Physical Society and demonstrated the transmission of Hertzian waves—as they were then termed—between different parts of the St.

Petersburg University buildings. There is evidence to suggest that on that occasion the words "Heinrich Hertz" were transmitted in Morse code and that the aural signals received were transcribed on a blackboard by the society's president, who was the chairman of the meeting.

During the academic year 1895–96 at the Torpedo School, however, Popov became interested in setting up experiments on Röntgen rays (X-rays), which had just been discovered. Therefore, he discontinued for a time the further development of his lightning, or thunderstorm, detector. During the summer of 1896, Popov worked as electrical engineer in charge of the power plant at the annual fair in Nizhnii Novgorod (presently Gorky) where the demonstration of his storm detector won him a prize. He then read the first newspaper accounts of Marconi's demonstrations in September 1896. It seems clear that neither Marconi nor Popov was aware of the close similarity between their experiments.

Similarity between Popov's and Marconi's experiments

The news of Marconi's work, as disclosed in his patent of June 1896, aroused Popov to fresh activity. Working in conjunction with the Russian Navy, he effected ship-to-shore communication over a distance of six miles (ten kilometres) by 1898. The distance was increased to about 30 miles (50 kilometres) by the end of the following year, during which he had also visited wireless stations in operation in France and Germany.

Popov was given remarkably little support by the Russian government until 50 years later, when national attitudes and enthusiasms had changed and on May 7, 1945, the Bolshoi Theatre was packed with a distinguished audience to celebrate the 50th anniversary of the "invention of the radio" by A.S. Popov. On the stage sat scientists, marshals, admirals, commissars, leaders of the Communist Party, and Popov's daughter. It was announced that in the future, the 7th of May would be celebrated as the day of the radio.

Although it is agreed that Popov's experimental work in connection with Hertzian waves is deserving of recognition, it has not been generally accepted that radio communication was actually invented by Popov. Thus, while it is true that historical research has brought to light indirect evidence that Popov demonstrated the transmission of intelligible signals in March 1896, there is comparable evidence that Marconi demonstrated the transmission of intelligible signals at an even earlier date, though not before an audience of scientists.

In 1901 Popov returned to St. Petersburg as a professor at the electrotechnical institute, of which he was later elected director. He died on January 13 (December 31, 1905, O.S.), 1906.

BIBLIOGRAPHY. CHARLES SUSSKIND, "Popov and the Beginnings of Radiotelegraphy," *Proc. Inst. Radio Engrs.*, 50:2036–2047 (1962), a comprehensive and critical review of the early days of wireless communication, with special reference to the Soviet claim that Popov invented radio; R.L. SMITH-ROSE, "Marconi, Popov, and the Dawn of Radiocommunication," *Electronics and Power*, 10:76–79 (1964), a review of the claims of the two pioneers, G. Marconi and A.S. Popov, with an assessment of credits as the founders of practical radio communications.

(R.L.S.-R.)

Popular Dance

The term popular dance refers to dance performed in public or in private solely for the enjoyment of the dancers. Such dance requires no audience nor any special occasion. Its movements are usually simple, but they are sometimes complex in rhythm. It is at times described as social dance, although that term may also embrace folk dances, which are social in function and the forms of which are ancestral to many popular dances. A detailed coverage of these traditional dances may be found in the article FOLK DANCE. In strictest definition, popular dance developed entirely within the West until, in the 20th century, its forms were carried around the world.

CHARACTERISTICS OF POPULAR DANCE

Unlike folk dance, popular dance is neither ritual nor functional, although it has developed or borrowed from traditional sources and may be danced on occasions that were once a part of folk rites. Unlike ballet and other theatrical forms of dance, popular dance requires no strict physical or mental discipline. Its movements usually require little or no training, nor do they express mood, emotion, or action. From the 13th to the 18th century, however, many popular dances originated in court ballrooms, where niceties of behaviour and of dress were, like the dances themselves, subject to rules of the dancing masters. These dances required much training and practice, and dance was an important part of the education of the young nobles.

Bases. The bases of popular dance are numerous. First, popular dance helps to fulfill the need for identity with communal activities. Second, it is related to courting, in which the male displays his ability to move his partner in harmony with himself. In much social dance the male role is dominant, the female that of a follower. Third, some popular dancing, especially before the 20th century, was to celebrate such events as weddings, harvests, or merely the end of a working day.

Change. The forms of popular dance are susceptible to change from many sources. The migration of the performers or the adoption of a dance by a different nation or social class may subject it to new social situations and political institutions. The process of urbanization, the broadening of education, and the ease of communication not only altered traditional forms but also disseminated new dances rapidly. The primary vehicle for change in popular dance, however, always has been youth.

Youth always has tended to rebel against the dancing masters, whose insistence on correctness and uniformity of execution are contrary to the spontaneous responses to rhythm sought by the young. The teaching of these dancing masters led to concentration on performance, to a kind of studied dance that became largely an entertainment for onlookers. Since popular dance has always been, at base, for the participants, demands for correctness of performance rather than enjoyment caused youth continually to seek less conventionalized forms, often from among the dances of another nation or social group.

Music. Unlike the ritual dancer, who guides the music or actually supplies it, the ballroom dancer follows the music. The timing of dance rhythms has varied as fashions have changed or as wars and political upheavals have altered the structure of society. No matter how the rhythms may be beaten out, the dancers have always responded to the lilt of the melody with moving feet and body until they are caught up in the spirit of the dance.

HISTORY OF POPULAR DANCE

The concept "popular dance" cannot properly be applied to most of the dancing of prehistoric or ancient societies, even in the West, because of the ritual elements that pervaded it. Even the Greek Dionysiac dancing at the grape harvest and the Roman floralia at the advent of spring were essentially ritual, despite their popular and orgiastic nature. There was some secular dancing in Greece, but its forms and its festivities placed it in the tradition of the folk dance. Some of these dances persisted into the 20th century and spawned such other dances as the Provençal farandole and the Catalan sardana. A dance of sacrificial ritual survived as a dance by men in Greek seaport taverns. The Cornish Furry Dance, a flower dance, is believed to be descended from the Roman floralia. In ancient Greece, as in later centuries, dance had a place in educating youth. Plato's dictum "dance begets a healthy mind in a healthy body" was echoed by the Greek slave Lucian to his Roman masters. The Romans, however, contributed little to dance and tended to regard it with disfavour.

The dances of the early Middle Ages were often thinly disguised pre-Christian forms that included animal dances, fertility rites, and the like. The church countenanced dancing on social occasions because this popular means of celebrating could not be ignored, but it tried to eradicate the pagan elements.

Early court dances. The court dance, from which the ballroom dance was to develop, evolved about the 13th

century from the common chain and circle dances of the townsfolk and peasants. These traditional dances were consciously changed by such mentors as the troubadours, who imposed codes of courtly behaviour onto the dances.

Chivalry in dancing

These dances followed the codes by emphasizing the strength of the male and the delicacy of the lady. They contained elaborate displays of courtesy between the sexes and indicated the order of social precedence, from the lord and his lady to the squires and ladies-in-waiting. In addition, the dances accommodated the robust peasant steps to the court costumes, with their greater weight and ornament. The "Lay of Robin and Marion" (1285) mentions such new dances as the carole, a solemn stepping in triple time in which ladies and gentlemen, standing alternately hand in hand, circled about gracefully; and the stately courtesia, in which the couples interrupted their slow procession to express their affections through a stylized sequence of mimetic gestures. Such dances usually were accompanied by religious songs of grave tone and measure suited to the slow, gliding steps and the gentle rise and fall of the bodies. A musical setting for the 109th Psalm, of folk origin, was a favourite tune for these dances.

The troubadours may be considered the first of the dancing masters, who subsequently became an institution of importance in the development of the court and, later, the ballroom dance. The earliest of these masters were fiddlers and men of considerable learning. They would frequently accompany their young pupils when the latter married into foreign courts. Such travels made the dancing masters instrumental in the spread of dance forms across Europe, from court to court, as well as in regularizing dance and social decorum. They also made social dancing into an entertainment for spectators, eventually transforming it into staged dances.

In 1416, Domenico di Piacenza, dancing master to the ruling Este family of Ferrara in Italy, wrote the first known treatise on social dance. It indicated that a definite technique had been established for upper class social dance. Originated at the courts of Provence in southeastern France, it was penetrating into the Spanish, English, and other European courts. The caroles and the courtesia were gradually giving place to the branles and *basses danses*. These were similar in quality but were often danced to especially written secular tunes that gave them a livelier movement.

From 1500 to 1650. The period in dance from 1500 to 1650 has been called the "age of the galliard," but this lively couple dance, or jig, with its *cinq pas* ("five steps"), or "sink-a-pace" as Shakespeare called it, was by no means the only quick dance to be performed by the aristocracy. By 1500 youthful courtiers had revolted against the solemn caroles and branles. They again sought change later in the century after the pavane, a slow, procession-like dance from Padua, and the Spanish sarabande, requiring elegant pointing of the toes, were introduced at court. It became fashionable to follow these slow steppings with a faster "after dance." The galliard was the first of these. It included the mimed gestures of courtship that had been eliminated from the slower dances, and its more complex hops and capers allowed the male dancers to show off their agility.

With the increased amount of court entertainment, social dance became more varied, particularly in France after the Fêtes of Bayonne in 1565, staged by the former queen, Catherine de Médicis. At this fete, groups of dancers from various provinces introduced a host of folk dances to the court. Among them were the bourrée from Auvergne, the courante from Périgord, the volta from Provence, and the gavotte and the rigaudon from Dauphiné. The bourrée, with its neatly executed steps, later became the most popular court dance, and the courante was a favourite with Louis XIV (1643–1715). The vigorous volta, in which the male lifted his partner off the floor and swung her, shocked the Spanish ambassador to England when he saw it danced by Elizabeth I, who wore skirts shorter than most noble ladies. The gavotte, one of the first dances in which the feet were lifted from the floor, remained popular for three centuries.

New peasant dances

These dances became popular in the courts only after dancing masters evolved techniques that enabled the nobles to perform the rapid and intricate steps with ease and dignity and to display the graceful turnings of their legs or wrists. These subtleties were the marks of ladies and gentlemen aware of the figure they were cutting in society. The book *Orchésographie* (1588) by Thoinot Arbeau (pseudonym of Jehan Tabourot) was an important record of contemporary dances. In it, Arbeau insisted that the dancer pay greater attention to the display of his figure, manner, and style than to his enjoyment of the dance.

From 1650 to 1750. In 1653 the *branle du Poitou* was introduced at the French court under the name "minuet." This dance, of folk origin, became the most popular of aristocratic social dances at European courts until the French Revolution in 1789. It often was revived until the beginning of the 20th century. Characterized by its tiny steps (*pas menus*), it had many different forms, every dancing master altering or adding to the figures made, the exact placing of the feet, the angle of the partners' bodies in relation to each other, and other details. The minuet allowed the couples, who no longer held hands as in the earlier caroles and branles, to woo one another with their glances.

Minuet

From the accession of Louis XIV in 1643 until just before the beginning of the Revolution in 1789, the fashionable dances of the French court dominated aristocratic ballrooms in western Europe. But in England another kind of group dance became a permanent part of popular dance. These English "country dances," derived from traditional folk sources, were arranged by John Playford and published as *The English Dancing-Master*, first in 1651 and in 17 subsequent expanded editions in the following 80 years. Dances for social occasions always had been easily exchanged between court and common people in England. Contemporary accounts of the dances at the Elizabethan court and of the masques by John Milton, Ben Jonson, and others, as well as numerous references in Shakespearean plays, show the country dances and the various jigs taking their place alongside the fashionable French court dances.

Dance democratized

These country dances served the purposes of the middle-class-dominated Commonwealth government in the 1650s. Its puritanical ethic tolerated dancing, provided it was seemly and showed good manners and style. The steps were simple and patterns were intricate and "occupied the mind." The dancers constantly changed partners, giving no occasion for flirtatious behaviour. Everyone stood equal in circular, square, or longways sets, and as many as wished could join in. Moreover, because such dances belonged to tradition and to all classes, a person could step more easily from one rank of society to another without having to be groomed by some foreign dancing master, whose artificialities frequently annoyed Englishmen. These country dances, notably the Scottish and Irish reels and jigs, had already been taken by emigrants to the various colonies, where they continue to be practiced. Many underwent sea changes and came to be known as square dances.

After the Restoration of the English monarchy in 1660, the country dances retained their popularity, both at court and in the lower class dance halls that began to spring up in cities. These dances became popular also on the Continent, where they were introduced at the French court in 1710. They found favour quickly among the aristocracy, and, under the name contredanses, they were enlivened with steps from minuets, bourrées, and courantes. The popularity of these dances was due in large measure to the ease with which they could be learned.

The English public dance halls originated at spas, after the Restoration, as entertainments for the upper and middle class patrons who were "taking the waters." Soon they had spread into the cities and across the Channel. Among these were such dance halls and "pleasure gardens" as the Parthenon, Vauxhall, Ranelagh, Mrs. Cornelys', and Almacks, all in London. Their dances were derived from traditional or court dances, but their popularity lay in the rousing tunes played by military or town bands. The

musicians made up for any lack of finesse by the robust beat and the lilt of their music. By 1789 Paris alone had 684 public ballrooms. They offered one avenue of youthful escape and allowed the court to maintain its decorum.

The dancing masters who ran these commercial ballrooms assumed the old court title master of ceremonies. As such, they introduced couples to one another and directed each dance, often calling out the steps and figures to be performed. But when the French dancing masters elaborated their contredanses and developed the cotillons, their insistence on formal behaviour and technical precision once again made aristocratic youth look elsewhere for entertainment.

From 1750 to 1900. The allemande, a lively, turning couple dance of German origin, had become the most popular dance at open-air balls and fetes by 1760. It required skill, for the man turned his woman under his arm as they progressed around the ballroom. It was soon popular in aristocratic society, for practically alone among couple dances it allowed partners to keep close contact throughout.

The allemande did not retain its ascendancy long, however, for an exciting new dance began to appear in public ballrooms before the French Revolution. Mozart was one of the first to note the growing popularity of the Austrian Ländler, also known as the Deutsche. It was based on a simple, traditional courtship dance in which the man strutted around the woman like a mountain cock at the mating season. To music in $\frac{2}{4}$ time, the couple spun around in each other's arms with a step and hop. Such steps could not be danced on a slippery ballroom floor nor in the dress of the day, so Austrian musicians gradu-

Waltz

ally smoothed out the lively tunes into a swift, lilting $\frac{3}{4}$ measure. Thus the waltz swept into the people's dance halls and was taken up by aristocratic society.

At the Napoleonic court there was an attempt to emulate the grandeur and dignity of the court of Louis XIV by giving pride of place to the more formal figure dances. The quadrille became a favourite in fashionable society of the empire, particularly in the more expensive ballrooms. It provided an excellent opportunity for displays of skill. But once the waltzes of the Strauss family and of other Viennese composers became known throughout western Europe, this swirling dance became queen of the ballroom and remained so into the 20th century.

In 1840 the polka, a lively peasant dance from Bohemia with a bouncing, turning step and strongly rhythmic measure, began to rival the waltz in European ballrooms. It was followed by the even faster and simpler galop in 1845, but neither could surpass the waltz. Goethe's description of the hero of his novel *The Sorrows of Young Werther* (1774) dancing with his beloved suggests the intoxicating quality of that dance:

> And never have I moved so lightly. I was no longer a human being. To hold the adorable creature in one's arms and fly round with her like the wind so that everything around us fades away.

The ballrooms of England were among the last to succumb to the waltz. True to tradition, English dancers preferred the older group dances, and at first the middle classes eagerly adopted the cotillons and quadrilles. These in turn, were replaced by the lancers, a quadrille which contained familiar figures from traditional English dances and Scottish and Irish reels. But by the end of the 19th century, with the number of public ballrooms growing even larger, the waltz had become the dance for all.

The 20th century. Dance forms had always moved between social classes with relative ease. In the 20th century differences in dance styles related to the "who" and "where" of performance were virtually eliminated. So-called pop music from phonograph records or the radio and the examples of dance stars performing on the stage, television, and film became available to nearly everyone. These conditions produced at first a uniformity of dances across society. The Official Board of Ballroom Dancing, an association of dance teachers, was founded in England in 1929 to standardize dance techniques and regulate dance competitions. At the same time, social dance, now rightly called popular dance, increasingly defied regimentation. Dancers, especially the young, found improvised steps and patterns preferable to following prescribed routines. In seeking different forms of expression they embraced new dances, quickly exhausted their interest in them, and discarded them for even newer forms. The popular dance became a phenomenon in rapid flux.

A major innovation was the lead taken by the Americas in providing dances to the world. The main sources were the dances of the American Negro, largely by way of the stage, and of Latin America. The latter dances were based on a blend of indigenous and Iberian music and rhythmic elements brought from Africa. The first major break with the older European forms came in 1910 with the arrival of the tango in North American and European ballrooms. It became extremely popular in middle and upper class European ballrooms when managers of hotels and other gathering places began to stage their "tango teas." At these, the dance was first demonstrated by a professional couple, then danced by the clientele, who enjoyed its new hesitant rhythm and seemingly daring steps. The dance could be mastered fairly easily. It provided an excuse, as had the waltz, for close contact with a single partner.

Latin-American dance

An even more powerful beat began to sound through ballrooms at about the same time. Jazz musicians of the U.S., South America, and the Caribbean gradually brought the sound of their spirituals and "blues," with their syncopated rhythms and improvised playing, into the world of social dance. The original fox-trot was devised by a dancer-comedian, Harry Fox, for the Ziegfeld Follies revue in 1913. The unique syncopation of its music, and the simplicity of its steps, immediately aroused the audience's enthusiasm. Its potentialities were quickly recognized by ballroom-dance teachers, and it spread rapidly before the outbreak of World War I in 1914. The steps were gradually standardized, largely by the famous U.S. dance team of Vernon and Irene Castle, and the fox-trot remained popular for many years.

Throughout World War I the popularity of jazz increased in Europe, especially through the influence of American soldiers. When peace came in 1918, American films, musicals, and revues, as well as dancers and bandsmen, found engagements throughout Europe. The fox-trot, the Boston (a variety of waltz), and the one-step (a fast fox-trot) required little training. But others demanded the services of dancing masters before they became popular or were even countenanced in European ballrooms; *e.g.*, the shimmy of 1923, with its shaking body; the Charleston of 1925, with its wildly flailing legs; and the black bottom and the blues of 1926, the former with its distinctively twisted contortions and the latter with its close embraces.

These dances, too, had originated with the American Negro. They had been given theatrical form without losing their spontaneity. The musicians and dancers used the printed music or the prescribed steps only as the starting point for the improvisation that gave individuality to their performances. Such individuality had no place, however, in English and other European ballrooms. There, decorous behaviour was still expected, and many dancers were still too self-conscious to break into spontaneous movement. The jazz dances were therefore swept aside by new Cuban rhythms, exemplified by a craze for the rumba in 1930. After some sophistication and regularization, the rumba, like the tango before it, was introduced in public ballrooms by exhibition dancers, who then gave the floor to the spectators.

In the early 1930s some dance halls attempted to revive group dancing to counter the decline of the ballroom as a social meeting place. In these dances a number of couples performed a popular dance while following carefully prescribed patterns over the floor, much in the manner of a corps de ballet. But the idea found favour only among expert dancers, leading to the "formation dancing" performed in ballrooms or on the screen by professionals. The man in the street had no wish to be regimented in this way.

The tradition of the dancing master continued, however,

especially in Europe, where he acted as master of ceremonies at dance halls or in hotel ballrooms. Many of the masters developed large and fashionable followings. In the early decades of the century, many dancing studios sprang up in cities in the U.S. and in England, but they made little headway on the Continent. The heads of such studios eventually persuaded local authorities to license and regulate all dancing sessions and thus give their ballrooms a more respectable status in their communities.

The English style

The so-called English style of the Official Board of Ballroom Dancing became ascendant with the first international match, between English and Danish dancers, in 1934. It was recognizable for its elegance, grace, and comparative simplicity. Wherever dancers, teachers, and managers felt that the formal techniques of ballroom dancing were the only things that mattered, it became an accepted standard and was used by trained dancers in the major dance halls of the five continents. National and international competitions were fierce, and the ballroom championship of the world was highly prized. But such dancing, like formation dancing, can no longer be considered popular dance.

Dancers seeking enjoyment rather than the regimentation of the average ballroom took new pleasure in the Lambeth walk, introduced in England in 1937. Originated by the comedian Lupino Lane in his popular musical *Me and My Girl*, it rapidly became popular throughout the British Isles. It was little more than a parade with mimetic passages and friendly, even familiar, gestures that allowed the couple to do more than hold each other sedately. Moreover, the music had all the elements of old London street songs. This friendly familiarity first endeared it to cockney Londoners, then to their mates in the armed forces, and finally to foreigners living in the British Isles during World War II. Three similar cockney dances, "boomps-a-daisy," "under the spreading chestnut tree," and the "hokey-cokey," the last based on a children's game, also became popular.

Just before World War II, in 1939, the jitterbug found its way across the Atlantic. A favourite in the U.S. during the 1930s, it featured improvised spins, twists, lifts, and other athletic interchanges between the youthful partners. It evolved also into the less acrobatic "swing" and "jive" characteristic of the "era of the big bands." True jazz dance, from which jitterbug was derived, was repugnant to European dance-hall managers, because of its unrestrained and improvised movements. Another feature of the 1930s in the U.S. was the dance marathon, in which couples competed for prize money by trying to outlast all others in nonstop dancing. Threats to dance halls that they would lose their licenses kept the phenomenon out of Europe.

Dance in wartime

World War II led to the final breakdown in class distinctions within popular dance. It also weakened the hold that the dancing master and the ballroom manager had exerted on social dancing since the first Provençal troubadours demanded propriety in the performance of caroles. In Germany all social dance was prohibited, as it had been in World War I, but in England dance became a release for persons living under stress, often in cramped underground shelters. A tune blared out of a radio often broke tensions and brought on spontaneous dancing. In such circumstances, dance became again a communal activity.

Since World War II. After 1945, popular dance became increasingly without, and perhaps without need of, aesthetic theory or formal technique. The dances were learned by direct observation of other dancers on the dance floor or on stage, film, or television. As if in revolt against the restrictions of the war as well as of social dancing, youth asserted its own ideas in the dance halls, discothèques, cafés, and wherever else they joined in social activities. They improvised as the music or their spirits moved them. They had danced like that during the war and thus they continued to dance—to the deep-throated Negro singers, to the crooners whispering into their microphones, or to the throbbing beat of electric guitars. Whether it was jive, rock 'n' roll, the twist, the frug, or any other newly fashionable dance, and whether the musicians were some of the dozens of "rock" or "pop" groups or folk singers with their songs of protest, the dancers sought eagerly for new ways to enjoy themselves, to forget their troubles. The flashing stroboscopic lights in many of the teen-age dance halls appeared to freeze their writhing forms in space and time.

DIFFUSION OF EUROPEAN POPULAR DANCE

Influence of colonization and conquest

Beginning with the voyages of exploration and colonization in the 16th and 17th centuries, popular dances found their way from Europe to other parts of the globe. The emigrants from the British Isles to North America, South Africa, and Australia brought their country dances to those areas. These dances often contained elements from other nations as well. In South Africa, for example, the country dances and reels exhibited both the English love of pattern and the Dutch feature of stamping the feet. The do-si-do (dos-à-dos, "back to back") pattern of the North American square dance clearly showed a French influence.

The influence of the Spanish explorers and conquistadors was everywhere evident in Latin America, where the mixture of European, Amerindian, African, and Asian cultures spawned the tangos, sambas, congas, and other popular dances of the 20th century. Mexico and Brazil offer striking examples of the mixtures of peoples and dances. The Mexican Indians were overwhelmed by Spanish, who brought with them not only African but also Asian slaves, from the Philippines and Malaya. With them came the syncopated rhythms that were to make Latin-America dances so popular. In the 1860s came the French armies, led by Archduke Maximilian of Austria. From the early 16th century until the revolution of 1911, Mexico was continually infiltrated by popular tunes and dances from different European countries. First came the Spanish zarabanda and later the fandango. Later, opera became the rage of the Spanish-French upper classes; and the waltzes, mazurkas, and polkas acquired a Spanish flavour and were given fresh colour and new African and Asiatic rhythmic elements.

Brazil offered a similar picture. The case of the samba indicates how the shifting and intermingling of races produced different results. In northern Brazil, *samba* was the place where social dances were—and still are—held. Until recently, these were waltzes, polkas, and fox-trots. In Bahia, in east central Brazil, the *samba* was a communal round dance performed at carnival time with a soloist turning and leaping in the centre of the circle. In the northwest the "samba" was a type of unsyncopated tune in $\frac{2}{4}$ time. In Rio de Janeiro, the popular ballroom *samba* made its debut in 1917 as a reckless couple dance.

Effects of technology

During the 20th century, the phonograph, the motion picture, the radio, and the television set made the latest Western rhythms the property of the world, nearly as familiar in Japan, Africa, India, or the Philippines, as in London and New York. The traveller from the West encountered familiar tunes and saw the same shuffling, shaking, and stepping that he had known at home. He was not, as he might have been a century earlier, an outsider to the fashionable local dances. The transistor radio became a status symbol for many persons in more primitive areas of the globe. Its sounds, together with other elements of Western civilization, weakened the hold of indigenous dances, which as the 20th century progressed, were in danger of becoming museum pieces in many areas.

BIBLIOGRAPHY. *The Dancing Times* (monthly), and *The Ballroom Dancing Times* (monthly), contain articles on all types of social dances at all periods, as well as articles by expert teachers on all past and present forms of social dance; *Journal of the International Folk Music Council* (annual), contains articles on the relation between folk and popular dance, the development of contemporary popular dance, and the relation of popular dance to its music; P.J.S. RICHARDSON, *A History of Ballroom Dancing, 1910–45* (1946); and CURT SACHS, *Eine Weltgeschichte des Tanzes* (1933; Eng. trans., *A World History of Dance*, 1938, reprinted 1963), both give comprehensive surveys of all types of social dance.

(Jo. L.)

Popular Literature

The term popular literature has several connotations: it may mean literature intended for the masses, or literature that finds favour with large numbers of people, or it may be a literature of the people. The types may overlap, or they may remain distinct. Thus, "popular literature" can describe ballads, verse, tales realistic and picaresque, romances and confessions, jest books and chapbooks, westerns and thrillers, science fiction and fantasy literature, allegory and satire, emblem books, strip cartoons, comics, and even picture postcards; it may be applied to pamphlets and certain kinds of journalism; and it encompasses a whole range of dramatic literature, from monologue and comic turn to full-length play.

Many of the terms used in discussing popular literature, such as people, folk, and culture, carry widely differing and even conflicting implications. A broad discussion of such concepts as they relate to all the arts will be found in the article PRIMITIVE, FOLK, AND PRIMITIVE ARTS. This article will treat the subject of popular literature primarily as it relates to other forms of literature and only in general terms, since any attempt to present a comprehensive detailing of its outcroppings would prove endless. An analogous article is FOLK LITERATURE. Many of the literary forms most commonly associated with popular literature, such as detective stories, gothic romances, science fiction, and "best sellers" are discussed at some length in the article NOVEL; and there is also some discussion of other genres of popular literature in the articles SHORT STORY; and LITERATURE, NONFICTIONAL PROSE. Discussion of what is popular in other arts may be found in the articles POPULAR THEATRE and POPULAR MUSIC.

Articles on the history of literature, such as LITERATURE, WESTERN; and LITERATURE, EAST ASIAN; and articles on the arts of various peoples, such as SOUTH ASIAN PEOPLES, ARTS OF; and AFRICAN PEOPLES, ARTS OF, also discuss popular literature.

THE DEVELOPMENT OF POPULAR LITERATURE

Modest in achievement though so much of it is, popular literature can boast that the origins of journalism and the novel are to be found in it. A vast number of tales and romances of a popular nature were published in Great Britain before 1740, the date Richardson's *Pamela* appeared, and the date used by many critics to mark the beginning of the English novel.

Before some of the attempts to define popular literature are described, it might be useful to examine examples from one particular cultural context—that of English literature—which have proved to be popular or which have been said by scholars with different interests to be expressions of popular culture. The examples and the evaluations are not comprehensive and are not intended to be typical. They may, however, illustrate how differently understood is the expression popular literature and warn against taking too uncritically the generalizations about it.

With the advent of printing, literary popularity of a new magnitude became possible. Early evidence of the extent of such popularity is usually incomplete: the number of copies printed of 16th-century books is not known and often whole editions are lost. Thomas Deloney's tale, *Jacke of Newberie*, was evidently published in some ten or a dozen editions between 1597 and 1637, but no copies of the earlier editions have survived. Deloney's popularity extended beyond his own age, for there were at least another half-dozen editions of *Jacke of Newberie* in the later 17th century, and his *The Gentle Craft* was even more popular, if the number of surviving editions is an indication of the number printed. Furthermore, it was turned into a play by Thomas Dekker a year or two after it first appeared and it survived in chapbook form in the 18th century.

Sermons, chapbooks, and broadsides

Even more popular in Shakespeare's time were the sermon and such moral tales as Richard Bernard's *The Isle of Man* (of which 16 editions are recorded up to 1683). The appetite for sermons, at least in towns, seems to have been insatiable. There are at least 128 known editions of the works of Henry, or "silver-tongued," Smith in various combinations, published between 1589 and 1637; and 128 editions of the works of the Puritan preacher William Perkins (1558–1602) are recorded before 1640. The equivalent editions of Shakespeare for this period number about 90, Marlowe 31, and Spenser 19. "Popular" may justly be associated with Smith and Perkins in ways other than that assessed by counting editions. Sermons were more than instruments of the salvation of mankind; they were the chief source of political information and political ideas for most Englishmen, and their "popular" significance in prerevolutionary England is considerable.

Another source for the history of popular literature is the chapbook, a wood-block illustrated pamphlet of tales, tracts, and rhymes circulated by peddlers in the 18th century. A rival to the chapbook would be the broadside ballad. The variety of subjects of broadsides was considerable. Subjects might be topical, political, or loyal; they might deal with calamities—shipwrecks, fires, or even the collapse of a building; but those that sold in largest numbers in the early 19th century (as many as 2,500,000 copies of a single broadside are reported) were those issued by James Catnach and John Pitts that reported trials and executions.

Although figures for the sale of broadsides in the early 19th century—and thereafter those for the novels of Dickens and Mrs. Henry Wood, and, outstandingly, Harriet Beecher Stowe's *Uncle Tom's Cabin*—are impressive evidence of this kind of popularity, the 19th century is even more significant in terms of popular literature for the birth of popular journalism. Though the origins of the popular press go back to the pamphleteers of the 16th century, it was the success of *Tit-Bits*, launched in 1881, and the *Daily Mail*, in 1896, that ushered in what has come to be called the "post Northcliffe" era of mass culture, after one of the most influential newspaper publishers, Viscount Northcliffe (1865–1922). Television now plays an even more significant part in popular journalism than the newspaper.

Popular journalism

This highly selective—and almost arbitrary—list of what has proved popular in terms of sales, or of what has been described by scholars as being of the people, or influenced by them, over the centuries, will indicate how varied is the range of literature which can be described as popular. It will be apparent how loosely the term is used and how broad, and unsure, generalizations about popular literature are likely to be.

FUNCTIONS AND EFFECTS

Popular literature can be approached theoretically in two sharply distinct ways. It can be thought of as deriving *from* the people, or it can be thought of as directed *at* them—with greater or lesser degrees of conscious calculation on the part of its originators. Those who attempt to define the first approach are likely to insist on the importance of the oral tradition. This tradition does not exclude the work of known authors, but it does require that it be refashioned by the community in order to make it a part of the culture of that community.

In discussions of this kind of popular literature, class and community distinctions are important for some scholars. Folk literature may be seen to be associated with agricultural communities or it may be rather more broadly defined as being associated with all the lower classes.

Popular literature, it has been suggested, is designed to entertain; it is immediate and demands spontaneous response. High, or elite literature, in contrast, is said to be formal and composed, an art for contemplation and *intended* to endure. It has also been argued that high literature may be judged by its ability to disturb and challenge, whereas the popular arts confirm known experience and values, seeking to reassure. This is an extremely useful distinction, but it attracts two comments.

Contrast with elite literature

Though there is no doubt that much contemporary critical evaluation regards highly the capacity to disturb and to challenge, this is nevertheless a limited approach. It fails, for example, to assign much merit to the capacity to delight (and to create delight without being trivial is,

perhaps, more difficult to achieve—and much more difficult to explain—than to challenge or disturb), and it disregards certain aspects of the expression of individuality and independence through language. Popular literature *can* give delight without being trivial and in its use of language, in, say, the absurd and the bad pun, its cultural contribution may be significant and lasting, awkward though it may be to assess in terms of high art.

Second, though it may do so infrequently, popular literature can disturb and challenge. This can be experienced in some of the earliest English dramatic literature that has roots in the community (as in the Crucifixion plays, which are a remarkable mixture of pathos and farce), in some ballads and music-hall songs and acts, and even in film and television comedy. When popular art is accepted by critics of high art, however, it is elevated hierarchically—it becomes work attributed, say, to the "Wakefield Master" or is lifted above the mass, as was the English music-hall art of Dan Leno (1860–1904) by Max Beerbohm, or the works of such detective story writers as Dashiell Hammett in America or Georges Simenon in France were by a variety of critics.

In Germany in the 1930s, then in America, and later in Britain, attention has been paid to the way in which popular literature has been affected—some would say corrupted—by mass media: newspapers, films, radio, and television. Many people express concern at the situation in which nations of the Western world find themselves in this respect. Alongside the fear of the corrupting effect of the mass media is an awareness of the progressive alienation of elite art from popular art. Yet the situation is not, perhaps, quite so desperate, nor so completely distinct as it may seem to modern man, confronted as he has been by three new media—films, radio, and television—in the space of half a century. Furthermore, the alienation of popular and high art forms is not uniform, and in some areas, this trend is being reversed.

Changing contexts of popularity. Elizabethan drama has been called a "popular form" not so much because it was a drama of the streets but because the plays had a *substructure* of the popular in works of high seriousness and in their use of popular attitudes and forms (such as revenge tragedy). It should be noted that "popular" is being used here in both of its principal meanings: popular *of* the people and popular *with* the people.

There are undoubtedly close links between popular dramatic forms and the plays of Shakespeare and many of his contemporaries. That these plays were popular in that they reflected attitudes popularly held, and that they enjoyed wide esteem is well established. But it is also clear that such plays are no longer "popular" either *of* the people or *with* the people. Full houses for Shakespeare in seasons at Stratford-on-Avon are not a true indication of either kind of popularity, and his contemporaries never fare as well.

Although their work has not fallen into the virtual oblivion suffered by the sermons of Smith or Perkins, it has nevertheless become representative of high art. Furthermore, the popular elements of the writing of these dramatists have to be searched out by scholars.

Thus, what was popular (in both major senses) in those dramatists has become, for today, something different, and much else that was popular—whether sermons or pamphlets or plays—can hardly be said to excite the attention of anyone but the specialist scholar. What was popular in the Elizabethan period, a modern scholar's understanding of what then was popular, and the response of a modern popular audience to such works are by no means coincidental.

The "threat" of popular literature. Those who are most sensitive to the high cultural tradition tend to see a threat to it from the popular arts of their own time, especially from the popular arts involving mass media: mass journalism, film, radio, and television. It is difficult to understand the effects of the new media while fully exposed to their innovations. Objective evaluations are complicated by inherited traditions of high culture, fears for their survival, the difficulty of detecting popular traditions in the new media, and perhaps, too, elements of Puritanism. Above all, it is impossible to predict how the human intellect and spirit will respond to these new media. Grounds for optimism may be found in the human response to the advent of printing, a challenge in its time no less remarkable than those occurring in the 20th century.

In the study of popular literature, the present is felt much more sharply than even the recent past. False comparisons with this past can easily result from this sensitive awareness of the present. Popular literature presents a much more difficult problem than any form of high art, which is, by its very nature, selective. In popular literature, some estimate of the whole, the mass, must be made. This presents physical difficulties and there must then arise the inadequacy and inaccuracy of data.

FORMS AND TECHNIQUES

It may seem to a modern scholar, studying the new media of the 20th century, that there has been a sudden shift from popular art to mass art and that this has led to depersonalization and the destruction of style. Mass art, in this use of the term, is not simply the art conveyed to a large audience through one of the mass communications media, but rather an art that is distinct from popular art in its relation to its audience. Popular art is seen as retaining the vital link between artist and audience that folk culture enjoys but mass culture relinquishes. Both appeal to large audiences, but the popular writer respects both his art and his audience, whereas the maker of mass art is ruled by his audience, which he is anxious to please—no matter how. The popular writer knows his audience as he knows himself. Although he may use a mass medium, he still expresses something essential to his own nature, and from the depth of this nature he brings up memories and feelings to which his audience, being of the same background as he, can respond.

Mass and popular art

That much of the literature fabricated for the mass media of today is depersonalized and lacks style is undeniable: but so, of course, was much of the literature created for large audiences of preceding generations (see, for example, the description of serial writing in Charles Manby Smith's *The Working Man's Way in the World*, 1853). Again, it is impossible to conclude with certainty that the proportion of mediocre writers to genuine artists is higher in the mass art of today than in the popular arts of the 19th century.

The romance probably has the longest and fullest tradition in popular fiction, extending from the early Middle Ages to the present. It is found in verse, prose, and dramatic forms. In popular literature, though it shows countless differences, its mainspring is some kind of forbidden love. The censure may apply in the world outside but not in the fictional world (this is notably so in pornographic literature, which frequently depicts societies freed from civil, and sometimes physical, restrictions and limitations). The censure also can be represented in the story, directly or indirectly; it may be legal or religious, or very frequently social (a class barrier, for example). A characteristic of such romance is that even when it appears to be realistic, it presents a world from which certain normal restrictions are removed.

Another principal kind of fiction is the realistic tale. Though the terms suggest a general difference between the worlds of romance and realism, it should not be assumed that realism necessarily reproduces the actualities of real existence, however carefully verisimilitude may be suggested. It is not actuality with which writers of popular realistic literature are chiefly concerned, still less with "real" issues or "real" relationships. Their aim (though there are exceptions), like that of the writers of popular romance, is to provide their readers with wish fulfillments, and in this respect it can be hard to differentiate between the two kinds.

Of the many other forms favoured by popular literature, fantasy is one of the most significant. Fantasy literature can be conceived of as an extension of the romance except that love is not the object of the inhibition. Utopian literature and science fiction are the two most obvious forms of fantasy writing and the latter is particularly intriguing in this context for two reasons. In the first

place, though its worlds may not be our known world, it is often at pains to be "realistic" within its own fantasy—for example, the careful, as well as imaginative, use of technological paraphernalia by the pioneer French science fiction writer Jules Verne. Secondly, while being in most of its manifestations a popular literary form normally disregarded by critics of high literature, it makes use of archetypal myths and gives expression to deep-seated urges and fears, which relate it not only to the popular subconscious but also to high art. The Soviet novelist Yevgeny Zamyatin, for example, using the science-fiction form in his novel *We* (1924), expressed fears of totalitarianism so effectively that he was forced to leave the U.S.S.R.

Values espoused. The prime aim of the writer of popular literature is immediate entertainment, and, in this, the tradition is as limited and as worthy as is the ancient craft of which it is a part: storytelling. The most serious criticism of popular literature (indeed, of all literature) concerned primarily with immediate effect—with wish-fulfillment—is not that its "literary standard" may be low, or its ideas stereotyped, but that its representation of human relationships and the values it espouses are false. The more seemingly in accord with verisimilitude it is, the more will its readers be encouraged to make unreal demands on life and the more frustrated may they become if their expectations are disappointed. Though this would seem to be true of most, if not all, products of formula literature, it ought to be borne in mind that most literature is mediocre or worse and that these criticisms can also be levelled at literature that is intended to be sophisticated, profound, complex, and enduring.

Attitudes toward love and sex

Nowhere is the difference between popular and high literature more striking in the latter half of the 20th century than in fiction dealing with relations between the sexes. Avant-garde literature and the fiction that gains critical recognition from the serious literary journals reflect an attitude to sexual relationships generally in accord with what has come to be termed the permissive society: frank, open, uninhibited. Yet at the same time there is a steady output of the novel and short story devoted to romantic love. Here, the standard set—the code of behaviour—not only is different at all points from that espoused in high literature, and *seemingly* acceptable in daily life itself, but has at no time come very near to actuality. In this genre (if two manifestations so utterly different can be comprehended in a genre) there is virtually no interaction between high and low, unless it be for the occasional parody of the popular by the high.

Qualities beyond literary analysis. On close verbal analysis of what is said and the way it is said, almost all popular literature will seem inferior to high literature. In its manipulation of words, in sophistication and complexity, in the use of irony, and so on, high literature will respond more readily and more satisfyingly to analytical techniques. It is apparent, however, that some popular literature, which is unresponsive to analysis of this kind, is not without artistic and human virtues. This is true of some folk poetry, of some ballads, and, of some of the tradition of nonsense and the absurd in music-hall entertainment. This last tradition fostered imagination and verbal delight when high literary art was concerned with realism and naturalism, and its significance is far greater than what is judged to be its literary merit. In English literature, especially popular literature, the irrational has played an important role: its expression of nonsense is tantamount on occasion to a poetical faculty of a kind quite distinct from that found on the Continent. This quality is also unresponsive to the kind of literary criticism appropriate to high literature.

Criticism and the bad pun

In its inability to deal with the bad pun, the inadequacy of such criticism can be seen to be particularly acute. The frequent use of the bad pun by the greatest of all English writers, Shakespeare, makes it difficult to dismiss out of hand. The delight afforded to those of all classes by the bad pun cannot be explained in high literary terms. It may be that this play on words, and the distortion of language that is frequently involved, is a means whereby individuality is expressed, due order and the rational are subverted, and personal independence asserted, while at the same time—in the response it evokes—a sense of community is shared. Though the bad pun cannot be accounted for in the terms of high art, its cultural importance (in the anthropological sense), and its popularity—transcending class distinctions—are of great significance.

The paradox of the bad pun epitomizes some of the peculiarities and problems posed by popular literature. Thus, although by high-literary standards the language of popular literature is often weak, analysis appropriate to high literature may not reveal its essential qualities, and what is different is not necessarily inferior.

INFLUENCE ON FORMAL LITERATURE

As significant as the differences between the characteristics of popular and high literary forms are the ways in which high forms develop from popular beginnings. The influence of the popular on the high is familiar enough in music but less obvious in literature.

Popular into high literature

One of the most interesting characteristics of certain popular genres of the 20th century has been the way in which they have shown signs of developing higher literary forms. Western novels, for example, have been largely third- and fourth-rate, but the works of such writers as Walter van Tilburg Clark and A.B. Guthrie are something more than "popular entertainment." More subtle has been the indirect effect on dramatists, such as Arthur Miller in America, and the Australian Ray Lawler, of that cliché of the Western, the isolation and independence of the frontier hero; both writers have used it, switching their locales not merely superficially to the city, but to the frontiers of national attitudes to life.

Without question, the clearest, though not necessarily the most significant, evidence for the growth of a high art form out of the popular is to be found in science fiction. The origins of science fiction, if the term is interpreted broadly, go back many centuries and the names of men of high ability have been associated with the literature of fantastic voyages and Utopian worlds. Nevertheless, it can be shown that in the form in which science fiction has proved so immensely successful in the 20th century, its sources are truly popular—of the people. Modern science fiction has sprung from the passionate enthusiasm of its fans, who exerted strong influence on the editors of science-fiction periodicals to ensure that the literature developed as its readers—its participants, indeed—wished. The way in which ordinary readers in this way played a part in the development of science fiction suggests that the concept of popular literature as being no more than a shadowy imitation of high art is not inevitably true.

Although the literature of science fiction may seem at first homogeneous, it is not only diversified but quite strikingly reflects the cultures supporting it. Thus, American science fiction is much concerned with what has been called the breaking of a culture (seeking new worlds in which man can develop further owing to the decline foreseen in Earth's civilization), while British science fiction is dominated by impending disaster.

The influence of popular art on high art can be seen most clearly in dramatic literature. It is a commonplace that popular drama has burlesqued the greatest art (Shakespeare, for example, by innumerable television comedy writers), and a little less obvious that popular elements have been used by serious creative writers (John Osborne's use of the music hall in *The Entertainer*). But the influence of popular drama on legitimate drama goes much deeper and further.

An examination of two of the most persistent dramatic traditions—the comic monologue and the two-man comic turn—tell much about this influence. The two occur in the earliest English drama, they are much in evidence in the plays of Shakespeare and his contemporaries, and they are to be found in the popular tradition up to the present. What distinguishes these forms from drama, in which disbelief must be suspended, is their continuous breaking of the continuity of the audience's involvement. In drama in which disbelief must be suspended, an actor who forgets his lines embarrasses his audience and the

integrity of the play-world is broken. In popular drama, the performer can use such an error as a source of comedy; furthermore, the effect is not to break the dramatic world but, paradoxically, to involve the audience in the performer's world more fully and more willingly. There are, too, differences between the relationship of the actor to his role in the two kinds of drama. The actor is involved in a legitimate role: he takes a part. In popular drama, the performer stands apart, retaining his own personality.

Many forces combined to separate the high and popular dramatic forms. In post-World War II English drama, however, they were reunited in the work of such playwrights as Harold Pinter and Samuel Beckett. What is so fascinating about this new relationship is that it makes intellectual use of popular dramatic material and forms as well as the popular dramatic relationship of performer to audience.

BIBLIOGRAPHY. The number of books and articles on popular literature and the mass arts is vast. The selection below will serve only as an introduction. Most of the books listed contain bibliographies and these should also be consulted. Volume 4 of the *New Cambridge Bibliography of English Literature* (forthcoming), contains a fairly lengthy selection of titles on the cultural debate, mass communications, reading and readership, journalism and the press, popular genres, content analysis, and the relations of literature to film, radio, and television, for the period 1900–50. Many of the studies included are very recent and some have implications for earlier centuries. HELMUT BONHEIM (ed.), *The English Novel Before Richardson* (1971), is a useful checklist of texts and criticism.

The best introduction to the common reader is R.D. ALTICK, *The English Common Reader* (1957), containing an extensive bibliography. Q.D. LEAVIS, *Fiction and the Reading Public* (1932), should also be consulted. RAYMOND WILLIAMS, *Culture and Society 1780–1950* (1958), and his *Long Revolution* (1961), are illuminating discussions of the interrelation of artistic and social developments. Undoubtedly the most stimulating and most moving account of popular culture is RICHARD HOGGART, *The Uses of Literacy* (1957). CHRISTOPHER PULLING, *They Were Singing* (1952), not only gives a good account of the music hall, but relates it in a significant way to the social life of the 19th and early 20th centuries. *The Popular Arts* (1964), by STUART HALL and PADDY WHANNELL is a pioneering critical study of the mass media; it is significant not only for what it says but for its attempt to stimulate further discussion by means of a series of projects intended for use in schools and adult education centres. S.L. BETHELL, *Shakespeare and the Popular Dramatic Tradition* (1944), is essential reading for an understanding of the relationship of popular and high dramatic art forms. It served as the starting point for P. DAVISON, "Contemporary Drama and Popular Dramatic Forms," in *Aspects of Drama* (1965). GEORGE ORWELL, *Collected Essays,* 4 vol. (1968), brings together much that has influenced later writers on popular literature in England; P. DAVISON, *Songs of the British Music Hall* (1971), gives words and music of 50 songs, with individual and general critical and social comment.

Five collections of readings are of special interest: B. ROSENBERG and D.M. WHITE (eds.), *Mass Culture: The Popular Arts in America* (1957); E. LARRABEE and R. MEYERSOHN (eds.), *Mass Leisure* (1958); D. THOMPSON (ed.), *Discrimination and Popular Culture* (1964); N.F. CANTOR and M.S. WERTHMAN (eds.), *The History of Popular Culture* (1968); and JEREMY TUNSTALL (ed.), *Media Sociology* (1970).

Finally reference should be made to those centres specializing in the popular arts which issue communications from time to time: The Centre for Contemporary Cultural Studies, University of Birmingham; the Centre for Mass Communication Research, University of Leicester; L'institut de littérature et de technique artistiques de masse, Bordeaux; and the Popular Culture Association (which publishes the *Journal of Popular Culture*), at Bowling Green University, Ohio.

(P.H.D.)

Popular Music

Popular music is music inhabiting the broad domain that lies between folk music and fine art music. It is a domain with no clearly defined frontiers, and at one extreme it merges into folk music; at the other, into art music. But for the most part, the distinction between folk, popular, and fine art music is clear enough.

In general, popular music—unlike folk music proper—is produced by professionals and mainly in the towns. At the same time—unlike fine art music—it is in many cases diffused by oral means. In some important examples, such as Spanish flamenco, central European gypsy music, and a great deal of jazz, it flourished without the aid of print. These three considerations, professionalism, urbanism, and the relative degree of its oral transmission, have for centuries affected the nature of popular music in many parts of eastern and western Europe, North and South America, and much of Asia. After the mid-20th century, with the global expansion and dominance of Western music from Los Angeles to Laos, from Prague to Peking, a worldwide uniformity of popular music began to be apparent, derived in large part from U.S. models.

EARLY HISTORY

Antiquity to the Middle Ages. In antiquity another layer of music was interposed between the music of the peasantry and that of the upper classes. Cities such as Rome and Alexandria attracted Greek minstrels, Syrian dancers, Negro musicians, singers from the Italian countryside, and former slaves or smallholders who made a living by providing music for the urban crowds at circuses or in theatres, in processions or dances. The phenomenon of the popular "hit" song, spreading from the theatre into the streets, was already familiar in Roman times. The Roman writer Suetonius reports that when Galba was made emperor in AD 68, the entire audience in a theatre roared out a song in vogue at that moment, satirizing Galba's avarice.

When, in 6th-century Europe, the by then degenerate theatre of antiquity finally disappeared during the barbarian invasions, dispossessed professional musicians were driven afoot to mingle with the bards of various peoples —Teutonic, Celtic, Slav—in that miscellaneous body of entertainers who haunted the towns and thoroughfares of the Middle Ages. Before long, a general European popular music idiom began to show itself, in which the national differences between music of Spanish, Italian, German, and Polish origin, for example, were to some extent masked by a stylization that resulted from the cosmopolitanism of the professional musicians' calling—but to some extent only. It is true that the popular performances of medieval Europe were far more "international" in character than those of folk music, which often varied enormously in style from country to country and even from province to province. Yet there were characteristic differences between the analogous melody forms and utterances of such varieties of poet-musician as the northern French trouvères, Provençal troubadours, German minnesingers, Spanish juglares, and English minstrels. A 14th-century proverb of French origin indicates the manner of popular singing among various nations: "The French sing, the English carol, the Spaniards wail, the Germans howl, the Italians bleat." In the high and late Middle Ages, popular music covered a wide social and artistic range. Minstrels might be of the court or the tavern, they might be large landowners or beggars, wandering scholars or illiterates, virtuosos or strummers. The repertory ranged from aristocratic hero ballads through lyrical songs, courtly or crude, to rustic dance tunes. Musicians' earnings varied as widely as their grades, and the wealthy troubadours were as scornful of the ragged jongleurs (French minstrels) as minnesingers and meistersingers were of the itinerant *Spielleute* of the German countryside and back streets. The fortunes of popular musicians in the Middle Ages were subject to the same ups and downs as those experienced by the star performers today. Thus the talented jongleur Bernart de Ventadour (flourished late 12th century), a baker's son, attained the courtly rank of troubadour and entered the aristocracy; on the other hand, his contemporary, the equally gifted Raimbaut de Vaqueyras, a dispossessed nobleman, sank down into the ranks of the juglares, playing and singing on fairgrounds and in market squares.

National differences

International traffic in popular music, exemplified in modern times in the worldwide spread of jazz and rock 'n' roll, was already foreshadowed in the Middle Ages by

itinerant professional musicians who might wander anywhere within a territory bounded by Edinburgh (Scotland); Santiago de Compostela (Spain); Cyprus; and Tallinn (Estonia) on the Baltic. Paris-based minstrels would operate in France between Roncesvalles and Metz; German musicians based in Nürnberg might be in Cracow (Poland) for the summer and in Verona (Italy) for the winter. French, Italian, Spanish, and Portuguese minstrels were frequent guests in England. In the towns on the Adriatic coast, Serbian, Greek, and Croatian musicians played alongside instrumentalists from the German principalities and from Transylvania (Romania). French popular musicians, settled in the Hungarian wine-growing districts, accompanied the annual Tokay caravans through Slovakia and across the Carpathians into central Poland. All these musicians exchanged melodies, musical ideas, and instrumental techniques.

Near Eastern influence

Nor was the influence of the Near East unimportant. Urban popular music of Arabic-Persian origin was early spread over parts of southwest Europe by the Moors in Spain, southern France, and southern Italy. From this culture came the short-necked lute with bent-back pegbox, the instrument now popularly associated with medieval minstrels. Later, with the Turkish invasions and occupations, music of similar provenance was spread among occupational musicians in the towns of southeast Europe, notably Bulgaria, Romania, Serbia, Bosnia, and Albania. It was carried, mainly by gypsy professional entertainers, into Central Europe, where it mingled with the local musical stock to create the peculiar, hybrid, semi-Oriental kind of popular music of which, in the 18th century, Mozart availed himself in his *alla turca* (Turkish-style) compositions.

The Renaissance. As feudalism merged into early capitalism (in western Europe in the 15th and 16th centuries, later elsewhere), popular musicians became more settled, and generally only those in the lowest reaches of the profession remained itinerant. An active traffic between town and country culture became the normal condition of popular song. Some of the agents of this process were amateur composers among the rural middle class, making formalized arrangements of folk tunes; domestic servants bringing village music into towns and taking town music back into the villages with them; fairground and marketplace singers; street showmen; booth-theatre actors; peddlers of cheap songbooks and broadsides (from the 16th century onward).

Various institutions helped in the formation of popular-music idioms. In England, from Tudor times, the municipalities employed "town waits," instrumentalists who combined the functions of perambulating night watchmen and town bandsmen. They played for official visits, for the mayor on market days, and for summer-evening concerts from the guildhall roof or a roadside scaffolding. In Coventry, for instance, they also played softly at various corners of the city five days a week between midnight and 4 AM, for the reassurance of citizens. Their repertory consisted mainly of adapted folk melodies, humble amateur compositions, popular dance tunes, theatre music, and marches. In France, from the time of François Villon and Rabelais onward—*i.e.*, from the early 15th to the mid-16th century—popular city cabarets were putting into circulation a large number of songs, sentimental, satirical, or comic, most of which enjoyed only a brief vogue, though some proved lasting.

An even more powerful source of popular music was the secular theatre that established itself firmly in France, Spain, and Italy during the 15th and 16th centuries. The theatres performed short plays, often four pieces in an evening, acted usually on open platforms of a movable kind by companies consisting partly of professionals and partly of amateurs, including women and at times children. Many of these plays began and ended with a song, in the manner of an overture-finale frame, and most of them were punctuated with chansons, villancicos, and villanelle during the course of the action. Some of the actors played instruments and at times would perform pieces from the regular minstrel repertory, in the absence of music especially written for the theatre.

In Germany, where, following the Reformation, printed music, the popular choral movement, and the activities of educated amateur musicians, parsons, and schoolmasters worked with especially powerful effect on lower class music making, a vast amount of folk song was transformed into popular song by the imposition of the conventions and usages of art music. This was accomplished by simplifying rhythms, converting old modes into modern scales, tidying up the structure of melodies, and encouraging a standard kind of voice production. A characteristic institution of the early capitalist period in Germany was that of the meistersingers—mostly small merchants, artisans, and tradesmen—who assembled, usually in the guildhall on Sunday afternoons, to practice their new often homemade songs that belonged firmly to the realm of popular music, being neither folk song nor fine-art compositions.

The Baroque. In 17th-century England, the powerful expansion of bourgeois ways of life and manners of thought was accompanied by a flourishing of urban popular music that became equally as important as folk music and art music. Significant landmarks were the extremely successful publication of John Playford's *The English Dancing-Master*, a collection of contredanse tunes, many of them based on popular song and ballad airs, that ran into 17 editions between 1650 and 1728, and Henry Playford's *Wit and Mirth. An Antidote against Melancholy*, published in 1682, gained great fame when it was re-edited in 1699 with the subtitle of *Pills to Purge Melancholy*. Rather unjustly, Thomas D'Urfey (*c.* 1653–1723) gets the most credit for this collection of popular songs, serious or saucy, of aristocratic, bourgeois, or lower class, origin. New volumes were added over the years and reprintings constantly ordered. By 1720, the work had grown to six volumes and contained more than 1,000 songs, giving a wide view of the popular music of the time in parlour and kitchen, tavern, and pleasure garden.

Influence of British pleasure gardens

The pleasure gardens that sprang up, notably during the 18th century, in the cities and spas, or watering-places, gave an important stimulus to popular music through their evening entertainments. Thousands of songs were written specifically for performance in these places and, at least from the middle of the century, were profusely printed on leaflets or in books. In London, between 1769 and the early years of the 19th century, at least one, sometimes two or three, books of Vauxhall Gardens songs were published annually, as were songs from Marylebone Gardens. One composer, James Hook, wrote more than 2,000 songs mainly for the "gardens crowd," and another, Charles Dibdin, wrote nearly 1,000; several other composers were scarcely less diligent. A comparable flood of lyrical popular music flowed at a slightly later time from the beer-garden entertainments of the German and Austrian towns.

Meanwhile, in most of the larger towns of 18th-century continental Europe, popular comedies with music could be seen in the puppet and marionette tents and on the fairground or back-street stages, often with stereotyped figures derived from the Italian *commedia dell'arte*, such as Pantaloon, Harlequin, and Pulcinella. A report from 1711 tells that at the fairs of Saint-Germain and Saint-Laurent in Paris (held on the respective saints' days), the theatre audiences habitually joined with the actors in singing the favourite songs of the day; the actors had only to sing the opening bars of the song, and the audience knew the rest.

Ballad opera

Later in the century, when the French royal opera forbade actors to sing in the popular theatres on grounds of its exclusive privilege for such performance, the actors would cause a scroll to unfurl each time they came to a song, the instrumentalists would introduce the melody, and the entire audience would roar out the song, reading the words from the scroll. Usually, in this kind of performance, the text was more important than the music, which more often than not was derived or adapted from well-known tunes, similar to the ballad opera. Probably the most important of the ballad operas and one that had great effect internationally was *The Beggar's Opera*, by John Gay, an aggressively satirical piece and a

huge favourite with the populace. It was first staged in 1728 and was an immediate success; within a few years it had been played on countless stages in England, Ireland, and Scotland. In 1733 it was performed in Jamaica and in 1750, in Paris and New York. Its triumph encouraged others. By the end of 1729 London had seen 15 ballad operas and by 1733, 70. Some 700 popular tunes were used in these works, with librettos sometimes written by hack writers, sometimes by such eminent literary men as Jonathan Swift, Henry Fielding, and Richard Sheridan. In later years, this form of light Baroque lyrical theatre was to have considerable influence on continental popular music, notably in Germany.

The French Revolution. The years of the French Revolution comprise a singular period in the history of popular music. Admired singers would set up their little platforms in the squares or near the bridges of Paris to amuse the public by singing topical songs. In many parts of the city the audiences were so large that coach-and-wagon traffic was seriously hindered, and the municipal authorities were at length obliged to provide special vantage points where the singers might perform. During the height of revolutionary enthusiasm there was hardly a gathering of any kind where patriotic, satirical, or general popular song was not heard. In his compendious work, *Les Hymns et chansons de la Révolution* (1904), the musicographer of the French Revolution, Constant Pierre, reports:

Political songs

> For a certain period the theatres were literally invaded and treated as an annexe of the public highway; the audiences displayed their opinions by themselves singing songs of political tendency or by demanding performances, there and then, of their preferred ditties.

Singing invaded the National Convention itself, with one deputation after another announcing its revolutionary fervour and then singing its favourite song to the accompaniment of drums and other instruments brought along for the purpose. Deputies had the habit of standing up in the Convention and intoning lengthy ballads written on some topical point or other until, at last, in March 1794, Georges Danton, a prominent revolutionary leader, proposed a resolution "That henceforth this platform hears nothing but reason in prose." The motion was carried, and from then on singers were rarely heard in the Convention hall, unless by invitation. The course of the torrent of revolutionary popular songs is not without interest. In 1789 a mere 100 songs on political or social themes were published. In the enthusiasms of 1793 the number rose to 590, and the next year, 701 such songs were issued. With the reaction that set in after the execution of the revolutionary leader Robespierre, the numbers declined steeply. Only 137 new songs on topical themes were issued in 1795; and by the end of the century the flow of directly political songs had to all intents ceased. This example serves to remind that popular music is far more sensitive to the important moments of history than either fine-art music or folk music, at least in the Western tradition. Thus, the American Revolution and the Civil War both produced a flood of popular songs but only a few occasional pieces for the concert hall, while—contrary to general belief—the reflection of these events in folk song proper was minimal.

19TH AND 20TH CENTURIES

The 19th century. The period of industrial capitalism saw the greatest spread of specifically popular music. The growth of large towns meant bigger audiences, which, in turn, gave rise to an entertainment industry of vast scope; at first in Europe, later in the Americas, and then in Africa and Asia. Conditions favouring an oral culture weakened, and the immense mass of working people evolved away from self-made folk song toward the products of music hall, dance hall, and popular theatre.

Particularly in the early part of the 19th century in western and central Europe and the Americas, popular social song developed enormously in all its forms—sentimental, patriotic, comic, satirical. The remarkable evidence of the democratization on musical life, after the period of the French Revolution and with the growth of the industrial proletariat, was the appearance of the British music hall, which was to have an even more decisive influence on the trends of popular music than the cabarets of Paris or the operetta theatres of Vienna. The music halls first sprang up in the big industrial cities as workingmen's beer halls with entertainment. Gradually, during the second half of the 19th century, the entertainment became more important than the drinking, and the beer halls evolved into theatres-with-a-bar, providing shows that largely consisted of humorous, satirical, or sentimentally emotional songs, mainly of lower class life. In the earlier years of the 19th century, visiting entertainers from the British music halls were instrumental in stimulating the American entertainment industry that in a later time, in the domain of popular music as in other fields, was to become the dominant influence throughout the world.

Music halls

During the whole of the 19th century the process by which music "from below" in the social scale is taken over and formalized by educated composers for the benefit of a middle and even upper class audience was vigorously under way. Out of the topical songs and festive dances (the carmagnole, for example) of the French Revolution emerged the tradition of Parisian cabaret song and such boisterous gallops as the can-can. Describing these songs and dances as he experienced them in lower class city dives, the German poet Heinrich Heine wrote:

> They make fun not only of sexual relations but of bourgeois relations too, and of everything else that seems good and fair . . . patriotism, faith, belief, family relations, heroism and godliness in every shape and form.

This originally proletarian satirical tradition was transformed to provide the basis for the lively bourgeois operettas of Jacques Offenbach, Hervé, Charles Lecocq, and others.

Also powerful in its effect on the popular musical theatre was the less mordant, more sentimental Viennese tradition represented by the younger Johann Strauss and by Franz von Suppé. This tradition had derived much from the *Liedertafel* (sociable-song) movement of popular harmony-singing clubs and from waltz music that had grown out of the Ländler dances of the Austrian and Bavarian peasantry put into evening dress and performed amid the glitter of high bourgeois candelabra. This style even invaded the realm of fine-art music and reached its peak in Richard Strauss's opera *Der Rosenkavalier*.

Even in the less developed parts of Europe and in Central and South America, popular music was expanding and undergoing transformation. In Spain, flamenco music was growing away from its folk origins, becoming progressively more showy and exotic in the cabarets and small theatres that sprang up around the middle of the century, notably in the Andalusian seaports. In Lisbon, the fado developed in much the same way. About the same time in the Balkans, the emancipation of serfs and the weakening of the Turks' grip meant that many private orchestras belonging to nobles and officials—playing an orientally tinged music—were broken up. This released a flood of popular instrumentalists seeking employment in the restaurants and night spots of the towns and cities. The result was the stylized, urbanized idiom loosely called "gypsy music," in which—as with flamenco—the exotic aspects were often exaggerated to satisfy the customers.

In Latin America the great developments in popular music were in the domain of the dance. Fashionable travellers from Europe brought to Havana, Rio de Janeiro, Buenos Aires, and Lima the minuets, gavottes, schottisches, polkas, and waltzes of Paris and Vienna. In the upper class ballrooms, these dances would be performed as received, but as they seeped down to the lower classes—which, from the Caribbean to Brazil, were mainly comprised of Negroes and mulattoes—the dances and the music accompanying them became altered with the choreography, melody, and rhythm current among the working people, producing the rumbas, sambas, and congas that were later to invade the ballrooms and dance halls of North America and Europe and were to have great influence on urban popular music over some parts of Africa.

"Black" minstrel shows

In the United States, an important development toward the middle of the 19th century was the emergence of the blackface minstrel shows, with white performers disguised as Negroes presenting mainly a burlesque of Negro song and humour. Touring groups, such as Daniel Emmett's Virginia Minstrels, who made their debut in 1843, and the Christy Minstrels, who followed shortly after, gave widespread currency to a vast number of popular songs, some of which are still sung. Companies such as these, as well as the growing number of vaudeville theatres in the U.S. towns, promoted on a vast scale the compositions of such songwriters as Stephen Foster, the composer of "Oh! Susanna" and "Old Folks at Home"; Henry Clay Work (1832–84), who wrote "Marching Through Georgia" and "The Year of Jubilo"; and George F. Root, whose best known songs are "The Battle Cry of Freedom" and "Tramp, Tramp, Tramp." A more peripheral but ultimately influential development was the emergence of a stratum of rural musicians who became entertainers with medicine shows and on the country-town vaudeville circuits, playing a stylized music of folklore origin. It was these minstrels who laid the foundations for the powerful hillbilly "industry" and its subsequent analogues, country and western and bluegrass in the mid-20th century.

The 20th century. *Operetta and musical comedy.* In the early years of the 20th century, the upper, more genteel reaches of popular music were still dominated by European operetta and opéra bouffe (French light opera). This tradition, characterized by Jacques Offenbach and even more by Johann Strauss the Younger, was prolonged in the 20th century by such composers as the American Victor Herbert, who wrote some 40 operettas including *Babes in Toyland* (1903) and *Naughty Marietta* (1910). Prominent among Herbert's rivals were Rudolf Friml, born in Prague in 1879, and the Hungarian-born Sigmund Romberg (1887–1951), both of whom had emigrated to the United States in the early years of the century. Among Friml's greatest successes were *Rose Marie* (1924) and *The Vagabond King* (1925). Romberg is best remembered for *The Student Prince* (1924) and *The Desert Song* (1926).

Transformation into the "musical"

At the same time that these talented composers were producing their European-style works—sweetly lyrical, making heavy use of waltz time, and in the main presenting unrealistic and exotic subjects—a new kind of musical play known as musical comedy or just "musical," was emerging in the United States. It is a genre that is somewhat brash in character, with breezy rhythms, usually offering scenes of American life in relatively realistic terms. Representatives of this vigorous trend were Jerome Kern, whose *Show Boat* became a classic; and George Gershwin, who had a brilliant run of successes from *Lady Be Good* (1924) to the quasi-opera *Porgy and Bess* (1935). The boisterous folkways of the American frontier provided two outstanding successes in Richard Rodgers' and Oscar Hammerstein's *Oklahoma!* (1943) and Irving Berlin's *Annie Get Your Gun* (1946). More sophisticated and silken were the works of the composer-lyricist Cole Porter, whose urbane musical comedies skillfully combined the traditions of New York, Paris, and Vienna. Leonard Bernstein's *West Side Story* (1957) was a continuation of the tradition of realistic American urban musical comedy, begun in the opening years of the century by the vaudeville performer George M. Cohan, with such works as *Little Johnny Jones.*

Throughout the first half of the 20th century and to an ever increasing degree, the popular music of the cities and towns in a great part of the Western world (and ultimately, beyond) was nourished mainly from the United States, with songs from the musical shows and other lyrical pieces composed by such well-known musicians as those named above. Although most of these compositions, following the common fate of much commercial popular music, proved brittle and ephemeral, certain of them have lingered on over the years to become standard songs in the popular repertory.

Jazz. Most of the creations of the musical-comedy composers leaned heavily on the conventions of "serious" music for their idiom and the manner of their performance. Nevertheless, as the half-century progressed, they were more and more influenced by the world of jazz. Indeed, the most powerful development in the field of popular music in the 20th century has been the emergence of jazz as a commercially viable entertainment for a public of the broadest social range. Jazz developed first of all as a synthesis of European ballroom dance music (via ragtime), brass-band music, and the songs of black labourers (especially folk blues) and—to a less extent—white workers. It became such a powerful trend in mass music because conditions in post-slavery America favoured a synthesis of black African and white European music, attractive through its exotic colouring and informal atmosphere. Moreover, the growing drift of country workers into urban areas meant that a new current began to flow into the culture of the cities, always avid for novelty. At the same time, from the beginning of the 20th century the most propitious economic and technical conditions developed in America for the marketing of this new music, for adapting it to suit the tastes of various layers of society, and for stimulating its mass production, dissemination, and global export.

Ragtime and blues

Significantly, the first kind of jazz to make an impact on the broad popular music market was precisely the genre lying closest to conventional "serious" music, namely ragtime. The ragtime composers were mostly Negro pianists trained in European music, often men of considerable artistic pretensions. The structures of ragtime derived largely from French or American light-music models rather mechanically infused with syncopations from Negro tradition. Prominent among ragtime composers were Scott Joplin and James P. Johnson. With its basically conventional idiom, ragtime was easily absorbed into the mainstream of commercial popular music, and by 1900 Tin Pan Alley—the professional purveyors of U.S. popular music—had taken it over. As early as 1899, a writer in *Musical Record* forecast "Ragtime will find its way into the works of some great genius and will thereafter be canonized," as if he foresaw Debussy's "Gollywog's Cakewalk" and Stravinsky's "Ragtime."

Only a few years later a deeper, more spontaneous and folk-rooted component of jazz, the blues, entered the popular music market by way of professional women performers on the music-hall stage. The pioneer blues singer "Ma Rainey" (Gertrude Pridgett) obtained her first engagements in 1902, and subsequently, in company with such singers as Bessie Smith and Bertha Hill, established a bold, expressive, essentially professional style that some call "classic blues." Eventually, however, in diluted and polished forms, with its realistic passions replaced by the poetry of bourgeois dreams, the blues was adapted for general show-business exploitation, whether in the most expensive night clubs or in the cheapest suburban dance halls. For all that, it was the lyrical dance music of small jazz bands rather than the piano rags or vocal blues that attracted the attention of large-scale commercial promoters. After 1910, as jazz music and jazz musicians moved northward from such centres as New Orleans to the large cities such as Chicago, Kansas City, and New York, elements of jazz flowed easily and to an increasing extent, into the "sweet" popular music that was the dominant product of the entertainment industry. Prominent among Negro composers who profited from the discoveries and inventions of a great army of nameless jazzmen, black or white, is Duke Ellington (1899–), whose compositions over the last half century have moved progressively closer to the realm of "serious" music. As far as the mainstream of popular music is concerned, the vitalizing effect of jazz was most felt on the standard popular lyrical song, with its 32-bar AABA chorus, the great mainstay of the 20th-century light-music industry.

It should be recognized that jazz is not a single category but a whole sphere of popular music, ranging from light salon music (Glenn Miller), "big band" dance music (Count Basie), popular symphonic (certain works of Gershwin and Ellington), to a kind of chamber music (Modern Jazz Quartet) or even to a form of musical Dadaism (Charles Mingus).

Folk- and jazz-derived styles. Jazz was not the only music coming up "from below" to mingle with and ultimately to affect the direction taken by popular music in the 20th century. Just as, in the lowland South, the rapid industrialization of the post-World War I period in the United States had resulted in the emergence of the predominantly Negro hybrid music called "jazz," so in the upland South—primarily in the Appalachian and Ozark Mountain regions—similar factors produced the predominantly white hybrid music called "hillbilly."

"Country music"

For much of the 20th century this music played but small part on the general scene of popular music. Shortly after World War I a number of the newly established U.S. country radio stations began to draw on the growing tradition of hillbilly musicians, interspersing local advertising with brief musical interludes provided, often free, by performers appearing in the neighbourhood. By 1924 it was realized that there was even a Northern market for this music; the first successful Barn-Dance program was established on WLS Chicago, mainly using country musicians from Kentucky. Around the same time the recording industry began to interest itself in the music, intending to issue discs mostly for sale in rural crossroads stores. To their surprise, they soon found there was some national demand, at first modest, for the product. Unlike jazz, which was based chiefly on wind instruments, hillbilly music relied on strings. Instruments such as violin, guitar, and banjo had become widely available in rural areas through urban contact and mail-order houses and were found to be relatively cheap, easy to combine in ensemble, and well-suited for a repertory that had grown to a large extent out of folk-fiddle music.

By the 1930s the place of "country music" was consolidated in the mainstream of popular music. The repertory of performers of "country music" represented a jumble of music deriving from folk-dance tunes, white adaptations of Negro blues, the sacred songs of fundamentalist Protestant sects contrasted with lyrics of drink and violence, ballad-like accounts of railroad disasters and hard times in the cotton mills, and, above all, a range of songs of commercialized pathos treating of dying mothers, orphaned children, bereft lovers, and lonely men far from home. During World War II "country music" expanded enormously through intense promotion by the popular entertainment industry, probably because the strong nostalgic bias of the music was particularly appealing during a time of wide-scale shifts of population into urban industrial plants and the armed forces.

By the mid-1950s, two contrasting styles of "country music" had become dominant: on the one hand, the Southeastern tradition had hardened into the impetuous "bluegrass" forms, giving prominence to a virtuoso five-string banjo style (perfected from North Carolina folk tradition by Earl Scruggs). On the other hand, the Southwestern tradition, less folkloric, more pliant, happily accepted elements from a metronomic, jazz-influenced city Negro blues backed by a heavy thumping bass (called "rhythm and blues"), and from this amalgam, rock 'n' roll emerged.

During the 1950s, the growth of a profitable market among adolescents, combined with the invention of electronic instruments allowing the easy production of a powerful and impetuous sound, stimulated the enormous diffusion of rock 'n' roll music, which relies largely on amplification for its exciting effects and which proved irresistible to many young people throughout the world. Subsequent modifications of "pop" music have retained the metronomic quality present in the music of Elvis Presley and later rock 'n' roll performers, but have also reinforced its blues ingredient. Rather paradoxically, what was originally an intensely American kind of popular music was to find some of its most characteristic modifications and its greatest commercial successes in England with such performers as the Rolling Stones and most notably and influentially with the four young men who comprised the group known worldwide as the Beatles.

Toward the end of the 1960s, a new force began to enter the world of professional popular music, coming mainly from the white middle class, particularly the more libertarian fringes of the intelligentsia. This music was derived in large measure from a parody of the blues, with the old Negro slum preoccupation transposed into bourgeois terms, complaining of spiritual rather than material deprivation and commenting, at times frenetically, on the general dilemmas of the world rather than on personal domestic tragedies. In its various nuances such as folk, acid, and hard rock, this form of music, sometimes called "Progressive Pop," existed successfully alongside the more conventional popular music, at times approaching the aleatory fragmentation characteristic of some present-day "serious" music.

Urban blues and "soul" music

Concurrently, the folk-rooted blues tradition persisted in the cities, reflecting the culture of the Negro ghetto-dweller, yet appealing also to audiences beyond the ghetto. The older city blues evolved into the urban blues style, marked by freer vocal phrasing and, often, larger, more electronically amplified ensembles; and, through urban blues, into "soul" music, which fuses blues with jazz and gospel music influences.

Initially formed mainly in the south and west of Europe, Western popular music has gradually spread throughout the entire world, at first to the ends of Europe, then to the colonized Americas, then to the cities of the Orient, and now into the maize and paddy fields and even the jungles of peoples who had hitherto remained primitive. The rise of the United States to a position of economic and political power accelerated the process by which Western popular music replaces the proper musical arts of Asians, Africans, and other peoples, a process already begun with the weakening of tribal and feudal traditions and the introduction of formal education through Western colonialism. Missions, schools, trade, radio, films, development aid projects, have all played a part in encouraging a taste for Western popular music. At the same time, throughout the Communist world, sovietization carried the more formal idioms of popular European music across a vast area of the globe, acclimatizing it by means of political rather than commercial propaganda.

Throughout the Communist world, popular music has become mainly identified—at least officially—with stylized folk music performed in a diluted manner, while in the Western world and throughout the sphere of Western influence, jazz music has seemed progressively to shrink into an intellectually rarefied atmosphere, replaced in the affections of the younger mass audiences by the emotional and musical simplifications of rock 'n' roll and its successors. Musical comedy, in theatre and in motion pictures, was still vital in the 1970s, but salon music was fading. For the global intrusion of Western popular music, the Orient extracted a mild revenge, with the introduction in the 1970s of one or two Eastern instruments (such as the sitar) and a few tentative Oriental intonations into popular instrumental ensembles of both the U.S. and Europe.

BIBLIOGRAPHY

General works: WATER WIORA, *Die vier Weltalter der Musik* (1961; Eng. trans., *The Four Ages of Music*, 1965), a concise and penetrating account of the evolution of music through four great periods—prehistory, antiquity, the growth of Western music, and the era of global industry—that sets the development of popular music clearly within the framework of general musical culture; REGINALD NETTEL, *Seven Centuries of Popular Song* (1956), a social history of English popular (as distinct from folk) song from the Middle Ages to the era of ballroom jazz (handy for the layman); S.G. SPAETH, *A History of Popular Music in America* (1948), a broad panorama of popular songs and the musical theatre from colonial times to the mid-20th century.

Specialized works: WALTER SALMEN, *Der fahrende Musiker im europäischen Mittelalter* (1960), a comprehensive social-historical study of the way of life of itinerant professional minstrels throughout Europe in the late Middle Ages, with a musicological account of their vocal and instrumental repertory—the best account of this important and little-understood class of popular musicians; C.M. SIMPSON, *The British Broadside Ballad and Its Music* (1966), a monumental reference work that relates some thousands of broadsides from the mid-16th to the early 18th century to the 500-odd

tunes for which they were designed; ANDRE HODEIR, *Hommes et problèmes du jazz* (1954; Eng. trans., *Jazz: Its Evolution and Essence*, 1956), an analytical work for the musically educated—brilliant, if a little overbalanced toward the modernist viewpoint; M.W. STEARNS, *The Story of Jazz* (1956, reprinted 1970), a serious study, regarded by some informed critics as the only history of jazz worthy of the name; B.C. MALONE, *Country Music U.S.A.* (1968), a broad-based account of the development of white rural professional music from its humble origins to its present state as large-scale commercial entertainment; CHARLES GILLETT, *The Sound of the City* (1970), an exhaustive survey and acute analysis of the emergence and growth of rock 'n' roll, and the consequent global changes in commercial popular music.

(A.L.Ll.)

Popular Theatre

To engage in an act of theatre, a living performer does not need a specially designed building nor elaborate equipment nor a written text nor a high artistic aim; all he needs is a place, a time, an audience, and himself. Performers in music hall, vaudeville, burlesque, follies, revue, extravaganza, circus and sideshow, musical comedy, and the like are as much a part of the theatre as those who appear in what is sometimes—rather grandly—called the "legitimate," "high," or "artistic" theatre. The singers, dancers, comedians, clowns, puppeteers, jugglers, acrobats, conjurers, ventriloquists, and others who make up their number refer to their profession as show business, or "show biz," and their unashamed intention is to please and astonish their audiences by the display of some special visual, vocal, or physical skill. They provide what the director Sir Tyrone Guthrie called the "theatre of delight," and they are part of a popular theatre that has at times itself been of the highest importance in the history of world theatre, reaching its peaks in the theatre of ancient Greece, in that of Renaissance England and Spain, and in the commedia dell'arte tradition of 16th- to 18th-century Italy. Outside the West, the Japanese Kabuki theatre, various kinds of puppetry throughout Asia, and certain other forms of Oriental dance or of dance-drama are examples of popular theatre that also achieve the status of high art. The present article considers popular theatre in terms of its impulse to generate exuberance and joy, the nature of its audiences, and its capacity for change; it then examines the basic ingredients of most popular theatre and concludes with a section on its internal continuity and its impact on other kinds of theatre. Popular theatre is set within the larger context of theatre art in the article THEATRICAL PRODUCTION, and its history side-by-side with "high" theatre is sketched in the article THEATRE, WESTERN and in general articles on the arts of other cultures. Its ability to transform sometimes crude facilities into a theatre is noted in the article THEATRES AND STAGES. For other general information, see THEATRE, ART OF. For more detailed articles on certain forms relevant to popular theatre, see CIRCUS; COMMEDIA DELL'ARTE; MIME AND PANTOMIME; MUSIC, THEATRICAL; PAGEANT AND PARADE; and PUPPETRY.

The people of "show biz"

SCOPE AND CHARACTER

Popular theatre adheres to the core of theatrical tradition in three crucial ways. First, it involves performances by one group of persons for another; second, it demands that the performers acquire at least a minimum of skills relevant to their activities on stage; and third, it usually follows an agreed-upon sequence and reaches a set conclusion. The last of these factors eliminates athletic or sporting events, which otherwise are skilled performances—a kind of "theatre of improvisation" or chance. With the almost certain outcome of the bullfight, however, an approximation of ceremony is achieved. (Religious ceremony may involve all three of the conditions, though a close analysis of the relation between religion and art will distinguish performance from ceremony, as well as the impulse to create a work of art from the impulse to worship.)

The notion of popular theatre should also be examined with reference to its basic impulses, its audiences, and its proclivity to change. Popular theatre has almost always involved an expression of physical pleasure and exuberance—indeed, of joy—that can be distinguished from the pleasure of the intellect or spirit. The harvest songs of praise and thanksgiving from which a primitive "theatrical" activity emerged were, among other things, expressions of joy that physical labours were finished and that physical needs were provided for. With regard to the makeup of its audience, popular theatre, in general, continues to be as broadly based as the theatres of Shakespeare or Classical Athens. Peasants doubtless used to watch the jugglers and comedians at a local fair in order to find momentary release from their toil, much as the intellectual today finds recreation in the circus or in the basement café, with its guitarists and folk singers. In the matter of change, fashion and taste prevail here as elsewhere in the arts: such forms as music hall, burlesque, and vaudeville have largely been replaced by various kinds of nightclub and television entertainment. Another kind of change occurs when a popular form is transformed into a more sophisticated brand of entertainment, less appealing to a broad public. The once widely popular art of the mime has been brought to a brilliant polish by such practitioners as Marcel Marceau in France, yet its audience has become restricted to a relatively small coterie. Puppetry, often a joy to persons of all ages, is now almost exclusively a children's entertainment in the West. On the other hand, the musical comedy, as it is understood in the mid-20th century, has drawn upon a long tradition of song, dance, and narrative to emerge as an art form in its own right while yet retaining its popularity and drawing its audience from a broad cross section of society (although the costliness of its presentation and the consequently high admission prices often exclude less affluent patrons).

Expressions of exuberance and joy

THE INGREDIENTS OF POPULAR THEATRE

Throughout its history popular theatre has made use of physicality and music as direct means of communication. The two are not always distinct; the movements of the dance, for example, are both physically and musically expressive.

The physical element. The physical element in popular theatre most often presents itself through eroticism, through distortion, or through the acrobatic.

Eroticism. The pre-theatrical ceremonies, characterized by excess, that accompanied spring and harvest festivals appear to have influenced the development of comedy in Classical Greece, where indecency and lewdness were integral parts of comic performances. In Roman times, the highly popular mime made such play of gross indecency that after the triumph of Christianity in the West during the 5th century AD this and all other forms of theatrical activity were officially banned. Theatrical traditions, however, were kept alive by itinerant bands of players, whose repertories undoubtedly made use of eroticism, though probably in humorous form. It was probably also eroticism, disguised as temporary lunacy, that was responsible for the dance manias that sometimes seized entire communities during the Middle Ages; and, as sentiment, it certainly pervaded the cult of chivalry and informed the songs of the troubadours and other wandering minstrels. Court entertainments in the 15th and 16th centuries, which evolved into ballet, tried to achieve in living form the idealized image of the human body that was represented in ancient Greco-Roman statuary. But the general direction taken by ballet tended away from the erotic. In spite of the glorification of the ballerina throughout the 19th century and the introduction of erotic themes in modern ballet, the form has virtually desexualized dance or else has taken sexuality into a rarefied realm outside that of popular theatre.

Influence of Greek spring and harvest festivals

It is the female figure in rhythmic motion, however, that has epitomized the erotic element from earliest times. In the 19th century, female dancing became an essential ingredient of popular theatre in four similar manifestations: revue, vaudeville, music hall, and burlesque. The first two, almost identical in form, go back in spirit to medieval France: the revue to street fairs, during which the past year was reviewed in satiric song and skit;

Revue, vaudeville, music hall, and burlesque

vaudeville to similar sources that, in the 18th century, were adapted to accompany the production of short pantomime plays by itinerant players. Revue and vaudeville had as many variants as there were companies performing them; their many offshoots tended either toward greater vulgarity or else toward greater refinement. Whereas the revue became most closely associated with France (though it provided the model for the "follies" type of show in most Western countries), vaudeville became most associated with the United States, in which it passed from a condition of low repute in the early 19th century to that of respectable family entertainment.

In England, the traditional music hall was virtually identical with vaudeville, originating in the "music annex" of working class alehouses but becoming standard fare for all classes of society. All three forms generally offered a variety of short pieces—sentimental and patriotic songs, dances of all kinds, comedians, humorous or pathetic skits, magicians and jugglers, tightrope walkers, and other acrobats. In burlesque, however, there was an insistent emphasis on the female dancer, and in the 1920s striptease became part of the entertainment.

Distortion. The gross deformity of a malformed body has often been used as a source of entertainment: theatrical wonder resides in the body itself rather than in anything it may do. People flock to such sights, perhaps as a means of alleviating their personal sorrows through a reminder of the greater misfortunes of others.

Fascination with deformity has a long history, social as well as theatrical. Primitive societies might either destroy or deify their defectives; barbarian people of odd appearance were herded along, strangely garbed, in the triumphal processions of the ancient world; dwarfish or humpbacked jesters, whether clever or dull-witted, provided kings with entertainment or served as butts for derision; genetic or birth defects resulted in people being called "village idiots" or in their being exhibited in circuses and sideshows; and the humour of "differentness" is still exploited by comedians who caricature, for example, a foreign accent or ethnic cliché. At the fullest artistic exploitation of distortion, Shakespeare often made his fools speak with greater understanding or compassion than did their masters.

Humour itself may be a distortion of reality—sometimes crude, as in the pratfall and the slapstick, sometimes kindly, as in the satires of a revue sketch, and sometimes corrosive, as in the barbed mockery of a modern-day "stand-up comic." Although at times humour reaches so far past the physical that it attains the sophistication of wit, it retains a basically physical impact. Its effect—earthy, ribald laughter—has been sought in all kinds of theatre, no matter how "serious," while the greatest playwrights have been able to blend the humorous and the serious in their reflection of life.

Use of the mask

On another level, distortion has served to generalize, to wipe out what is particular to the individual personality and to substitute a universal face that can be funny, pathetic, tragic, or anything else. The mask is the traditional means: wearing it, the performer is depersonalized, his character more obviously seen as symbolic. The Greek tragic and comic actor, the ancient mime, the Nō actor in Japan, the Harlequin and Columbine and Capitano of the commedia dell'arte all used the mask to diminish themselves and to extend their significance. Perhaps the ultimate mask is that of the puppet, where the individual is abandoned altogether. Sometimes a performer, playing in the popular tradition, adopts a private mask of such universal significance that the gap between popular theatre and high art is bridged. Outstanding examples in this connection are Charlie Chaplin's Little Tramp and Buster Keaton's deadpan mask of Everyman.

The acrobatic. The West seems to have lost almost all touch with the experience of magic, especially that of the shaman, the prophet-priest-medicine man whose intense and disciplined training allowed him to mediate between the people of his tribe and the mysterious, invisible forces of nature and who was also a physical and spiritual healer. His skills probably included almost the entire repertory of a modern circus. Many qualities invested in him survive in the West in a fragmented form. All artists, to a greater or lesser degree, follow a regimen such as that of the shaman, disciplining body and mind to the unique demands of their art, and all art may be said to reveal characteristics of the old magic: it excites wonder. Ballet dancers, opera singers, and musicians must keep their various instruments in fine tune, and actors must keep alive the openness that allows many voices to speak through them.

Nowhere, however, is there a greater and more conscious attempt to transcend the limits of the body, to attain an ever more astonishing level of accomplishment than in the art or craft of the acrobat, whether he is a juggler, a tumbler, a trapeze or high-wire artist, a knife thrower, an equestrian, an ice skater, or even a magician-prestidigitator. The shaman, whose maskings and whirlings were a precursor of drama and dance, was not, of course, a performer in the theatrical sense. Nor were his actions and feats carried out as demonstrations of prowess, as ends in themselves. Yet, in the desire to perform an act far beyond the reach of the average person or at least to create the appearance of stretching the human body beyond its capacities into some unknown new area, the acrobat and the prestidigitator retain some small element of the shaman's magical powers of casting spells. Like the shaman, the acrobat must punish his body to attain his skill, and his concentration must be absolute; for one inexpert move or one flicker of inattention can destroy him. And the drumbeats of the shaman to get the attention of the gods might be compared, in their effect on their human audience, to the drum rolls that intensify the suspense surrounding the acrobat's preparation for his next astounding feat of skill and daring.

Such displays of physical skill have been a mainstay of theatre from its earliest days, though they have been granted little esteem as artistic endeavour. With the demise of vaudeville and the music hall, they have retreated to the circus, the sideshow of the carnival or county fair, the ice show, and the television studio.

The importance of music. Few experiences in the theatre can equal the uncomplicated and unabashed enjoyment of "singing along" with the entertainer, whether in a crowded taproom with only an upright piano or in a gilded auditorium with a large orchestra. Indeed, there is little else that can encourage interaction between audience and performers as can music, the very presence of which tends to remove any pretense of realism, thereby joining performer and viewer even more closely in a shared event. Apart from its occasional function as a background accompanying other skills, music in the popular theatre has been employed chiefly in various kinds of musical-dramatic performances, from which today's musical comedy has evolved; in song-and-dance routines characteristic especially of the minstrel show and of vaudeville and its relatives; and in the stage singing and instrumental playing of cabaret, jazz, country-western, folk, rock 'n roll, and other "pop" performers who have gained favour at various times in the 20th century.

The effect of music

Musical drama. Musical comedy, as it is known today in the West, has had no straight line of development but has emerged from a wide variety of musical, dramatic, and dance styles. The use of relatively simple ballads to reinforce the narrative or dramatic structure of a play was known to the Elizabethan dramatists and was very popular during the early 18th century in works like John Gay's *The Beggar's Opera*. The tradition of Viennese operetta and that of the English comic operas of Gilbert and Sullivan was exploited in the United States in the early 20th century and Americanized in the musicals of Jerome Kern and others. *Oklahoma!*, in 1943, saw the introduction of balletic dancing and a tighter unity of song, dance, humour, and narrative, while in the late 1950s a more serious vein began to appear in such works as *West Side Story* (1957), *Man of La Mancha* (1965), *Hair* (1968), and *Company* (1970).

Song and dance. The entertainer known as a "hoofer," who could satisfy an audience by executing a good song-and-dance routine, was thought especially important in vaudeville and the music hall and on the showboats

that carried live entertainment down the rivers of the U.S. Middle West and South during the 19th century. Also in demand were sentimental and patriotic singers, including much-loved personalities such as Yvette Guilbert and Maurice Chevalier in France; Florrie Forde, Sir Harry Lauder, and Marie Lloyd in Britain; and George M. Cohan and Al Jolson in the United States. The work of such performers is inextricably linked with the history of show business in the late 19th and early 20th centuries, when entertainment became big business. In these decades, with improved communications in most Western nations, professional entertainment became widely available, most of it under the control of a few great theatrical empires that were created around 1900. London, Paris, New York, and other centres became the hubs from which variety and dramatic companies toured provincial theatres. Internecine warfare was practiced to control both talent and theatres: finally, one circuit tottered, then another, until the motion-picture industry began to grow, restricting the living theatre to urban areas.

The heyday of show business

Songs and instruments. "Pop" singer-guitarists and nightclub crooners are part of "show biz"—the latest in a line of singing entertainers that stretches back for centuries to the balladeer. Remnants of very ancient ballad traditions still exist, mainly in isolated regions of less industrialized nations. The urbanized West, however, has received watered-down versions of the old or musical materials manufactured to order (and to formula) through a commercial and competitive branch of show business once known as "Tin Pan Alley."

The advent of the phonograph and its major allies, radio and the jukebox, has increased demand for this kind of popular theatre in live form. In a small town, it is easier to bring together a number of players and a singer than to gather the diverse talents needed for a vaudeville show or a play, and such a group has an assured appeal, especially among the young. The modern balladeers, however much their art may fall short of the epics of Homer or the lyrics of the troubadours, continue to sing of love and peace as their ancestors sang tales of tribal wars, heroes, and heroes' loves. The erotic is often brought into play here, as, sometimes, is humour, while not infrequently the listeners respond with dance—as in those pretheatrical tribal ceremonies when performer and audience were one.

CONTINUITIES AND IMPACTS

Remnants of the popular theatre. Throughout much of the West and in other regions of the world in which Western (especially U.S.) influences have been strong, traditions and instances of live theatre seem to be on the wane despite public and private efforts to inject new life onto its stages. The traditional face-to-face confrontation of performer and audience has generated spontaneity and joy, however shopworn the materials. But the sweeping technological advances of the 20th century have altered, in a few decades, the centuries-old traditions of popular theatre. Circuses and ice shows may visit large cities once or twice a year; a vigorous puppet theatre may help to enliven the arts in a few areas of the world; various entertainments in cabarets, clubs, discotheques, and theatres still may be quite widely available. But the overwhelming proportion of what was the live popular theatre, both physical and musical, now comprises a major part of television's grist. Whether audiences and performers experience the same spontaneity and joy is doubtful. No matter how physically rambunctious, suggestive, or broadly satirical television may aspire to be, it remains hobbled by various kinds of censorship in ways the live theatre outside the family living room rarely was.

Popular theatre as the grist of television

Popular and literary theatre. Of the countless thousands of performers who have worked in the theatre, the great majority have done so only to make a living or to satisfy some personal drive to showmanship. Their displays of acting, singing, or acrobatics that have been given today must be repeated tomorrow. Before the arrival of film, only the playwright could survive in anything more than reputation. Popular theatre, with little or no literary base, has had to rely on tradition passed from one generation of performers and audiences to the next.

In his book *The Empty Space*, English director Peter Brook has linked popular theatre with what he calls the "Rough" theatre, saying of it that in the many forms it has taken there has been one constant factor:

> . . . a roughness. Salt, sweat, noise, smell: the theatre that's not in a theatre, the theatre on carts, on wagons, on trestles, audiences standing, drinking, sitting round tables, audiences joining in, answering back: theatre in back rooms, upstairs rooms, barns; the one-night stands, the torn sheets pinned up across the hall, the battered screen to conceal the quick changes—that one generic term, *theatre*, covers all this. . . .

Brook relates it to the freedom of Shakespeare's theatre, and contrasts it, by implication, with the respectable sterility that stifles much of today's legitimate theatre (which he dubs "deadly"). The "Rough" theatre, however, has consistently given refreshment and inspiration to the more "serious" theatre with which it exists side-by-side. In the present century, for example, such a director as Joan Littlewood in England has made uninhibited use of its sometimes vulgar but always vigorous current, borrowing its immediacy and excitement. Toward the end of his life Bertolt Brecht in Germany urged that the theatre must rediscover its naivety: his work had already borne witness to his point of view. The small and deliberately tinny orchestra that accompanied his *Threepenny Opera*, for example, was reminiscent of the limitations that a small cabaret might impose, and the format of this and many other of his works had much in common with a series of loosely linked vaudeville skits. Brecht's theory of dramatic alienation, formulated with the aim of forcing his audience into an objective analysis of the drama, owes a good deal of inspiration to fairground entrepreneurs, whose frank and unabashed methods of presentation are far removed from the comfortable theatre of illusion, where audiences are encouraged to sink back into a world of make-believe and think it "real." The clown and the jester of popular theatre, too, their comic images and antics turned to grimly serious purpose in the works of such playwrights as Samuel Beckett and Eugène Ionesco, still play a significant part in modern dramatic theatre, just as they did in the plays of Shakespeare and of Molière.

BIBLIOGRAPHY. Since popular theatre as discussed in this article is not a generally recognized genre, no critical study has been applied to the area as a whole; the closest is ABEL GREEN and JOE LAURIE, JR., *Show Biz* (1951, reprinted 1971), the story of show business in the United States for most of the 20th century as seen through the eyes of the industry's "bible," the weekly newspaper *Variety*. M. WILLSON DISHER has written extensively on aspects of popular theatre, especially *Music Hall Parade* (1938), *Blood and Thunder: Mid-Victorian Melodrama and Its Origins* (1949), and *Pleasures of London* (1950). ENID WELSFORD, *The Fool: His Social and Literary History* (1935); and DOUGLAS GILBERT, *American Vaudeville* (1940), remain valuable. *The Journal of Popular Culture* (quarterly) is a rich resource. Numerous motion pictures have probed into facets of show business; especially perceptive are *The Blue Angel* (German, 1930) and *The Entertainer* (English, 1960). Books containing more general discussions include MARTIN ESSLIN, *The Theatre of the Absurd*, rev. ed. (1968); PETER BROOK, *The Empty Space* (1968); and THEODORE ROSZAK, *The Making of a Counter Culture* (1969), especially ch. 8, "Eyes of Flesh, Eyes of Fire."

(G.W.W.)

Population

Any population, whether human or animal, constitutes a collection of elements modified by increases (births and immigrations) and losses (deaths and emigrations). A "natural" human population—that is, one that is unable to cope with death by means of efficient therapeutic methods or with birth by means of birth control and that is restricted to one geographical area—will find itself in approximately the same position as an animal population. Such a population will be unable to expand beyond a certain total; a herbivorous population, for example, will be roughly limited by the annual growth of grass. For purposes of illustration, it is helpful to consider a population, originally small, living in a restricted natural environment. Such a population increases because of its

multiplying powers, its birthrate being superior to its deathrate; at a certain stage it reaches a degree of saturation (lack of food, water, or other necessities) that results in an increase of deathrate, often accompanied by a decline in the birthrate; at some point the population reaches its position of equilibrium within the environment and it fluctuates in the neighbourhood of this maximum under the effect of the seasons (heat, humidity, and so forth), of diseases, or of other tribulations that may temporarily afflict it.

This "natural" balance can be lastingly modified, either through a change in the environment or through a change in the species. The environment may improve or deteriorate, mainly under the influence of the climate. The species may become better adapted to the natural environment, with a resulting increase in its maximum number; however, if such progress allows the species to consume its own capital (for example, by exterminating the prey upon which it lives) the results may be fatal.

Fertility

In addition to the conditions for natural balance other factors affect human populations. Some of these involve *fertility*—*i.e.*, the actual frequency of births in a population—as distinguished from *fecundity*, or a population's ability to bear children. (In French and other Latin languages, demographers tend to reverse the meanings of these terms—*fécondité* meaning "fertility" and *fertilité* meaning "fecundity.") It has been estimated that if all of the couples in any given population were to reproduce from puberty through the woman's menopause at an approximately natural rate, the population would multiply about six times every generation. The reasons that the human species has never actually achieved this rate of multiplication, however, are the following: (1) a fraction of the population does not marry or procreate; (2) not all males and females couple upon reaching puberty; (3) some couples are separated by death or by some other cause; and (4) the new generations are reduced in number by death before all of them, in turn, reach puberty or the age of procreation. In a natural population this last factor alone is enough to decrease the rate by more than half.

It is not easy to separate completely the biological and sociological factors listed above and to imagine a population that might use its powers of fecundity to the full. Theoretically, the only system allowing such total use might be a kind of human stud farm. In actual fact, random or promiscuous mating does exist among certain populations, but always in conjunction with very poor hygienic conditions that compensate for its generative advantages. Practically speaking, the matrimonial system that puts the multiplying powers of the human species to the best use is usually thought to be monogamy; a clear historical illustration of this is given by the French Canadians who, within three centuries, increased from some 3,000 initial inhabitants to several million.

Mortality

Another important determinant of population is mortality, or frequency of death, as it relates to average life-spans. In a human population that does not suffer from chronic excessive hunger or from overwhelming disasters but does exist under conditions of poor hygiene and inefficient therapies, the average life-span of an individual is some 25 or 30 years. In addition to this more or less natural mortality experience, human beings also experience an accidental super-deathrate, which may stem from disasters of nature (floods, changes of climate, etc.) but more generally appears under three conditions: acute famine, epidemics (from exterior sources), and wars or massacres. These three scourges can, in fact, be found to interlock, with war often provoking a state of famine, an epidemic spreading more easily among an undernourished population, and so on.

Finally, it is a fact that human populations are not merely arithmetical totals. They are characterized by variations in total numbers and also by changes in structure, including changes in the distribution of ages, occupations, and various qualitative, physical, and intellectual attributes (such as average or predominant height, weight, blood types, colour, nationality, intelligence, and level of education).

I. History of population

PREMODERN AGES

Until the contemporary period, the total population of the globe throughout millennia in general experienced a slow increase in numbers, with occasional, sudden periods of regression in numbers, in most cases following such phenomena as wars, epidemics, and famines.

Among primitive cultures and in ancient civilizations fertility was usually honoured, probably for the very reason that life is opposed to inevitable and yet so dreaded death. This feeling, which today is still strong among certain rural populations, nevertheless yielded to other considerations in the Greek republics. Ancient Greece was the first known, and relatively well-assessed example of a population that declined for reasons other than catastrophes—namely, because the culture favoured the family practice of having few children and, in fact, even favoured celibacy. The politician and historian Polybius actually denounced the decline of procreation in Greece, considering it tantamount to suicide.

That the concept of birth control existed in the earliest days of the Roman Empire seems evident in the fact that the emperor Augustus instituted laws to encourage marriage and, above all, childbearing. Although no proof exists, birth control, along with other Roman influences, may have spread throughout the empire beyond Rome. In any case, if the imperial population had been governed by natural multiplication, it should, in four centuries, have increased so sharply as to swarm out beyond the borders or, at least, to have densely populated the existing territories of the empire. But it did not, for there was a shortage of men to face the barbarians, themselves undoubtedly relatively few in number. Even the recurrent and widespread plagues, particularly from AD 166 to 180, do not satisfactorily explain the shortage.

Another ancient country, China, whose demographic history remains to be written, seems to have been densely populated from very early times; overpopulation was commonly deplored by ancient Chinese writers.

From the fall of the Roman Empire to medieval times, Europe was beset by massive invasions and by vast disorder and destruction, and populations tended to decline. Beginning with the 9th century, however, a certain degree of human stabilization and some signs of progress may be observed in Europe. These became more evident in the course of the 11th century and are almost certainly confirmed by the Crusades, which could not have been undertaken by sparse populations. The 14th century, on the other hand, was a time of severe crisis, which stemmed from famines and wars and particularly from the Black Death, the plague that ravaged all of Europe from 1347 to 1351 and resulted in a mortality varying in the different regions between one-eighth and two-thirds of the population; on an average it may possibly have killed off as much as one-third of the population. After some increase in population during the 15th century there ensued an uneven period, a time of bloody wars and the beginnings of emigration away from Europe. On the American continent the demographic contribution of the white man was at first often more than offset by the decline of the native populations, which were subject to decimation and massacres, new diseases introduced by Europeans, and expulsion from their homes and customary means of livelihood. As contacts with the other parts of the world became more frequent, Africa was revealed as a half empty continent and Asia as a teeming mass of human beings.

The 17th century was marked by two important additions to knowledge: in 1662, one John Graunt established the first genuine table of mortality, and various other authors estimated the world population (a Dutch scholar, Isaac Vossius, suggested a figure of 545 million).

THE MODERN ERA

The first demographic revolution

In the 17th century all populations, even those of western Europe, could, from a demographic point of view, still be described as "natural." Medical knowledge had made great strides in anatomy and physiology, but not in therapy, and the efficiency of antimortality techniques as a

whole was negligible. At the same time, contraceptive practices, as well as being inefficient, were reserved for so small a fraction of society (prostitutes and the aristocracy) that they had no influence upon global statistics. The normal annual deathrate reached 30 to 35 per 1,000, and the birthrate approximately 35 to 40, limited by partial or total celibacy. Thus from the earliest days of mankind, its biological conditions had scarcely changed.

This was the period that witnessed the start of a far-reaching transformation—known as the demographic revolution—which was to modify the life and reproduction of the species, thanks to a significant decline in both the deathrate and the birthrate. This transformation would have two phases, first among the advanced countries, and later throughout the rest of the world.

Decline of the deathrate in the West

Changes in the Western world to the 1930s. In the course of the second half of the 18th century, with such innovations as smallpox inoculation and vaccination, therapy became more efficient. A slight decline in the deathrate had occurred even earlier, however, as a result of improved hygienic conditions and, possibly, of improved food supplies. Furthermore, the super-deathrate from acute famine and epidemics virtually disappeared during the 18th century in western Europe because of agricultural advances, international trade that improved the availability of all resources, and better hygienic defenses (the famine of 1847 in Ireland was atypical). It was also in the course of the 18th century that epidemics of the plague ceased in western Europe, though they persisted for long periods of time in the less developed countries—as did cholera, which persisted even in Europe until the middle of the 19th century.

The decline of the normal deathrate became notable toward the end of the 19th century, after Louis Pasteur's experiments and the discovery of various vaccines; but it was a phenomenon still restricted to developed countries. On the eve of World War I, expectation of life was over 50 years in the most advanced countries.

Although the decline of the birthrate started in France at about the same time as the decline of the deathrate (because of the spread of contraceptive practices, particularly the use of the condom and coitus interruptus), the decline was experienced only much later in the rest of western Europe. In England the birthrate did not begin to decline before about 1875. Because of this time lag of nearly a century between the decline in the birthrate and that of the deathrate, Europe's population increased at a tempo hitherto unknown, at least over any long period of time and over such a vast area. The natural rate of increase of 1 percent was frequently exceeded. Even with the heavy emigrations to the Americas between 1801 and 1901, populations increased from 11,000,000 to 38,000,000 in the U.K., for instance, and from 25,000,000 to 56,000,000 in Germany.

Decline of the birthrate in the West

After 1875 the decline of the birthrate became more pronounced and spread to all the developed countries. A few typical rates are shown in Table 1. Birthrates reached their lowest levels in the course of the 1930s. Generations no longer provided their own successors, and there was already being advanced the theory of a permanent decline of the Western races. The decline was accentuated by the economic depression.

Table 1: Annual Births per 1,000 Inhabitants

	France	U.K.	Germany	U.S.
c. 1880	24.6	33.9	38.0	36.0
c. 1900	21.4	28.5	35.0	30.1
1911–13	18.8	26.3	28.0	27.0
1933	16.2	19.8	16.7	16.5

In the developed countries the falling birthrate caused a significant aging of the population. The proportion of sexagenarians rose to about 10 percent, as compared with 6 percent in the 18th century.

In the rest of the world there was no demographic revolution; the birthrate remained in the neighbourhood of the natural level and the deathrate declined only slightly.

Table 2: Pre-World War II Annual Birthrates and Deathrates per 1,000 Inhabitants

	birthrate	deathrate	natural increase
Africa	40	36	4
U.S., Canada	18	11	7
Latin America	42	24	18
Near East	42	33	9
Southern Asia	42	32	10
Japan	27	17	10
Far East	42	32	10
Southern Europe	23	16	7
Eastern Europe	33	22	11
Rest of Europe	17	12	5
Oceania	20	10	10
World total (*c.* 1939)	36	26	10

In Table 2 are shown regional approximations of the birthrates and deathrates per 1,000 inhabitants in 1939 on the eve of World War II.

The second demographic revolution

Changes after World War II. World War II and the years immediately following witnessed far-reaching changes: (1) The birthrate resumed an upward trend in the developed countries. The significance of the phenomenon lies in the fact that it constituted a change in a century-old trend, which earlier had been considered irreversible. (2) The deathrate underwent a sharp decline in the less developed countries. The antimortality techniques of advanced civilization significantly outstripped economic and birth-control techniques, because the most important of the former (such as water purification and vaccination) demand neither heavy capital investment nor much cooperation from the population. Though very uneven (depending on the countries involved) the deathrate at times has declined in major proportions; life expectancy sometimes reaches beyond 60; the deathrate falls often below 10 per 1,000, a lower rate than that of the developed countries because of the relative youth of the populations in the less developed countries.

Because births have remained at their traditional rates (or have sometimes even risen slightly because of the decline of puerperal mortality), the less developed countries have experienced a far more rapid population increase than did the European countries in the 19th century. The rate of 2 percent, previously unknown in history, is exceeded almost everywhere; many countries reach 3 percent (and even 3.5 percent in parts of Latin America). A rate of 2.5 percent represents a doubling of a population within one generation. A typical family, which in days gone by had five children, of whom two survived, may now have six, with four surviving.

Thus, the world, which until now had known two types of population—one traditional, young, highly fertile, but subject to a high deathrate, and the other modern, older, its birthrates and deathrates both low—is today experiencing a third type of population, young, endowed with both a high birthrate and a low deathrate. Such a condition if it lasted very long would result in earth's population being multiplied 32,000 times in only 500 years. Though present rates must, therefore, be considered temporary or transitory, just as Europe's population increases were temporary in the 19th century, the length of the transitional period can be of great importance.

Table 3: World Population Growth

year	population estimates
1000	340,000,000
1650	545,000,000
1800	907,000,000
1900	1,610,000,000
1950	2,509,000,000
1970	3,650,000,000

The aging process that had been affecting the developed nations for about a century continued after World War II, despite the upward trend of the birthrate. The proportion of sexagenarians in some developed countries is now over 17 percent and should continue to rise. The less

developed countries, on the other hand, have more or less retained their former age distributions, so that not only is their combined population two and a half times as great as that of the developed nations but also, with a more youthful population of procreative ages, their potential growth is accelerating. The less developed nations have three to three and a half times more people who have not passed out of the childbearing age.

Table 4: Present-day Age Distributions
(percent)

age groups	developed regions	less developed regions
0 to 14 years	27	42
15 to 64 years	63	55
65 years and over	10	3

It might also be noted that worldwide population distribution by sex never reveals more than minor variations. The approximate proportion of 105 boys born for every 100 girls (observed in all countries) is sooner or later corrected by a differential in their deathrates. Even when the balance between males and females is occasionally disturbed by wars or migration, it is gradually re-established by the rise of younger generations.

From a geographic point of view, the most striking 20th-century change in developed countries has been the migration from rural to urban areas as a result of the relative decrease in the need for labour in the primary sector of the economy (agriculture and fisheries) in favour of the secondary (industry) and tertiary (services) sectors. In the past 20 years, however, the less developed nations have also experienced shifts to urban areas.

Table 5: Rural-Urban Ratios
(percent)

	1950	1970
Developed regions		
Rural population	49	32
Urban population	51	68
Less developed regions		
Rural population	84	24.5
Urban population	16	75.5

Social characteristics of contemporary population. *Characteristics of modern fertility.* Standards of living and cultural traditions greatly influence fertility during the transition from a demographically primitive to a highly developed society. Birth control begins among the wealthy and the highly educated and in cities, but at a later stage it embraces all social classes and, among European countries, an opposite trend eventually may be observed: although the influence of urbanization generally keeps the birthrate low, prosperous couples in the cities again tend to have a larger number of children. In Table 6 are shown the number of children per 1,000 couples in West Germany in 1962, categorized by level of income and by the size of the community.

Birthrates among social classes and in cities

Table 6: West Germany, Number of Children per 1,000 Couples

income (Deutsche Marks)	townships of fewer than 3,000 inhabitants	towns of 3,000 to 50,000 inhabitants	cities of 500,000 inhabitants and over
Under 600	1,955	1,701	1,385
600 to 800	1,961	1,771	1,494
800 to 1,200	1,858	1,798	1,550
1,200 and over	1,944	1,882	1,630

Aside from noting the influence of urbanization, there has been insufficient exploration of various social factors influencing deliberate birth control. It is known, however, that housing conditions, for instance, seem to have different effects in the poor and the rich nations. Among a generally impoverished population, in which individuals live in overcrowded conditions, there is usually a rather high natural fertility, whereas in an economically developed population, unsatisfactory housing may well reduce the number of a family's children. It can also be noted that although traditional religions, such as Roman Catholicism, have not been powerful enough to put a stop to birth control and the reduction of the number of children born in wedlock (as influenced by economic and social evolution), they have been able to apply brakes to such reductions (as in Canada, Italy, and Spain). The fertility differential resulting from this braking effect has actually resulted in a change in the religious balance of certain countries—between the Catholics and Protestants in The Netherlands, for instance. When interpreting such statistics in terms of religion, however, it is important to keep in mind the fact that faiths are often linked with social status—the Protestants in Northern Ireland, for instance, enjoying higher social and economic status than the Catholics.

Characteristics of modern mortality. From the 18th century onward, writers have tried to assess the influence of professions and of social standing upon the deathrate. Although social-class inequality in the face of death has always existed, to the extent that the lower classes have been relatively undernourished, such inequality must have become even sharper during the 19th century, when a more efficient, though expensive, medical science began to afford the middle and upper classes a greater life expectancy. The deathrate was possibly twice as great among the lower classes. Today, however, this gap appears to be closing, at least in highly developed countries, because of improvements in housing and in the food available and because of the extension of medical care through welfare service. The disparities that persist in developed countries seem to be related more to levels of education than to levels of income. Nevertheless, poverty and malnutrition still plague various minority groups in such affluent countries as the United States.

Social class differences in mortality

In the underdeveloped world rampant death unquestionably still stalks the poor. A survey by the Food and Agricultural Organization of the United Nations in 1963 found that at least 60 percent of the people in less developed areas were undernourished and that half the world's people suffered from malnutrition. Also, although such large-scale preventive medicine as the Ceylonese eradication of malaria and other diseases have indeed made possible a startling decrease in the deathrate, the lack of adequate physicians, hospitals, and medical services under most welfare programs has reduced the possibilities of individual attention to the poor. Thus, class disparities in mortality seem to be still very much a part of the less developed countries.

There remains the question of whether or not, in the advanced countries at least, the drop in the deathrate that began in the 18th century is now nearing its end. Opinions on whether or not man's life-span can continue to extend and, if so, how far, are widely divergent. The optimists (the Soviets especially) currently hold that longevity will still notably increase, though they recognize the present impossibility of assigning any precise limits to such growth. The majority of scholars and medical specialists in gerontology think otherwise, although they do not rule out the possibility of entirely new discoveries. For the time being it can be assumed that progress is being made in the struggle against specific diseases, cancer in particular, but that the slow attrition of the body, which begins at a very early age, has not decreased its tempo. At the present level of knowledge, the average extreme limit of age of the human species would seem to remain invariable at some 107 years, the deathrate increasing sharply with the latest years of life.

Characteristics of modern international migration. With the total absorption of all land areas by sovereign states and with the imposition of restrictions on immigration (and sometimes on emigration), the 20th century has added new characteristics to patterns of global migration. Although the free movement of individuals has always been impeded by poor health or ignorance, attachment to habitual sites, the costs and risks of migration, the availability of economic opportunities, and, to

New immigration restrictions

some extent, political barriers, these political barriers lately have assumed major importance. Although the sick, the physically handicapped, and the insane, for example, have always been hindered from migrating through a kind of process of natural selection, nations have now generally imposed a stricter process of physical selection, demanding that immigrants submit to medical examinations. Similarly, not only have the young always emigrated more willingly than the old but now most host countries flatly restrict the entry of old people. Host countries also usually specify occupational or economic conditions, often receiving migrants only from among those professions in which they find themselves shorthanded.

For countries of origin, emigration has generally served, in favourable terms, as a means of releasing population pressures, of extending influence overseas (as in the days of exploration and colonialism), and of providing economic and occupational opportunities. In unfavourable terms, it has occasionaly meant a "brain drain"—the loss of highly qualified individuals—which has often been very pernicious (as in France, when the Protestant Huguenots were expelled in the 17th century, or in Nazi Germany, from which many Jews and intellectuals fled when faced by persecution). The brain drain continues today, though more slowly and principally in the direction of countries already highly advanced industrially and professionally. But emigration can scarcely any longer solve the pressures of population. In the heavily populated areas of the underdeveloped world, the sheer magnitude of population and the force of natural increase exclude the possibility that international migration can be a major factor in raising levels of living, at least for the foreseeable future. Only in a few countries pressured by moderate overpopulation, principally in southern Europe, can emigration still relieve demographic problems.

For the host country, immigration under current conditions of restriction can be advantageous. Screening applicants at the point of entry permits a country to improve its own professional profile (and ultimately actually reduce unemployment). The country also profits by the arrival of mature men, who have reached their fullest "value" and whose upbringing and education it has not had to pay for. Restrictive policies also avoid the problems of having to assimilate large numbers of people of culturally different backgrounds.

Characteristics of modern internal migration. The most important migrations within the boundaries of individual states have been those from rural areas to the cities. Throughout history, the rural exodus has been due largely to demographic factors (for a long time the deathrate was higher in the cities, and later on, the birthrate was higher in the country). Today the rural exodus in advanced countries is mainly the result of technological progress: fewer men are required to till the soil and operate the productive machinery of modern agriculture, and more men, conversely, are demanded by the urban systems of industrial production, distribution, and management. Even in less technologically developed countries, the considerable increase in total population has resulted in an overflow of the surplus from the overpopulated rural regions toward the cities. (The reverse pattern of movement from urban to rural areas, incidentally, is rare and not always spontaneous. The government of the People's Republic of China, for instance, has on different occasions forced a return to the land.)

But the rural exodus is not the only variety of internal migration. With improvements in transportation, communication, power sources, and various distributional techniques, manufacturing and service industries need no longer be concentrated in urban areas. Furthermore, there is often the lure of lower taxes in nonurban areas. Thus there is evidence of some dispersion of population or at least the possibility of such dispersion. One or the chief reasons why industries—and, with them, populations—have not dispersed as much as they could is that these populations generally resist moving from the traditional centres of urban culture.

II. Population beliefs, theories, and policies

PREMODERN BELIEFS

Pronatalism among the ancients

From the earliest times when men first began to live together in groups, they were faced by problems of numbers. To help ensure their own survival most ancient peoples and civilizations fostered fertility cults and showed lively concern for the numbers of their descendants. Ancestors were honoured; sterility was considered a curse. The laws of Manu in Hindu mythology, for instance, stressed the holy duties of matrimony and paternity, just as did the Persian Zoroaster and his followers. On the other hand, Confucius was aware of overpopulation in northern China and may have been the first known writer to express the concept of an "optimum population." Various later Chinese writers expressed alarm over, or warned against, overpopulation.

The first legislators of ancient Greece—in particular, the legendary Spartan Lycurgus—were concerned with population levels, as were the philosophers Hippodamus (second half of the 5th century BC) and, particularly, Plato (*c.* 428–348/47 BC), who actually conceived of a republic whose members were to be kept at an optimum number (5,040 citizens) by means of birth control. The Romans, on the other hand—troubled by a falling birthrate and primarily concerned with national power—were, generally speaking, pronatalists. Many wrote about it, and the emperor Augustus promulgated laws encouraging marriage and childbearing.

Although Judaism had been guided by the injunction "Be fruitful and multiply," Christianity in its early days (from St. Paul to St. Augustine) attached little importance to population increases and, in fact, actually extolled celibacy. Later, Thomistic Christianity was to manifest greater interest in prolific families and to oppose all birth-control practices. The parson of Chaucer's *Canterbury Tales*, for instance, scorned "adultery in marriage" ("avowtrie . . . betwixe a man and his wyf")— that is, putting pleasure in sex before the desire for procreation. Taken as a whole, Europeans in the Middle Ages were pronatalists, as were their antagonists, the Arab Muslims; the historian Ibn Khaldūn favoured dense populations and elaborated a complete economic, demographic, and political system for the life of empires.

BEGINNINGS OF MODERN POPULATION THEORY AND POLICY

Population theories involve the effort to explain the lives of populations in their economic, social, and political environments. They attempt to render historical trends and events intelligible and thus to provide a basis for qualitative judgments and policies. Although this kind of deliberate wedding of theory and policy occurred sporadically throughout previous history, it was not until early modern times, beginning in the 16th and 17th centuries, that theories of population and its control began to acquire the interest and importance that they have today.

Mercantilism. One of the earliest, most broadly based, and most widely accepted population theories was contained in mercantilism, a school of thought dominant from the 16th to the 18th centuries, which was pronatalist and more concerned with national (or royal) power than with the standards of living of the inhabitants. According to this most consistent doctrine, each nation should wish to have a large population, for a large population would provide a supply of labour, a market, and a source of soldiers for effecting expansionist policies.

When mercantilism began to decline rapidly in the 18th century, it was supplanted by a confusion of views. Not characterized by a uniformity of ideas, the 18th century was a period of uncertain groping for policies as the earliest elements of the Industrial Revolution began to emerge. The century nevertheless remained pronatalist on the conservative governmental level in Europe—the state's aim being to increase the power of the sovereign and the wealth of the governing classes. As Turmeau de La Morandière, a noted populationist, cynically wrote, "We must multiply the number of subjects and cattle."

Laissez-faire theory. Perhaps the major 18th-century school of thought touching on population questions was

physiocracy, often regarded as the first "scientific" school of economics. With it is associated the well-known phrase laissez-faire, which perhaps suggests the physiocrats' belief that the state should lift its vast complex of mercantilist rules and regulations and let nature take its course. The physiocrats did, however, recognize the fact of growing populations and urged private investments (such as in clearing forests for agriculture) in order to provide expanding employment. It is also notable that, although the physiocratic influence on Adam Smith has often been cited, Smith himself evinced little interest in population matters, beyond claiming that an increase in a nation's aggregate wealth fostered increases in population and that population growth generally contributed to overall human happiness (if not to every individual's happiness).

Laissez-faire attitudes toward population policy

Malthusianism. Although various authors in 17th- and 18-century Italy, Germany, and England had warned against the dangers of overpopulation, the first definite and thorough association of these problems with social issues occurred at the very end of the 18th century. The 19th century produced the long-drawn-out debate between the exponents of two conflicting dogmas, expounded by Thomas Malthus and by Karl Marx, about relations between population and economic conditions.

Malthus, who in 1798 published the first edition of *An Essay on the Principle of Population, as It affects the Future Improvement of Society, with Remarks on the Speculations of Mr Godwin, M. Condorcet, and other Writers*, held that the increase of population caused by "the attraction between the sexes" constantly tended to exceed the food supply and other means of sustenance. The population was always subject to the "positive checks" of famine, war, and disease. Each technological advance was absorbed by a consequent increase of population, thus preventing any rise in the general standard of living. Increases in production, in Malthus' view, could never equal much less exceed increases in population. He later modified this position, however, by suggesting that under some conditions population might be brought into a more favourable balance with production through a "preventive check"—the postponement or avoidance of marriage and procreative activity.

The French philosopher Marquis de Condorcet had suggested that fertility might be controlled within marriage. This solution was repudiated by Malthus not only on moral grounds but also for economic reasons, because he viewed large families (and the consequent poverty) as a necessary stimulus to economic activity. He indeed opposed certain welfare measures such as the English Poor Laws, which he viewed as futile and dangerous, and he repudiated the idea of improving the condition of mankind through changes in economic and social institutions. In general, he defended private property in such a fashion that he provoked violent reactions among outstanding contemporaries, such as the political philosopher William Godwin and particularly among the doctrinaire socialists of all schools—Robert Owen, Charles Fourier, Pierre-Joseph Proudhon, and Karl Marx. These men reproached him with trying to lay at the door of the poor the guilt or responsibility for their own indigence, which, they contended, was on the contrary the result of the selfishness of the rich. "There is only one man too many on this earth," wrote Proudhon, "and that is Mr. Malthus."

Marxism. Whereas Malthus—in the laissez-faire, classical tradition of economics—had opposed deliberate intervention or change in social and economic institutions, Marx viewed revolutionary changes in economic and social institutions as inevitable and as leading toward an ideal society. He violently attacked Malthusian theory. He equated overpopulation with actual or latent unemployment; he described relative overpopulation in this sense as essentially and purely a phenomenon of the capitalist system. In other words, Marx—and indeed most orthodox Marxists and communists up until very recent times—tended to deny the existence of overpopulation and asserted that what seemed to be overpopulation was actually capitalism's failure to provide adequate employment. If such a view now seems simplistic, it must be remembered that Marx was reacting not only against what he considered to be a vicious Malthusianism but also to the relatively less critical population conditions of his time. Life expectancy at birth in his day was scarcely over 40 years in Europe (as against 70 today), and the number of deaths between 0 and 20 years of age was 35 per 100 births (as against 5 per 100 today).

Marxist theory of seeming "overpopulation" and unemployment

Exponents of Marxist doctrine have generally viewed the control of population and changes in economic institutions as alternative and incompatible approaches to the advancement of levels of production and income. In fact, the later division of Marxists into socialists or social democrats on the one hand and communists or orthodox Marxists on the other contained, among other things, a disagreement over the question of population control. The social democrats (and also the anarchists) came out in favour of birth control by means of contraception or even abortion. The communists, faithful to Marx, interpreted this attitude as capitulation before the capitalist system. According to communists, a socialist state would be at all times capable of feeding its population, however large it might be. Even today, although they have modified their attitude, the socialist countries still basically pay homage to Marx's theory.

CONTEMPORARY TRENDS

Modern theories. Most 20th-century demographers and economists view population size and environmental resources, age distributions and social security, and the labour force and education (or the development of "human capital") all as complementary elements in arriving at population theories and policies. It is true that until recently the social aspects of population (or the "quality" of life, as perhaps undermined by crowding and pollution) have received less attention than the economic aspects, but it is also true that these have now come to be seen by some as possibly even more important. Because of the difficulties still existing in trying to measure such qualities, however, the discussion here will be limited largely to standards of living and collective or national considerations.

Optimum population. The term optimum population, introduced as a scientific concept early in the 20th century by a Swedish economist (though resting, of course, on very ancient ideas) and since then elaborated upon by demographers and economists in many countries, is based on the notion of the happy medium. Simply put, it may be said that, since states of overpopulation and underpopulation both exist, there must also be an intermediary position corresponding to an "optimum." In other words, in any given area or country at a given time, there must be a certain population size that provides the best welfare for the nation or people, as measured by some established goal. This level of population is somewhere between overpopulation and underpopulation.

The criteria of optimum population

The real problem begins in trying to apply criteria of social objectives. Without going into the evidence of economic marginal analysis and mathematical models, it is apparent, for instance, that a population size that ensures the highest per capita production (that is, the highest average standard of living) is lower than a population size that ensures the highest collective power. This explains why dictators or authoritarian governments, more concerned with power than with standards of living, are generally more pronatalist than are democracies, and it explains why in times of war or even of heavy collective expenditures, a country finds it in its interest to have a larger population than in normal times.

Also involved in a discussion of optimum population are concepts of a "minimal population" and a "maximal population." A minimal population refers to that number below which a group in a given environment cannot go and still survive in the long run. A minimum number of people is required for division of labour, a sustaining number of marital partners, and so forth. A maximal population refers to that number above which a group in a given environment cannot go. Its standard of living would be the lowest compatible with human life; the addition of further inhabitants would consequently result in a higher deathrate. In both minimal and maximal pop-

ulations the people live at a bare subsistence level, which obviously is not conducive to the enhancement of either individual welfare or national power.

Criticisms of optimum population

The theory of optimum population has been criticized, usually on the following counts:

(1) The theory is too static and could scarcely apply to any but a traditional agricultural population.

(2) An optimum population is difficult to determine; no scientifically achieved results are available. Historical examples, though, might possibly be adduced: after the Black Death in the 14th century, for instance, the inhabitants of certain regions in Great Britain and France enjoyed an improved standard of living and even a certain degree of emancipation from the authority of their overlords. This may suggest that the population before the plague had been higher than the economic optimum.

(3) The political application is unclear, even assuming precise and reliable assessments of what is optimum. If a government of a particular country of 30,000,000 inhabitants were to determine that its optimum population should be only 20,000,000, how does it go about reducing the figure by the necessary 10 million? Similarily, if the optimum assessment were 40 million, how could it reach this figure in a reasonable time even with the help of immigration? In both cases a time factor intervenes; theories involving a more dynamic approach seem, therefore, preferable.

The effect of economic and demographic investments

Population growth and demographic investments. When a country's population remains constant, an efficient distribution of economic investments should raise the national income by about 4.5 percent a year on the average (though varying from 7 percent down to zero from year to year); such a growth is normally attained in many countries. This growth rate, however, is no longer the best when the population increases or decreases. In such an event "demographic investments" must be taken into account—investments, that is, which keep the fewer or greater number of people at least at the same standard of living as before. When the population increases, for instance, new arable lands must be brought under cultivation, factories and housing must be built as well as new hospitals, schools, and so forth. These demographic investments are approximately proportional to—and sometimes even greater than—the increases in population. For housing alone, an increase of 1 percent in the population might absorb, in one year, as much as 1.5 or 2 percent of the national income. In general this type of analysis was well summarized in a 1954 statement made by the UN Secretariat:

> To sum up, population enters into the problem of achieving satisfactory standards of living in the underdeveloped countries in three principal ways. . . . First, their high birth rates create a heavy load of dependent children per adult. This makes it difficult to save enough over and above what is required for the support of the workers and their dependants, for needed investments in equipment for economic development. It also seriously complicates the problem of providing the children with the education that is essential for social and economic progress in the long run. This aspect of the population problem is common to all the under developed countries. . . . Second, falling death rates with high birth rates bring about a rapid increase of population. Large investments must be made to keep the growing numbers of workers equipped even with the same inadequate amounts of working equipment per man as they have had in the past. So the possibilities for investments which would improve the equipment and raise productivity per worker are diminished. This speeding up of population growth, aggravating the shortage of capital, is now taking place in very many underdeveloped countries, wherever successful public health campaigns have greatly reduced death rates. . . . Third, many of the under-developed countries have an excessive density of agricultural population in relation to the area of cultivated land. The average farmer has too little land to make a satisfactory living for himself and his family. Not all underdeveloped countries face this difficulty, but it exists in some which have large amounts of unused land, as well as in those where nearly all the cultivable land is fully occupied. At least some of the countries now suffering from acute agricultural over-population might be able to employ all their numbers to good advantage, and benefit in the long run from a substantially larger population, if they were better equipped to utilize the land resources which they possess, or if they could industrialize. But neither the necessary improvements of agriculture nor the development of industry can easily be accomplished, and the difficulties are increased by high birth rates and rapid population growth. . . . It appears that even in a country where population growth would be economically advantageous in the long run, economic progress will be hindered if the birth rate is so high and if the population grows so rapidly as to put an excessive strain on the economy. (From UN Dept. of Social Affairs, "The Determinants and Consequences of Population Trends," *Population Studies*, No. 17 [1953], and "Population Growth and the Standard of Living in Under-developed Countries," *Population Studies*, No. 20 [1954].)

The question of adequate food production

As reasonable as these predictions may sound, however, there are demographers and economists who insist that they have not been entirely borne out by events in the years since 1954. These scholars note that, although available capital has been far less than had been considered indispensable and although the rate of population growth has exceeded all forecasts, the average standard of living has, generally speaking, not fallen; it has in fact grown higher in most areas of living. The notable and, admittedly, vital exception involves the production of food, which has not kept pace with the population increase—falling so far short of doing so, particularly with respect to high-protein food, that some of the more pessimistic forecasters have begun to see the beginning of wide-spread, chronic famine sometime in the future, their estimates ranging from some 5 to 30 years from now. Optimists, on the other hand, are pointing to an acceleration of an already visible "green revolution," the development of higher yield, higher nutrient food that can accommodate larger populations.

In other areas of production, the pessimistic forecasts have not all proven true, and this divergence between theory and practice may be explained in several ways: (1) The theory rests on the assumption that the productive output of an additional inhabitant is nil without supplementary capital. This is not always true. Two individuals, working with the same initial capital, may frequently be more productive than one. (2) An increasing population seems to adapt more easily to various changes in social structure (such as geographic or professional distributions) that are required by technological progress and economic evolution. (3) Certain social conditions seem frequently to benefit from a certain population growth. A nation, for instance, whose deathrate is declining but which continues to produce a young population avoids the economic drawbacks of an aging population. Because such social factors as these are neither easily submitted to experimental study nor easily adapted to individual cases, they are all too often neglected by economists and demographers who elaborate theories of growth.

Optimum rate of population growth. In noting that many demographers believe that a steady increase in population can be more advantageous than a constancy or decline in population, one must further note that such demographers do put a limit on the rate of increase. A runaway population can drain all or most of a national income into demographic investments. Certain demographic investments can sometimes increase even faster than the population, as noted earlier with reference to housing—an annual increase of 1 percent in population demanding in some cases as much as a 1.5 or 2 percent increase in housing investments. As a result there has come to be formulated what is called an "optimum rate of growth" in population—that is, a rate ensuring the highest annual increase in the standard of living.

Though perhaps more easily determined than a static "optimum population," the optimum rate of growth is still difficult to pinpoint scientifically and it has critics among those who are convinced of the desirability of a stationary population. In any case, the figure most often cited, one reflecting the experience of western Europe, is an annual increase of 0.5 to 1 percent in population—an increase that seems to afford the best economic advantages at present, subject as it is, though, to future technological, social, or political changes.

The rate of increase of the population—including the rate of its short-term variation—may of course depend largely upon special circumstances: when national resources, for instance, are abundant and easily exploited, the increase in population is more readily accommodated. Similarly, some scholars believe that population increases are not as easily accommodated in developed countries where there is a high accumulation of capital per individual—the reason being that each new addition to the population, each new worker, must himself be afforded a very high measure of added investment in order to keep him up with his fellows and to avoid drawing everyone down. Scholars with this point of view thus consider that a less developed country can handle higher rates of population growth because it depends less on capital and more on labour. The argument against this point of view might appear to be that wealthy countries tend to save more; they have greater surpluses of income, which can be applied to additional units of population. In any case, one can readily see that the complex nature of the problems involved and the difficulties inherent in trying to isolate the population factor entail considerable uncertainties.

Composition of the labour force. In economics the term labour force denotes all those members of a population who are at work or are available for work; it embraces workers of all grades and occupations, and it extends to both those who are unemployed and those who are underemployed (that is, not employed to their fullest normal capacity).

Of primary concern are those in the labour force who are not sufficiently employed. Generally speaking, the less developed countries are marked chiefly by underemployment, usually resulting from the demographic condition of overpopulation in rural areas. Industrial countries, on the other hand, are often characterized chiefly by unemployment, resulting largely from faulty economic or social adaptations and not from overpopulation. All countries, however, industrialized or not, are subject to inconsistencies between their effective labour forces and their "desired labour population," which is the distribution of workers by occupational categories that is considered appropriate for a country if it is to respond fully to the demand for its different products and services. For various reasons—technological or social changes, inadequate development, misdirected manpower training, or whatever—there may be extended periods of unemployment in certain professions and a lack of manpower in others. The desired population in certain professional categories, in other words, is not reached and can only be corrected by such measures as training or retraining, migrations, or work connected with increased foreign trade.

The issue of a "desired labour population"

Outside the membership of the labour force are, among others, the very young and the very old—that is, those in the "inactive ages." The dividing line between the "active age" and the "inactive ages" naturally varies among cultures and can remain constant or vary even within a single society, depending on the influences of social and economic change. Social security plans are, of course, dependent on assessments of the inactive ages. In any case, a manual worker's economic value usually begins to decline at about the age of 40 to 45, and that is precisely the age over which most countries refuse immigrants, unless they are possessed of adequate capital.

Modern governmental policies. *Policies regarding fertility.* A pronatalist policy—one designed to increase the birthrate—may call upon many different means: the encouragement of marriages, family allowances, various methods of helping the family (schooling, health, housing, and so forth), tax abatements based upon the number of children, and propaganda. The effectiveness of such measures varies. For example, an intensive program of housing construction or a general raise in salaries may have no effect on the birthrate; whereas allocation of housing to a family upon the birth of a third child or family allowances granted to the second and third child seem to wield a far greater influence.

In modern times most such policies derived from demographic trends alluded to earlier. During the late 1920s and the 1930s, when birthrates were falling rapidly in Europe and North America, there was much alarm about "the retreat from parenthood"—reinforced in some countries by considerations of national security. Some economists, moreover, viewed the slackening of population growth as tending to cause economic stagnation and chronic unemployment. It was, in any case, apparent that demographic changes were increasing the proportion of aged persons in the population.

These considerations led nations to public policies that promoted the maintenance and growth of population. In France, concern about population trends, in conjunction with other interests, led in 1939 to a comprehensive system of family allowances and related measures, called the *Code de la Famille* (revised in 1946), whereby the government subsidized the cost of additional children in many French families. Some programs of financial aid to families with children have been adopted in most European and many non-European countries, including Canada. These programs are usually designed to serve various ends, and it is often difficult to say to what extent they express population policies. Moreover, most governments have been reluctant to authorize expenditures on a scale comparable to that accepted in France. It may also be noted that laws against the promotion or sale of contraceptives, as well as against abortion, are supported in some countries by demographic as well as religious considerations, and that such laws have the effect of pronatalism.

Before World War II, several countries had permitted complete freedom in the use of contraceptives (the Soviet Union had even authorized abortion), but none had aimed at the restriction of the birthrate as a check on population growth. After the war, Japan launched the first birth-control policy of modern times (in the Eugenic Protection Act of 1948), based upon contraception and abortion. From that date onward, a considerable number of countries adopted policies aimed at extending family planning and reducing the birthrate. By the beginning of 1970, for instance, 22 developing countries had instituted a national policy to promote planned parenthood—countries as diverse as India, Pakistan, Sri Lanka, Malaysia, South Korea, Taiwan, Turkey, Egypt, Tunisia, Morocco, Mauritius, Kenya, Honduras, Puerto Rico, and Chile. Their plans were administered either directly through public agencies or through public assistance to private agencies. And in addition to these countries with official programs, many other countries granted approval to privately sponsored programs of planned parenthood. Furthermore, the General Assembly of the United Nations in 1962 adopted a resolution calling attention to the importance of population problems, and the UN Economic and Social Council in 1965 voted to increase "the amount of technical assistance in population fields available to governments of developing countries upon their request." The World Health Organization in 1965 also gave limited approval to this principle.

Birth-control policies

The results of official national programs—notably, for instance, in India—were at first discouraging. The obstacles to changes in personal behaviour in a vast illiterate population dispersed in small villages are enormous. Subsequent events have provided grounds for greater optimism. Methods of organization and communication have been improved on the basis of early experimentation. Appropriation for such work was increased in India from 1,500,000 rupees in the first five-year plan (1951–56) to 265,000,000 in the third and 300,000,000 in the fourth plan (1966–71). Equally intensive programs have been inaugurated in some other countries.

Such countries as Sweden, Denmark, and the United Kingdom, and some states of the United States have legalized abortion under regulated conditions. The Soviet Union and most of the eastern European nations have legalized abortion and promote contraception, even though they have maintained that such measures are intended only for the welfare of individual families and are not influenced by economic considerations or a desire for population control. Moreover, the Communist regime in China has lately encouraged delayed marriage and the control

of births within marriage as a factor in achieving its economic objectives.

Leading Roman Catholic theologians have also recognized the serious economic and social consequences of the persistence of high fertility, in association with declining mortality, in nations with limited resources. Accepting the regulation of fertility as a family goal, they continue to insist that the means of achieving this goal must conform to approved moral principles (see also BIRTH CONTROL).

Policies regarding mortality. Although in principle all national policies aim at reducing the deathrate to the lowest point possible, in no country are the means adequate, for there is always a compromise more or less candidly expressed (if nowhere else than in the budget of expenditures), between efforts to achieve the highest possible degree of longevity and other aims. Even more thorny questions can arise, however, regarding the desirability of therapeutic abortions, euthanasia, and sterilization, for example.

Policies regarding migration. Although an immigration policy may have a demographic purpose, when it is a matter of peopling a sparsely populated territory (as in Australia and New Zealand) or when cultural or political reasons encourage it (as in Israel), such a policy more often has economic aims and acts selectively, if not exclusively, in giving priority to workers in certain professions or to certain nationalities.

The main host countries of immigration after World War II were Australia, New Zealand, the U.S., Canada, Argentina, Brazil, Venezuela, and Uruguay. Highest rates of net inflow existed in Australia and New Zealand; immigration was responsible for two-fifths of the increase of population in Australia. Migrants in the Americas were absorbed most readily by Canada and Venezuela, though in absolute numbers the U.S. still took by far the largest total. Jamaica lost nearly a quarter of its natural increase through emigration, while Mexico had a net outflow of 250,000 (mainly to the U.S.), which however, was only 3 percent of the natural increase.

One of the striking developments after 1945 was the influx of European and Indian populations into certain parts of Africa. In the Federation of Rhodesia and Nyasaland (later Rhodesia, Zambia, and Malawi) the number of Europeans increased from 105,000 to 274,000 between 1946 and 1957, and more than two-thirds of this was due to immigration. A similar process was at work in the Belgian Congo (later Zaire) and in Kenya, Tanganyika (Tanzania), and Uganda. Except in white-dominated countries such as Rhodesia and South Africa, however, such immigration tended to decline after the country had achieved independence.

Australia's "white Australia" policy, officially termed the "restrictive immigration policy," though somewhat liberalized after 1965 to permit the entry of some non-Europeans, remains perhaps the most noted current example of the exclusion of coloured peoples. In a world in which the volume and direction of migration have been heavily influenced by restrictions unfavourable to coloured peoples, however, some channels remained open, as shown by the immigration of Puerto Ricans to the U.S., and, until passage of the restrictive Commonwealth Immigration Bill in 1968, of West Indians to the U.K. In these cases the fact of common citizenship made possible a relatively large increase in the movement of coloured migrants into these countries.

Other policies. Although people often fail to recognize a "population policy" in their own country (other than the obvious one involved in restrictions on migration or in the stimulation or restraint of the birthrate), all countries are in fact governed by a great number of economic, social, fiscal, and tariff measures that together may exert a quantitative or qualitative influence upon the growth and movement of population, without necessarily having such a purpose. Without enumerating their effects, the possible influence of the following may be significant: (1) agricultural policies, involving conservation, land tenure, land allotment, and the like; (2) national defense policies involving inequalities with respect to the draft or the distribution of recruits by race or national origin; (3) general economic policies, regarding such things as full employment, economic development, housing, and the like; (4) educational policies, involving particularly the teaching of children of foreign immigrants; (5) general welfare policies, involving social security, social and health services, naturalization, delinquency control, and the like; and (6) labour policies, particularly those involving occupational training and the recruitment of temporary labour of foreign origin. Such a sampling suggests that the whole domestic policy of a country can be considered, from one perspective, a population policy.

III. Future prospects of population

POPULATION PROJECTIONS

Population growth in the future

In 1969, the UN extended its projections of world population to the year 2050, on the basis of three estimates, "low," "medium," and "high"—the medium being deemed the most likely. The results are indicated in Table 7.

Table 7: World Population Projections

year	"low" variant (000,000)	"medium" variant (000,000)	"high" variant (000,000)
1950	2,500	2,500	2,500
2000	5,400	6,100	7,000
2050	8,500	11,000	13,500

Source: *World Population Situation*, Note by the Secretary General of the UN November 1969.

For the first time the tempo of growth was seen to slow down; in the medium variant, for instance, the increase decelerates in half a century from 144 percent in 1950–2000 to 80 percent in 2000–2050. Projections for the future have been carried even further: there are estimates that the growth of world population could come to a stop in the year 2050 (a low estimate) or 2100 (a medium estimate) or 2200 (a high estimate). On the basis of such hypotheses (uncertain ones, to be sure), the number of inhabitants on the planet would level off and become stable at 8,500,000,000 in 2050 or 13,500,000,000 in 2100 or 30,000,000,000 in 2150. Whatever the nature of the hypothesis, however, growth will be uneven among the developed and less developed nations. Whereas present growth ratios are roughly 2:1, in the year 2000 they would become 3:1.

Table 8: Population Projections in Developed and Less Developed Nations

	1970	2000
Developed regions	1,082	1,441
Less developed regions	2,510	4,688
Total	3,592	6,129

The aging process in the future

With regard to age distributions, in 1980 the proportion of individuals *over* 65 years of age should still be only 4 percent in the less developed regions, as against 10 percent in the developed regions (13 percent in Europe). On the other hand, the proportion of individuals *under* 15 should be 39 percent in less developed countries, as against 25 percent in developed countries (23 percent in Europe). After 1980, and particularly during the course of the 21st century, the number of older people will grow, because the great numbers of young people in the less developed countries today will have a greater life expectancy. Whether the percentage of older persons increases depends upon the birthrate—that is, upon whether or not the number of young people increases as rapidly as the number of old. The likelihood is that there will be a significant aging process; the birthrate—and thus the number of young people—will go down. All the countries of the world will find themselves faced with larger numbers of old people whose productivity in modern economies is less desired. In the meantime, all countries and particularly the less developed countries will undergo

jolts and jars trying to adjust to the new distributions in ages. The course of population growth and distribution probably will be very chaotic in the course of the 21st century in most of the less developed countries.

The population implosion—the further concentration of peoples in urban areas—should continue unabated. In Table 9 the vast increase of the urban population of less developed regions is noted; this will have quadrupled in 40 years. The drop in the rural populations in the more developed regions is also observed.

Table 9: Urban-Rural Populations

		urban population	rural population	total
More developed regions	1960	502	352	854
	2000	1,010	280	1,440
Less developed regions	1960	490	1,647	2,137
	2000	2,030	2,740	4,675
World total	1960	992	1,999	2,991
	2000	3,090	3,020	6,110

PROBLEMS OF RAPID POPULATION GROWTH

The rapid increase of population in the less developed regions presents a far more serious problem than the population increase in 19th-century Europe for the following reasons: (1) the increase is far more rapid (2.5 to 3.5 percent instead of 1 percent); (2) it affects a far greater population; (3) it is not susceptible to any compensation in such migratory and economic outlets as Europe found in the New World and Australasia.

The increased speed in the spread of antimortality techniques has rapidly raised the question of how an ever greater number of human beings, multiplying without controls of any kind, are to be fed. Faced with this vast problem, observers separate into two camps: the pessimists, with their unfavourable and even tragic prognoses (such as general famine), and the optimists, confident that science and technology can overcome these difficulties. A second conflict of scholars, not far removed from the first, exists between the "economic developers" and the "birth controllers." The former group favours an absolute priority for economic and cultural development, which, they contend, cannot fail eventually to trigger a decrease in fertility. The latter group hopes to work for birth control at the family level to reduce national population burdens—burdens that they assume must be alleviated before economic development can successfully be undertaken.

A defensible conclusion with respect to the matter of economic priority, even from the strictly economic point of view is difficult to ascertain. The cost of contraceptives capable of preventing the birth of a child can be much lower than the cost of investments and consumer goods necessary for that child's economic life (housing, food, tools, etc.)—but only if the contraceptives are reasonably efficient. When their efficiency is virtually nil, as may be the case among indigent and ignorant populations, economic development yields greater profits.

The difficulties met by such populations in the use of contraceptives have led to the search for new methods, such as the more efficient oral contraceptive (the Pill) and intrauterine devices, but even these are not yet entirely satisfactory. What are needed, perhaps, are pills with longer term action and intrauterine devices less susceptible to expulsions and complications. Some countries have resorted to what are usually considered to be "radical" birth-control techniques including surgical sterilization (India and Puerto Rico) and abortion (Japan, Tunisia, and the Communist countries). What may yet prove to be a more efficient means of encouraging birth control, however, may involve instilling in adults a love and appreciation of children and persuading them that they can extend better care to a smaller number of children (see BIRTH CONTROL).

That birth control can succeed without new contraceptive practices is proved by the experience of some countries, particularly in the Far East, whose birthrates had declined even before the use of the Pill and intrauterine devices, and continue still to decline (see Table 10).

Table 10: Selected Countries—Births per 1,000 Inhabitants

	1956	1970
Singapore	44.4	23.0
Hong Kong	37.0	20.1
Taiwan (Formosa)	44.8	28.1
Mauritius	43.3	27.3

The alternative to a direct program of birth control—that is, a program giving priority to economic development—also involves frustrations. Even in a very sparsely populated country richly endowed with natural resources (Brazil, for example, or Zaire), overrapid growth of the population brings in its train an excessive need for demographic investments—that is, investments just to keep pace with the extra mouths to feed and bodies to clothe and shelter. Such quantitative investments reduce the availability of qualitative investments intended to raise the standard of living. Although it is always difficult to measure optimum growth, that point in most of the less developed countries seems to have already been passed (the figure of 2 percent per year might well be taken as an upper limit).

International political debates

International discussion of population growth has been marked by wavering opinions on the necessity for measures on birth control. In February 1947, at the first meeting of the United Nations Population Commission in New York, the representative of the Soviet Union strongly opposed any restriction on the birthrate, thus opposing the U.S. The representatives of Roman Catholic countries similarly opposed those of the Protestant countries. At this time and for many years to come, the less developed countries—with their age-old traditions regarding fertility and with their new thirst for power and prestige—also generally opposed family planning. By 1962, however, after some 15 years of debates, the position had changed radically. A vote in the UN General Assembly in that year somewhat surprisingly split the nations into roughly three equal groups—those for birth control, those against it, and the remainder abstaining from opinion. In favour of a restriction of births were many African and Asian countries (particularly the Arab States), a new persuasion that prompted a cautious abstention from the Soviet Union. Since 1962 the tide of nations favouring the limitation of births has steadily swelled, particularly in Asia.

The continuing diversity in votes and opinions reflects, in part obviously, the great diversity among countries with respect to deathrates, birthrates, and levels of economic development or potential. Purely global studies of population have therefore become almost pointless; future studies must bear increasingly on regions and even on individual countries. India, for instance, faces almost terrifying population pressures as related to area, intense poverty, and density of population (436 inhabitants per square mile). In 1965 a threatened famine, following on two poor harvests, was narrowly avoided. Birth control has thus become a matter of national public welfare. In sub-Saharan Africa, on the other hand, the problem of the immediate future is far less urgent although malnutrition is at times dramatic enough. In Latin America the demographic and economic characteristics vary widely from country to country—with a relatively fine show of achievement in a country like Mexico, as contrasted with the increasing poverty of ever-larger numbers in a country like Haiti. Everywhere in Latin America, however, the growth of the population is very rapid and tends to aggravate the problems of political stability. Generally speaking, whether or not predictions of the danger of a sudden general worldwide famine are accepted, serious local problems do remain unsolved, problems quite capable of provoking political complications.

BIBLIOGRAPHY. E.E. ARRIAGA, *New Life Tables for Latin American Populations in the Nineteenth and Twentieth Centuries* (1968), historical life tables for 17 Latin American countries, computed by a new method based on the stable

population theory; B. BENJAMIN, *Social and Economic Factors Affecting Mortality* (1965), a study of mortality in terms of employment, social status, housing, nutrition, marital status, education, and culture, and *Demographic Analysis* (1969), a short manual explaining demographic methods to the general reader and showing the relationship between population trends and social change; B. BERELSON *et al.* (eds.), *Family Planning and Population Programs* (1966), a comprehensive assessment of family planning programs throughout the world; G.H. BEYER (ed.), *The Urban Explosion in Latin America* (1967), a series of articles on social, demographic, political, economic, historical, and other important aspects of urbanization; D.J. BOGUE, *Principles of Demography* (1969), a comparative work stressing demography as a unique systematic discipline, extensive bibliog. follows each section; W. BRASS *et al., The Demography of Tropical Africa* (1968), a demographic discussion concentrating on case studies for the Republic of the Congo, French-language and Portuguese territories, the Sudan, and Nigeria; J.C. CALDWELL, *Population Growth and Family Change in Africa* (1968), an analysis of the urban elite in Ghana, their population characteristics, and reproductive behaviour; A.M. CARR-SAUNDERS, *World Population: Past Growth and Present Trends* (1936), a world population history during the modern era up to the early 1930s, particularly in Europe and "Europe Overseas," with a detailed analysis of migration; M. CEPEDE, F. HOUTART, and L. GROND, *Population and Food* (1964), a discussion of population growth and food supply, life expectancy, standards of living, aid to underdeveloped countries, and the need for family planning to help provide enough food for everyone; C. CLARK, *Population Growth and Land Use* (1967), an analysis of selected problems of population growth; W.W. COCHRANE, *The World Food Problem: A Guardedly Optimistic View* (1969), a review of the world food problem in historical perspective, with attention to social, political, economic, and demographic factors affecting the problem; O.A. COLLVER, *Birth Rates in Latin America: New Estimates of Historical Trends and Fluctuations* (1965), an historical study of fertility trends in Latin America, based on data from censuses and vital statistics; R.C. COOK and J. LECHT, *People: An Introduction to the Study of Population* (1968), an introduction for the general reader to the problems of world population growth; P.R. COX, *Demography*, 3rd ed. (1959), an account of principles and methods in the study of population, including the collection, analysis, and use of data; R.F. DASMANN, *Environmental Conservation* (1968), a study of man's impact on nature in the U.S.; H.I. ELDRIDGE, *The Materials of Demography: A Selected and Annotated Bibliography* (1959), a list of significant works in the area of population analysis, with special emphasis on works suitable for teachers, researchers, and students of demography; R. FREEDMAN (ed.), *Population: The Vital Revolution* (1964), authoritative essays on major aspects of contemporary population trends; N. FROTHINGHAM, *Population Concerns and Family Planning in Africa, South of the Sahara* (1968), a presentation of the current status of demography in sub-Saharan Africa; H.K. GEIGER, *The Family in Soviet Russia* (1968), a detailed view of the changing Russian family; P.M. HAUSER, "World Population Problems," *Headline Series No. 174* (1965), a description of the problems of population growth and the benefits of controlling this growth; D.M. HEER, *Society and Population* (1968), a brief text limited to the consideration of population of nation-states and of the world as a whole; G.W. JONES, *The Economic Effect of Declining Fertility in Less Developed Countries* (1969), a study of the implications of different population trends for economic development in low-income countries; N. KEYFITZ and W. FLIEGER, *World Population: An Analysis of Vital Data* (1968), demographic statistics for different countries over varying periods of time; S.S. KUZNETS, *Economic Growth and Structure: Selected Essays* (1965), comparative essays on geographic areas and their economic growth, with a chapter on population change; S. MUDD (ed.), *The Population Crisis and the Use of World Resources* (1964), a collection of articles documenting pressures of population growth on natural resources; NATIONAL BUREAU OF ECONOMIC RESEARCH, *Demographic and Economic Change in Developed Countries* (1960), a collection of papers on the causes of economic and demographic change in developed countries; PAN AMERICAN HEALTH ORGANIZATION, *Migration of Health Personnel, Scientists, and Engineers from Latin America* (1966), a study of migration and the forces involved, with consideration of how trained personnel can be induced to remain in their own countries; A. SAUVY, *Fertility and Survival: Population Problems from Malthus to Mao Tse-Tung* (1961), an analysis of some theories of population, world population problems, and proposed solutions, and *General Theory of Population* (1969), a study of the relationship between the demographic-biological characteristics of societies and their economic and social circumstances; J.M. STYCOS, *Human Fertility in Latin America: Sociological Perspectives* (1968), a discussion of fertility in Latin America based on empirical studies considering such topics as church and politics, attitudes toward family planning and family size, employment of females, education, culture, urbanization and other factors affecting fertility; WORLD HEALTH ORGANIZATION, "Trends in the Study of Morbidity and Mortality," *Public Health Papers 27* (1965), a discussion of the study of morbidity and mortality in developed and developing countries; D.H. WRONG, *Population and Society*, 3rd ed. (1967), a brief introduction to the main aspects of the population dilemma.

UN Publications: *Demographic Yearbook* (annual), basic demographic statistics for over 200 countries and territories assembled by the Statistical Office of the UN; *Statistical Yearbook* (annual), statistical data for over 200 countries and territories covering a wide range of economic and social subjects; "The Situation and Recent Trends of Mortality in the World," *Population Bulletin No. 6* (1963), a detailed statistical analysis of global trends in mortality.

Periodical literature: *Milbank Memorial Fund Quarterly*, a journal of scholarly papers analyzing demographic trends and family planning programs; *Population Index* (quarterly), annotated listings of publications in demography, fertility control, and many related fields; *Population* (bimonthly), a journal published in French with English and Spanish summaries, includes scientific articles on all aspects of population dynamics; *Population Bulletin* (bimonthly), issues and interprets well-documented facts about population trends in the U.S. and abroad, with discussions of the consequences for governments and individuals; and *Population Studies* (3/yr.), articles by demographic professionals from all over the world.

Population, Biological

A population of animals is the number of individuals of a species living in a restricted area. The term restricted area raises difficulties because all the individuals of a species, or at least of a subspecies, may intermingle to some extent. In a small area, even one with natural boundaries such as a forest or a pond, there is usually some movement of animals in and out. Perhaps it is only on nearly inaccessible islands that a biologist can hope to study a complete, self-contained population of large animals, and this has rarely been done. He may, of course, enclose a self-contained population of small animals in his laboratory, but this seldom replicates a natural situation.

One may also speak of populations of plants. In a plant population, however, there are often great variations in individual size, so that the total number of plants present may be less significant than their combined weight or "biomass." There is a further difficulty in counting the numbers of a plant that reproduces asexually (vegetatively) and in which the new shoots remain joined to the original plant. One bramble bush (*Rubus*) may extend for several miles, and the dog's mercury (*Mercurialis perennis*) covering the floor of a whole wood may be, effectively, one individual.

Animal and plant populations tend to be stable—that is, their fluctuations are small compared with what would be theoretically possible if no limits were placed upon their increase. Only a relatively small number of species are either expanding in numbers or decreasing to extinction. Stability is possible only if numbers tend to rise when low and to fall when high, through the action of density-dependent factors. This article attempts to summarize what is known about the critical factors concerned.

Human populations are currently unstable. Primitive tribes were perhaps limited by the density-dependent effects of starvation, disease, and fighting. In the last century or two, however, human populations have not been effectively regulated but have been increasing to an unprecedented extent. This has come about because of improved food supplies and the eradication of diseases. But a population must eventually cease to increase. It can do so in one of two ways, either through higher mortality (from starvation, disease, or war) or lower natality. While a further increase in food supplies may be desirable on humanitarian grounds, it can at best provide only temporary alleviation.

This article is divided into the following sections and subsections:

I. The study of populations
Reasons for studying populations
Species studied
II. The characteristics of biological populations
Age, sex, and genetic differences
Age distribution
Sex distribution
Genetic differences
Numbers and density
Natality
Mortality
III. Changes in population characteristics
Growth of populations
Fluctuations in stable populations
Seasonal fluctuations in numbers
Irregular annual fluctuations
Density-dependent natality
Density-dependent mortality
Predator–prey oscillations
Movements
Migration
Emigration
Dispersion
Dispersal
IV. Interactions of populations
Interactions between members of a population
Cooperative behaviour
Competitive behaviour
Interactions between populations of different species
Competition among species
Negative direct interactions
Positive direct interactions

I. The study of populations

REASONS FOR STUDYING POPULATIONS

Perhaps the most important reason for studying animal populations is scientific curiosity. Animals are able to reproduce so rapidly that any species might conceivably overpopulate the world in a short space of time, but in fact this does not happen. All species remain far below the numbers that their reproductive rates would allow, and the reasons for this provide a scientific challenge. Linked with this problem are fascinating subsidiary questions, such as why some species remain scarce for years and then suddenly increase explosively, why a few species show regular fluctuations in their numbers, and why some species are always scarcer than others. These questions lead to such further questions as why some animals lay many more eggs, and some plants produce many more seeds, than others; why some species eat out their supplies while others do not; and why some species engage in mass movements while others do not.

Apart from its theoretical interest, population dynamics has important practical bearings. Entomologists have been concerned with reducing the numbers of various pests that destroy crops or commercial forests; often they do this by introducing natural enemies, for which knowledge is needed of how the numbers of a predatory species interact with those of its prey. Fisheries biologists have been concerned with raising the weight of fish brought to the market, for which knowledge is needed of the influence of fishing on the age distribution of the fish and of the complex interaction between their growth and reproduction rates and their food supplies. Conservationists seeking to preserve large herbivorous mammals have been concerned with the problem of whether or not, or at what rate, to destroy natural enemies.

Complex relationships involved in population study

It cannot be stressed too strongly that the answers to practical problems of this sort are often complex, and that detailed study may be needed if effective practical measures are to be developed. This is largely because an action effective for a short time may prove disastrous later on. Tourists in the game parks of East Africa have complained to authorities in the dry season that young lions (one of the main attractions) are starving and that the wardens ought to feed them. They fail to see that if the young lions were fed, more would be present to starve in subsequent years. In the early days of conservation in North America and South Africa, large carnivores were killed in order to increase the numbers of deer or antelope, only to have the latter outrun their food supplies and decline catastrophically through starvation. Again, the British government for many years paid a bounty to encourage farmers to shoot wood pigeons (*Columba palumbus*), which are a pest of agriculture, but this measure was ineffective because, although many birds were killed, shooting did not bring numbers below the limit set by starvation in late winter. In the early development of a fishing industry, the weight of fish brought to market increases with the efforts of the fishermen, but beyond a certain point the catches drastically decline because the fish can no longer replace their losses. In Fiji, a native leaf-mining beetle (*Promecotheca reichei*), which eats coconut leaves, did no economic harm because its numbers were held down by a native insect parasite. But a predatory mite, *Pediculoides ventricosus*, was accidentally introduced. This mite also preys on the beetle, destroying virtually all the larvae at one time but leaving the adults untouched. At this stage, both the mites and the native parasites find no further prey, so that they die in large numbers; the next generation of beetle larvae is virtually untouched and does serious harm to the coconuts. Hence, under some circumstances, the introduction of an enemy may greatly increase, rather than reduce, the numbers of its prey. The problem in Fiji was solved by the introduction of another species of parasitic insect. To give a final example, the scale insect *Icerya purchasi* did enormous damage to the California citrus orchards until its enemy, a ladybird beetle, Vedalia (Rodolia) *cardinalis*, was deliberately introduced, after which, for many years, the scale insect caused no appreciable harm. But the use of chemical insecticides against mosquitoes, near the citrus orchards, once again made the scale insect a pest, because, though the poison affected the scale insects, it had a much greater effect on the ladybugs.

The interactions between different species and the factors controlling their numbers are often subtle; immense care must be taken before introducing new animals or plants if the balance of nature is not to be disturbed in unexpected ways, many of which may be disadvantageous to man. Because such interactions, though vital, are often inconspicuous, the factors controlling numbers in existing situations may be hard to discover. The common European Saint-John's-wort, for example, was accidentally introduced to California, where it made huge areas of grass pastures unsuitable for livestock, to which it is harmful. The numbers of the plant were later controlled by the introduction of a small beetle (*Chrysolina quadrigemina*), and both plant and insect are now so scarce in California that, without knowledge of the very recent past, it would be extremely difficult to discover that the one controls the other. Again, it has been shown, through the recovery of marked songbirds, that about half the adults die each year; yet this huge mortality passes unnoticed, mainly because the dead bodies are so quickly removed by scavenging animals.

The human population problem

The fundamental principles of population dynamics apply as much to man as to other animals. Some economists have argued, for example, that the human population problem could be solved by bringing new areas, such as the Amazon forests or the seabed, into cultivation or by further improving the genetic strains of crops; but the study of animal populations makes it reasonably certain that the mere provision of extra food would lead to a further increase in human numbers unless other things were done. Even the simple point that, in a stable population, the birthrate and the death rate must be equal, tends to be overlooked with respect to mankind: efforts to reduce the death rate tend to ignore the fact that for stability it will be necessary to reduce the birthrate. It should be added, of course, that the population problem is more complex in man than in other animals, principally because of the interplay of biological and ethical principles. The populations of small songbirds are stable because death rates equal birth rates, and about half the adults die each year; but such a high mortality among humans would be considered appalling.

SPECIES STUDIED

Each species has advantages and disadvantages as a subject of population research. Most birds are diurnal; they

can be readily watched, whole populations can be marked individually, and nests can be found with comparative ease. On the other hand, dead birds vanish quickly, non-breeding adults are hard to detect, and individuals moving outside a study area may go unnoticed. The often-studied small rodents are very abundant, but they are mainly nocturnal, and it is hard to catch whole populations. Big mammals present great advantages in that they can more easily be caught for marking. Fish, like small rodents, cannot be observed, but large samples can be caught and marked, and like various large mammals, their ages can be determined by structural characteristics; further, much money is available for fisheries research. The chief advantage in the study of insects is that large samples can be killed without significantly affecting the population. In some instances, it is not necessary to kill the insects; *e.g.*, certain insect larvae, called parasitoids, feed on other insect tissues in such a manner that the latter is not killed until larval development is complete. In this case, the living larvae can be examined. Individual marking, however, is not normally practical, and there may be a stage in the life history that is hard to study. Almost any species of animal may, in fact, yield results of value to a skilled researcher: planarian flatworms have provided some of the best evidence for the limitation of numbers by food shortage.

A great practical difficulty in the study of natural populations has been that biologists tend to change their posts and move from one area to another. Most of the long-term field studies have been carried out at research institutes rather than by particular individuals, and chiefly on animals of economic importance.

II. The characteristics of biological populations

AGE, SEX, AND GENETIC DIFFERENCES

Populations consist of individuals of different ages, of two sexes, and of varying heredity.

Age distribution. In some animals, the young are similar to the adults except in size; they eat similar types of food and are subject to similar predators, so their numbers and those of the adults may interact with each other. But in other kinds, the young are extremely different in appearance, live in different habitats, eat different types of food, and are subject to different predators, so that adults and young have no direct influence on each other's survival.

In many small invertebrates, notably many insects, the eggs are laid in one short period of the year, hatch into larvae more or less synchronously, pupate synchronously, emerge as adults together, and die after breeding. Hence at any one time, almost all the population consists of animals of the same age, and the breeding stock of one year is derived wholly from eggs laid the previous year. The situation is slightly more complex in species that have a similar life history but raise two successive generations each year; or, alternatively, that take several years to raise one generation. Most sockeye salmon (*Oncorhynchus nerka*), for example, live for four years. Each individual breeds once and dies; its young move from freshwater to the sea to complete their growth and then return to breed. Thus the populations coming up the rivers of British Columbia in four successive years are nearly separate. The point has commercial importance because in some areas one of the annual populations in a four-year period is much larger than the other three, probably because the average chance of survival is greater for a young fish hatched in a peak year, when they are much more numerous—relative to the numbers of predators—than in other years. In the European eel (*Anguilla anguilla*), which also breeds only once in its life, the age of breeding is more variable; the young of this group leave English ponds for the sea at between five and 20 years of age, largely depending on how fast they have grown.

Age distribution in salmon populations

The age distribution (*i.e.*, the proportion of individuals in each age group) is much more complex in most other vertebrate animals and in some of the larger invertebrates. In these groups, most adults breed every year until they die, so that the breeding population at any one time consists of individuals of a mixture of ages. The age distribution in a bird population with a high annual mortality is shown in Figure 1. In certain other birds the proportion

From *Journal of Wildlife Management* (1940)

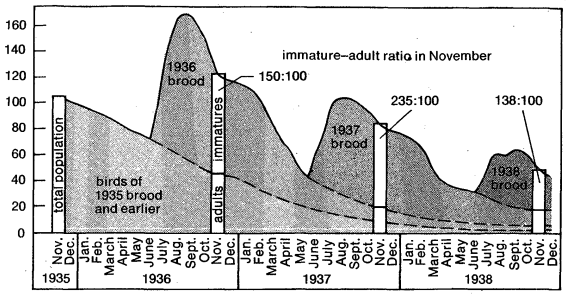


Figure 1: Age distribution of California quail (*Lophortyx californica*).

of younger individuals is much lower; there may be subadults, similar in appearance to adults, that do not breed. In most stable populations the age distribution remains relatively constant. But in certain fish and marine invertebrates, many more young are produced in some years than in others, and the predominance of a particular year class may persist for some years. A good example is the 1904 year class of herrings (*Clupea harengus*) in the North Sea that long outnumbered each of the later year classes, remaining predominant until 1919.

Various small plants, like various small animals, are annuals; they winter as seeds, breed once, and then die. Other species are biennials, and many more, including trees and bulbs, are perennials. Some large trees live much longer than any known animal—as long as several hundred or, in rare instances, 1,000 years.

Sex distribution. In most animals the male–female sex ratio is approximately equal. In various animals with an almost equal sex ratio at birth, one sex may become more common than the other later in life. In many monogamous birds, for instance, adult males are rather more common than adult females. This is probably because the females take the greater part, and in some species the whole part, in incubating and raising the young and thus are subject to heavier predation than the males. On the other hand, in polygynous birds (*i.e.*, those in which each male has several mates), breeding males are much scarcer than breeding females, perhaps in part because their exceptionally conspicuous plumage and behaviour patterns render them more subject to predation than the females. In one such species, the boat-tailed grackle (*Cassidix mexicanus*), the conspicuous males disappear faster than the less-conspicuous females in winter, perhaps through predation. Among humans, about 51 percent of all births are male, but in modern Western civilization, females survive slightly better than males, so that during adolescence the sex ratio becomes equal, and among adults there are slightly more women than men.

Variations in the sex ratio

In certain arthropods, notably aphids and water fleas, some generations consist solely of females that lay parthenogenetic eggs, which develop without fertilization by a male, though at other times of the year males are present and fertilize the eggs, which are of a different type. In another group of insects, the social hymenopterans (bees, wasps), most of the population consists of workers, which are sterile females. In various rotifers, only females are known. The sea worm (*Bonellia*) is first a microscopic male and later turns into a sizeable female, and a few other species also change their sex with age. Earthworms and snails, along with a few other animals, are hermaphrodites, with male and female organs on each individual. Hermaphroditism is, of course, common in flowering plants, both sets of organs usually being on the same flower. Asexual reproduction is common in plants, but in nearly all of them sexual reproduction is also the rule, and only a few species have dispensed with it entirely.

Genetic differences. In addition to the differences linked with sex, individuals of the same population differ somewhat in other genetic characteristics. These com-

prise the subject matter of population genetics, but it should be noted that the extent of genetic variation may vary with the state of the population; in particular, it may be unusually great in an expanding population because selection is temporarily lightened. When, for example, the fish *Bairdiella icistius* was introduced to the Salton Sea, California, it increased rapidly. In 1953 the population included many abnormal individuals, some blind, others with malformed jaws; a year later, when numbers had decreased, there were proportionately fewer abnormal individuals. Similarly, when the marsh fritillary butterfly (*Euphydryas aurinea*) greatly increased in the 1920s, for unknown reasons, in the north of England, there were unusual varieties in colour, pattern, and shape, but these had been absent when the population was small. Some researchers have argued that the regular population cycles discussed later in this article result from the production, in different years, of individuals of different genetic makeup; but there is no positive evidence for this, and it seems unlikely to be an important cause of cycles (though some genetic change is to be expected in populations that alternately increase and decrease in numbers on a big scale).

NUMBERS AND DENSITY

The numbers of large animals can often be counted directly, and in recent years aerial photography has been developed as a method for counting both big mammals and large birds. The numbers of various other animals have been assessed by capturing some of the population, marking and releasing them, and in a further series of captures finding the proportion of marked individuals in the new sample. Given large samples and random means of catching the animals, this "capture–recapture" method can provide accurate estimates of numbers. The results can be misleading, however, if the samples caught are biassed as to age or sex, particularly if this varies seasonally, and also if previous capture affects the chances of later capture—*e.g.*, if an animal learns to avoid being caught. In the case of some animals, including small but abundant species, total numbers are normally estimated from those counted in sample areas.

Estimates of numbers

Some animals are much more numerous than others. In general, large animals are sparser than small, because the larger the animal, the more food needed to sustain it; but there are so many exceptions that this generalization is of little use. Predators are much scarcer than their normal prey and are usually larger in size. Among the smallest existing populations are those of various land vertebrates on remote islands, though this is partly the result of destruction by man. Where man's destruction is checked, even a population of just over 20 individuals may persist, as shown by the conservation and recent small increase of the whooping crane (*Grus americana*) on the American mainland. The most numerous large animal in the world is man. It would be extremely hard to determine which is the most numerous small animal in the world.

The population density of a species typically varies in different parts of its range. Population biologists have tended to study animals where they are common—in the middle parts of their range and in optimal habitats—but their relation to their environment may be very different where they are sparse. In the few cases in which a population has been studied in two areas, it has been found that its density, natality, mortality, main predators, and even main foods may differ from one area to another. Most animal populations have inevitably been studied under conditions affected by man. This need not affect the validity of the general conclusions reached, but the proportionate importance of various factor influencing birthrates or death rates may be very different in man-modified and in natural habitats.

The equality between natality and mortality

Natality. The reproductive rate—in mammals the birthrate, in plants the rate of seed production—is the rate at which new individuals are produced. It should be distinguished from the recruitment rate, which is the rate at which new adults (not new eggs or new young) enter the breeding population. Most animal populations are not, on the average, either increasing or decreasing markedly, and in such populations the recruitment rate equals the adult mortality, while the natality or reproductive rate equals the total mortality of eggs, young, and adults.

Some biologists have argued that the equality between natality and mortality is an evolutionary mechanism to prevent overpopulation. This suggests that natality is smaller than it might otherwise be, particularly in long-lived animals. A different view holds that the implication of evolution by natural selection is that those hereditary types (genotypes) that leave most offspring must come to predominate over the rest. In other words, the reproductive rate of each species has been evolved through natural selection to correspond with that which gives rise to most surviving young per adult. This does not mean that each species must produce an immense number of eggs or young. In some this may, indeed, be the most efficient means of reproduction, but others are unable to form sufficient food reserves for many eggs, or cannot find food for a large family without serious risk to themselves, or need large food reserves for the young at hatching, and hence produce few (but often large) eggs or seeds. On this view, natality and mortality are roughly equal because mortality rises when the population is high.

Natality differs greatly in different kinds of animals. An oyster can probably lay 100,000,000 eggs during its life, and even so complex an animal as the cod *Gadus callarias* produces some 4,000,000. In contrast, an elephant matures late, has one offspring at a birth with an interval of several years between births, and may produce only five young in its life. Similarly, the number of seeds produced by a plant differs greatly in different species.

Birds. The average clutches of different species of birds vary from one to 15; the largest in a species that feeds its young is ten, in the blue tit *Parus caeruleus*. Clutch size is characteristic for each species and presumably depends on hereditary factors, although there are also phenotypic (nonhereditary) variations, the same individual laying clutches of different size under different conditions.

Egg numbers

In birds that feed their young, clutch size corresponds to the greatest number of young for which the parents can find enough food without serious risk to themselves. For instances, the European common swift *Apus apus* usually lays three eggs. If it is experimentally given four, then, as shown in Table 1, so many nestlings starve that each pair raises, on the average, fewer young than pairs starting with three. Hence, any tendency to lay a clutch of four will be eliminated by natural selection. Similarly, the most frequent clutch size in first broods of the European starling (*Sturnus vulgaris*) and the great tit (*Parus major*) corresponds with that brood size from which the most young survive; in these species, however, the young in larger broods do not starve in the nest, but leave it underweight and die soon afterward.

Table 1: Number of Young Raised by the European Common Swift Apus Apus

brood size	broods observed	number hatched	number fledged	proportion raised (percent)	number raised per brood
2	72	144	139	97	1.9
3	20	60	51	85	2.6
4	16	64	27	42	1.7

Source: Observations by C.M. Perrins at Oxford in the years 1958–61. Broods of four were not natural but were made up by adding newly hatched young from other broods, which does not disturb the parents. Some broods of two came from clutches of two, but others from clutches of three, either through failure of an egg to hatch or because a nestling was removed to make up a brood of four.

Many phenotypic variations in clutch size are adaptive. The Eurasian nutcracker (*Nucifraga caryocatactes*), which stores the food for its young the previous autumn, lays four after a good and three after a poor season for nuts. Various hawks and owls that prey on lemmings or voles lay unusually large clutches in the occasional years when their prey is extremely abundant. Regular seasonal variations in the clutch size of songbirds in north temper-

ate regions broadly fit the seasonal variations in the availability of their food. Individuals breeding for the first time lay slightly smaller clutches than older individuals and are also known to be rather less efficient at finding food. If such phenotypic adaptations were perfect, each species would always lay the clutch best suited to its circumstances, but this would require precise prediction of future feeding conditions for the young.

In hawks, storks, and other large birds, incubation starts with the first egg, so that the young hatch at one- or two-day intervals. This is an adaptation, for if food is sparse, the older nestlings are so much stronger than the younger that they get all of it, while if the food is plentiful, the larger nestlings are sated and the younger are fed; hence the brood size comes quickly to correspond with the feeding conditions at the time, and food is not wasted on young that would starve later. Another adaptation is the growth rate of the nestlings, for the same quantity of food per day can be used to raise a larger brood more slowly or a smaller one more quickly. This is probably why hole-nesting songbirds have larger clutches than those with open nests; the latter suffer such heavy predation that there must be strong selection for rapid growth, but fewer young can then be raised together. A slow growth rate may also be advantageous for species that find difficulty in bringing enough food for even one nestling, such as various large raptors and seabirds.

Another general tendency is for clutches to be larger at higher than at lower latitudes. At the higher latitudes, summer days are longer, so the parents can collect more food per day. Many species breed with the flush of insects after the winter dearth in which many adults have starved, so that food becomes plentiful in relation to adult number, and large broods can be fed. Similar conditions hold in tropical savannas, where birds breed with the rains after a long dry season. In tropical evergreen forests, however, food may be available in nearly the same quantities throughout the year; hence, the number of adult birds may stay close to the limit set by food, there being only enough extra food for small broods.

Although ducks, unlike most birds, do not feed their young, there is a broad inverse correlation between the number and size of the eggs in different species. This suggests that the female's food reserves may be used for more and smaller, or fewer but larger, eggs. Presumably, natural selection has favoured the evolution of a larger egg (and therefore a smaller clutch) in those species in which it is advantageous for the young to be larger, or to carry greater food reserves, at hatching. In a few other birds, including the common swift (*Apus apus*) and the red grouse (*Lagopus scoticus*), the clutch is slightly larger when more food is available for the laying female, but this is a very minor factor.

Natality is also affected by the number of broods raised. Most species raise one a year; many songbirds raise two, some even three, and the Senegal fire finch (*Lagonostricta senegalla*) raises four. At the other extreme, probably because their food is sparse, a few large raptors and seabirds take more than one year to raise a nestling; they breed successfully only once every two years.

Deferred maturity

The young of some finches and doves first breed when two to four months old; most birds when just under a year old; various water birds, raptors, and others not until several years old; and large albatrosses not until about the age of ten. Such "deferred maturity" (like a small clutch) has been seen as an adaptation to reduce natality in long-lived species. But it can as readily be explained in terms of natural selection, provided that those individuals that defer maturity ultimately leave more descendants than those that start breeding when younger. This will hold if (1) younger individuals are unlikely to raise young, and (2) the attempt to breed involves a greater chance of death for the adults. It has been found that in pelicans and herons, which have deferred maturity, immature individuals catch food less efficiently than adults and thus might be unable to feed a brood. Again, in the royal penguin (*Eudyptes schlegeli*), subadults weigh less than adults, presumably because they find it harder to obtain food, and only the heaviest birds breed; presumably this is because big food reserves are needed by the female to form eggs and by the male to take the first long incubation stint. Many species with deferred maturity lay only one egg, which suggests that they find it hard to obtain food for their young. In various seabirds, however, deferred maturity may simply result from an insufficient number of safe nesting sites. Finally, deferred maturity is also found in the males but not the females of various polygynous and promiscuous birds, in which there is such intense competition for mates that younger males may well fail to secure them and in which the elaborate plumage displays probably involve extra mortality from predation. One may conclude that, in birds, the age of first breeding may well have been evolved in relation to the maximum number of young that the adults can raise during their lifetime. Still, the alternative view—that breeding age is adapted to mortality—has not been disproved.

Natality in mammals

Mammals. Mammals produce from one to about 12 young at a birth, the latter being the figure for the African multimammate rat (*Rattus coucha*). In some species, individuals breeding for the first time have rather smaller litters than older parents; in others, such as hares and rabbits, there are seasonal variations in litter size; and in some species litter size varies with latitude. These parallels with birds suggest that similar factors operate in both birds and mammals. It has been shown that, in laboratories, infant mortality in the guinea pig (*Cavia porcellus*) rises with increasing litter size, but the extent to which it may do so in the wild is not known. Similarly, in man, the proportion of stillbirths is lowest for single babies, higher for twins, and much higher for triplets; it is likely that, in primitive conditions with poor food supplies, infant mortality also rises with the number of young at a birth.

Many rodents have several litters a year; many other mammals breed once annually and some only once in two or more years. The female meadow vole (*Microtus arvalis*) is fertile when 25 days old, many mammals when one year old, and other mammals not until several years old. As in birds, larger mammals tend to mature later than smaller ones, but there are many exceptions: at least two species of large whales first breed when two years old, whereas elephants and men mature in their teens. Maturity is also influenced by phenotypic factors. Thus a reduced diet delays maturity in various domestic and wild species; this is because it retards the growth rate, and the species concerned breed at a particular size rather than a particular age. Humans are capable of producing young when quite undernourished, but in the near-starvation conditions of some concentration camps in World War II, many women ceased to menstruate and could not have conceived.

Fishes. In fishes, as in ducks, there is a broad inverse correlation between the numbers and sizes of the eggs; this implies that egg production is limited by the food reserves of the female, and that larger eggs may in some circumstances be sufficiently advantageous to offset the disadvantage of smaller numbers. In general, marine fishes with planktonic larvae have many small eggs, and the larvae hatch with abundant food around them so do not need large food reserves; on the other hand, various deep-sea fishes and those that breed in polar coastal waters have fewer and larger eggs, because food is less abundant for the young and it is advantageous for them to be larger at hatching. Some species that have only a few large eggs protect them by making a case for them or even by staying beside them, and a very few fishes are viviparous (producing their young alive from eggs retained in the body). This has given rise to the idea that, because the eggs are protected, fewer are "needed"; but as suggested earlier, natural selection must favour those individuals leaving the most offspring. Presumably, the species in question leave the most offspring when they utilize their food reserves to produce a few well-nourished and well-protected eggs rather than many small and unprotected eggs.

Most temperate-zone fishes, once they mature, breed in successive years until they die. A few species living in

lakes that dry up each summer breed once and die, their eggs withstanding the drought. More remarkably, various salmon and eels grow for several years and then undertake a long migration from the sea to freshwater (or conversely in the case of eels), where they breed once and die. The long distance between the growing and breeding grounds of salmon and eels doubtless makes it disadvantageous for the female to breed in successive years. If a female can put virtually all her food reserves into eggs, she can produce more at one time than if she has to survive afterward; even so, most species of fishes can probably lay more eggs in a lifetime by breeding in successive years.

The relation between breeding and food supplies

Most fishes, like various mammals, first breed at a certain size rather than at a certain age, and they grow faster and mature earlier when food is more plentiful. Each individual also lays more eggs when it is larger. A striking illustration is the history of the North Sea herring (*Clupea harengus*). In the years around 1925, most of these herring first bred at the age of five; between 1950 and 1952 they bred at the age of four; and after 1952 they bred at the age of three. This was linked with an increase in the supply of their main food, the crustacean *Calanus finmarchicus*. The young of many other marine fishes have also grown faster and matured earlier after commercial fishing made their food more plentiful by removing many of the larger fishes.

Freshwater crustaceans. The inverse correlation between egg number and egg size is also found in crustaceans. Thus the "clutch" (retained in a brood pouch) of the freshwater copepod *Eudiaptomus graciloides* consists of nine to 18 small eggs in spring and of four larger eggs in summer, the total volume being in both cases the same. Food is plentiful for the larvae hatching in spring but not for those hatching in summer, for which food reserves in the egg are therefore advantageous. Similar variations have been found in other species. In the water fleas (*Daphnia*), as in fish, breeding starts at a particular size rather than a particular age, and the number of eggs laid tends to increase as the body size increases. Water fleas also have asexual generations, and the descendants of one female (a single clone) have the same genetic constitution. Individuals of different clones living in the same environment at the same time may differ consistently from each other in egg number, which suggests that there are hereditary differences in the size of the clutch.

Insects. In insects there is also a general tendency for species in which the larvae hatch with abundant food to lay many small eggs and for other species to lay fewer but larger eggs. This reaches its extreme development in the few viviparous species (in which the young develop inside the parent), of which the tsetse flies (*Glossina*) produce only one young at a birth. In some insects, such as the bordered white moth (*Bupalus piniarius*), the number of eggs laid is proportional to the size of the female's body and hence to the amount of nourishment she received as a larva.

In insects that lay their eggs on a limited source of food for the larvae, such as a flow of sap from a tree, a ripe fruit, a cowpat, or a corpse, there is some adjustment of the number of eggs laid to the duration of the food supply and to the number of eggs or larvae already present. An insect is likely to leave more survivors if it seeks another source of food than if it lays its eggs in a crowded or deteriorating source. Aphids, for instance, lay more eggs on fresh young leaves rich in nutrients than on old leaves poorer in nutrients, and a predominance of old leaves may inhibit breeding. Comparable adaptations are found in parasitoid insects; *i.e.*, those that lay their eggs in the larvae or pupae of another species, and in which the larvae obtain their food by consuming the host's body. Those that parasitize species of similar size to themselves lay only one egg in each host and can also detect whether another individual of their species has already laid an egg there. In contrast, species that are much smaller than their host lay many eggs in it. Finally, in the specialized social insects, such as ants and bees, many sterile workers provide food for, and protect, the larvae hatching from eggs laid by the queens. This greatly increases the number of eggs a queen can lay, since egg laying is thus her main or sole duty; the modification reaches an extreme among those termites in which a queen lays 6,000 to 7,000 eggs every 24 hours for a period that may extend over years.

Flatworms. Small animals are often thought of as having more young than large animals. But a freshwater flatworm in the genus *Polycelis*, which is less than 1 centimetre (0.4 inch) long, produces only 2.7 young per cocoon; and the average natality is only 1.0 per adult because many adults do not breed. This low reproduction rate has been shown to be the result of a limited food supply: in one experiment, natality rose by seven times when much extra food was provided in a pond.

Flowering plants. In certain fishes and insects, as already mentioned, the number of eggs laid by different individuals of the same species varies somewhat with their body size. This variation is far greater in flowering plants. In the poppy *Papaver rhoeas*, for example, the number of seeds produced by different individuals of the same genetic constitution, but grown under different conditions, varies between one seed capsule with four seeds and 400 seed capsules with 2,000 seeds each. In some trees, the number of seeds produced by the same plant varies enormously in different years, depending partly on climatic conditions and partly on a tendency for a good crop to be followed by a small one.

Seed variations in plants

When species of plants are compared, it is found that the size of the seeds tends to be smaller when there are more of them. Small seeds are advantageous, not only because more of them can be produced but also because they are dispersed more readily in the air. They are characteristic of those species that colonize broken ground. Because their seeds disperse so readily, species characteristic of broken ground, such as many of the daisy family

From D. Lack, *Natural Regulation of Animal Numbers*, published by Clarendon Press, Oxford

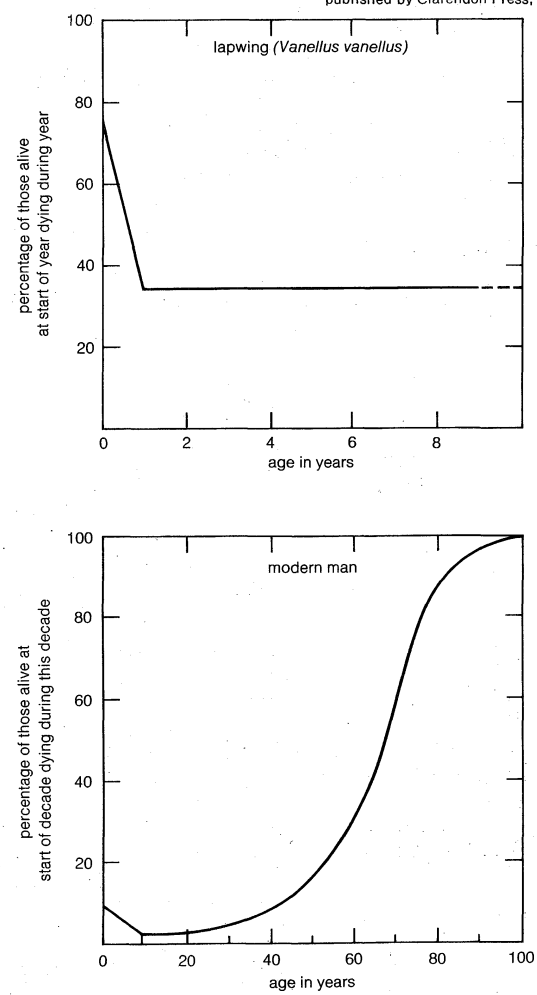


Figure 2: Mortality rate for (top) lapwing (*Vanellus vanellus*) in Britain and (bottom) white men in U.S. between 1929–31.

Compositae, are among the first to reach oceanic islands where, in the absence of competitors, some of them have evolved into trees that have larger and much less dispersible seeds. The seeds of forest trees usually have to germinate in deep shade, so it is advantageous for them to have large food reserves to give them a good start. Some plants, as already mentioned, have two means of reproduction: a seed requires fewer nutrients from the parent plant than a vegetative shoot but has a much smaller chance of survival than a shoot that remains attached to the parent plant until it is viable.

Mortality. The death rate varies at each stage of the life cycle. In long-lived species, the "potential age" to which survival is possible is much higher than the average age at death. Hence it is useful to calculate the mortality and the expectation of further life separately for each age group.

If an animal population is kept under favourable conditions in the laboratory, with plentiful food and without enemies, the annual adult mortality is much lower, and the expectation of further life much higher, than in a wild population of the same species. In the laboratory, mortality may be fairly high in the small young, but it is low in large young and in young adults; and while it rises with increasing age, it is high only in the elderly. At the opposite extreme, among wild birds in natural conditions, the losses of young and juveniles are extremely high. Adult mortality is also relatively high but does not rise with increasing age, nor does the expectation of further life decrease, until the individual reaches senility, as in the case of the lapwing (Figure 2). Relatively few individuals reach that age, however. A similar contrast is found in mankind: mortality in Western civilization (Figure 2) is similar in pattern to that of a protected laboratory animal, whereas that of Neanderthal man and of men living in Roman times seems to have been similar to the pattern of various wild mammals, though not quite so extreme in its characteristics as that of wild birds. Curves for the expectation of further life under different conditions are compared in Figure 3.

Birds. Two methods are available for determining the mortality of birds. The whole of a particular breeding population may be marked, and the number returning in each later year recorded; or the age can be determined of all those birds, banded by many different observers, that are later found dead and reported by others. The former method gives a lower figure for the annual mortality than the latter, probably because breeding birds comprise the most successful part of a population.

In all the birds so far studied, mortality is much higher in juveniles than adults, and it is particularly high soon after the young become independent of their parents. The highest known adult mortality is 70 percent per annum (which means an expectation of further life, at any given time, of only about 11 months), found in the Senegal fire finch (*Lagonosticta senegalla*). The lowest figure is 3 percent (expectation of further life, 33 years) in a very small population of the royal albatross (*Diomedea epomophora*) in New Zealand. As already mentioned, in a stable population, mortality equals natality; correspondingly, the Senegal fire finch has four broods each year with a mean of 3.4 young per brood, while the royal albatross raises, at most, only one young every two years. The mortality of various birds is summarized in Table 2.

Table 2: Annual Adult Mortality in Various Birds

species	average annual mortality (percent)	expectation of further life (years)
Fire finch *Lagonosticta senegalla*	70	nearly 1
Typical songbirds, ducks, gallinaceous birds, and doves shot for sport (northern Europe, North America)	50 (45–55)	1½
Herons, gulls, terns, waders, plovers, and one species of pigeon, in northern Europe and North America	30–40	2–3
Swifts *Apus melba* and *A. apus.*	18–20	5
Penguin *Megadyptes antipodes* (New Zealand), kittiwake *Rissa tridactyla* (Britain), manakin *Manacus manacus* (Trinidad)	10	9½
Gannet *Sula bassana* (Britain) and Australasian shearwaters *Puffinus tenuirostris* and *P. griseus.*	6	16
Manx shearwater *P. puffinus* (Britain)	4	25
Royal albatross *Diomedea epomophora* (New Zealand)	3	33

Mammals. The age of some mammals can be told from the annual rings on their teeth or horns, and that of a few others has been determined by marking. In wild rats, mice, and shrews, virtually all the adults die within a year, while at the other extreme only about 3 to 5 percent of adult African elephants (*Loxodonta africana*) die each year, varying somewhat with conditions. In all the wild mammals so far studied, adult mortality rises steeply with age after the first few years. In the Himalayan tahr (*Hemitragus jemlahicus*), a goat introduced into New Zealand, about half the young die in their first year; in the second year mortality is only 1.3 percent, after which the figure rises steadily to 38 percent in the 12th and 13th years; omitting the first year, overall average mortality is 15 percent. In the male waterbuck (*Kobus defassa*), annual mortality in the second to fourth years is only 2 percent, but in the eighth year it is 14 percent; thereafter it increases steeply, and the overall figure is similar to that in the tahr. Because mortality rises with increasing age, the average annual mortality is a less meaningful figure in wild mammals than in wild birds. In young adult mammals, mortality ranges between 50 and 67 percent in various larger rodents and smaller carnivores; it is between 15 and 40 percent in various deer; but it is apparently only 5 percent in the mountain sheep (*Ovis dalli*) and slightly less in the hippopotamus (*Hippopotamus amphibius*).

High mortality among young fishes and insects

Fishes. The ages of fishes that live in high latitudes can be determined from the annual rings on the scales, bones, and the otoliths in their ears. Fish fry have a huge mortality. In the few unfished populations so far studied, adult mortality rises with age as it does in mammals. In the whitefish *Coregonus clupeaformis*, for example, the figure rose in one lake from 8 percent in the 12th year to 17 percent in the 20th year and to 45 percent in the 27th year; in another lake it rose from 41 percent in the 7th year to 59 percent in the 13th year. Such figures show that annual mortality may differ greatly in different

From D. Lack, *Natural Regulation of Animal Numbers*, published by Clarendon Press, Oxford

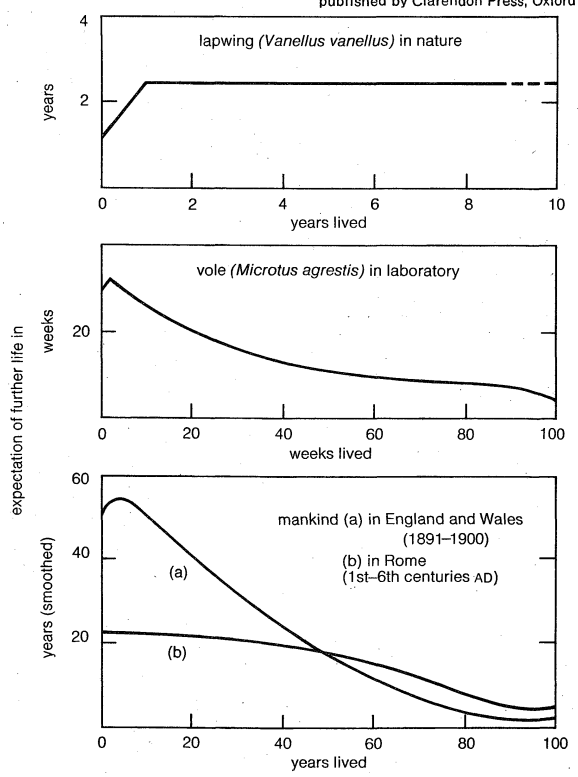


Figure 3: Life expectancy for (top) lapwing (*Vanellus vanellus*), (center) vole (*Microtus agrestis*), and (bottom) man.

populations of the same species. In the Queen Charlotte Islands' herring (*Clupea pallasii*), annual mortality increased from 28 percent in the 6th year to 72 percent in the 11th year. With heavy commercial fishing, however, annual mortality becomes much higher and may be almost independent of age. In one lake, the most common age group of the whitefish had been those aged 6, but after five years of heavy fishing the two-year-olds became most common. The annual mortality of adult plaice (*Pleuronectes platessa*) in the North Sea was 10 percent during World War II, when fishing was suspended, but rose to 50–60 percent afterward.

Insects. In most insects, as already noted, the adults die after breeding; interest therefore centres on the losses in earlier stages. The spruce budworm (*Choristoneura fumiferana*), the larvae of which eat the balsam fir, may be taken as an example. Each female lays about 200 eggs, of which an average of 19 percent are lost (nearly half through insect parasitoids and one-third through predators), leaving 162. Of the newly hatched larvae, 82 percent are lost (mainly through failure to establish on suitable parts of the food plants), leaving an average of 29.2. Of the surviving larvae, 86 percent are lost (nearly half through insect parasitoids and over one-tenth through predators), leaving 2.7 to emerge as adults. Of the adults, about one-fifth are lost before they breed, leaving one male and one female to repeat the cycle.

III. Changes in population characteristics

GROWTH OF POPULATIONS

If a few fruit flies (*Drosophila*) are put in a milk bottle with banana mash suitable for their larvae, or a few flour beetles (*Tribolium*) are put in a tin of flour, or a few water fleas (*Daphnia*) are put in a jar of pond water, they increase through breeding, first slowly, then rapidly, then slowly again. If their numbers are plotted against time, as in Figure 4, they form an S-shaped curve with the ends flattened. The initial slow, and later rapid, increase may actually represent a constant rate of change. A population that doubles its numbers in each generation seems to increase slowly from two to four to eight, and much more rapidly from 256 to 512 to 1,024. This type of population growth is called exponential or geometric.

From W.C. Allee *et al.*, *Principles of Animal Ecology*, 1st ed. (1949); W.B. Saunders Company

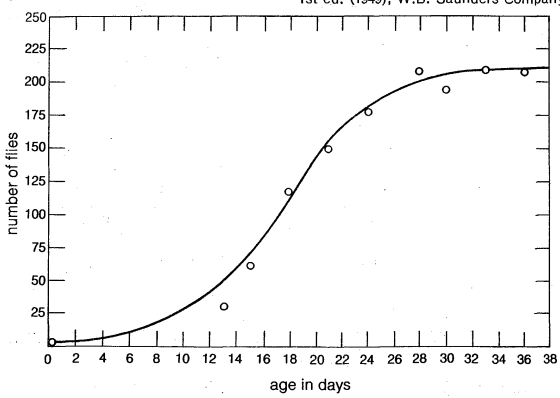


Figure 4: Logistic growth of a laboratory population of the small fruit fly (*Drosophila melanogaster*). Open circles indicate the points resulting from one experiment.

Variable rates of increase

When an animal is introduced by man to a new region suitable for it, it increases at first slowly and then rapidly, probably in an exponential way. After an animal has become abundant where introduced, its numbers in that area may stabilize while at the same time it is spreading into new areas. In this way the house sparrow (*Passer domesticus*) spread from the Atlantic to the Pacific coasts of the United States in less than 40 years. Similar spectacular increases have been recorded among plants: the water fern *Salvinia auriculata* was introduced to Ceylon in 1939, and only 16 years later had put more than 25,000 acres of rice paddy out of production.

Some laboratory populations eventually level off in numbers, but the situation is artificial in that their food supply is constantly renewed. The evidence suggests that, in nature, a successfully introduced population usually declines after a time and does not again reach its first high level, probably because its food supply declines. A few reindeer (*Rangifer tarandus*) introduced to St. George Island in the Pribilovs in 1911 increased to 200 by 1922, fell to 50 in 1926, and fluctuated somewhat below this figure in succeeding years. Another 29 reindeer introduced in 1944 on St. Matthew Island increased to 6,000 by 1963, decreased to less than 50 in the following year, and two years later were practically extinct. They had destroyed their food supply. In the late 19th century the cattle egret (*Ardeola ibis*) crossed from the Old World to the New, where it is increasing and spreading rapidly, but in a habitat (grass pasture with cattle) made by man. Man has also caused the rapid increase of various species in their natural habitats by changing their food situation. The herring gull (*Larus argentatus*) has been increasing rapidly on the coasts of Europe and North America; in the latter continent it has been doubling in numbers every 12 to 15 years since 1900, mainly because it now feeds extensively on fish waste and on rubbish dumps outside towns.

Examples of rapid increase by natural means in natural and undisturbed habitats are far less common because there are usually no ecological niches to be filled. A rapid though temporary natural increase can be observed when numbers have been greatly reduced by a catastrophe, such as a severe winter; the species involved usually attain their former numbers within a few years.

While the rate at which a population can increase is influenced by natality, this is not the main factor determining eventual numbers. Species with a high natality are not likely to be more abundant than those with a low natality. As pointed out by Darwin, the fulmar petrel (*Fulmarus glacialis*) is extremely abundant (and, since he wrote, has greatly increased in western Europe) but has an exceptionally low natality since it lays only one egg a year and requires several years to mature.

FLUCTUATIONS IN STABLE POPULATIONS

A stable population defined

All animal populations eventually stabilize at a level around which they fluctuate between relatively narrow limits compared with what is theoretically possible. The extent of the fluctuations, however, does vary greatly; for example, in studies over a period of 40 years of the gray heron (*Ardea cinerea*) in the Thames basin in England, the largest breeding population seen was about double the smallest. At the other extreme, in studies of the bordered white moth in a German pine forest, the largest population in a period of 60 years was 100,000 times the smallest. But even the largest observed variation is negligible compared with what is theoretically possible. The duckweed (*Lemna minor*) can double in numbers each 2½ days, and at this rate an initial square inch of duckweed would cover one acre in 55 days: yet the waters of the world are not all covered with duckweed. A pair of European robins (*Erithacus rubecula*) producing ten young a year could theoretically increase to over 120,000,000 in 10 years. Since even the largest observed fluctuations are negligible compared with what is theoretically possible, animal populations are said to be "stable"; it should be understood, however, that this term allows for fluctuations of the size mentioned, provided that, in the long term, numbers tend to fall when high and to rise when low. The only unstable populations in nature are those recently introduced to a new area and those decreasing to extinction.

Most biologists consider that populations as stable as those found in nature must be regulated by "density-dependent" factors; *i.e.*, factors that tend to reduce numbers when they are high and to allow them to rise when they are low. The contrary view, still advocated by some workers, is that animal numbers fluctuate irregularly by chance and should not be spoken of as controlled or regulated. The conflict in opinion has been partly the result of a difference in interest. Some biologists have been interested primarily in annual fluctuations (particularly if a pest is numerous only in

occasional years), which may be caused by climatic or other factors acting independently of population density. Others have been mainly concerned with the level about which such fluctuations occur, and it is in determining this level that density-dependent factors are important. Density-dependent factors may affect natality, mortality, or movements of the population; though some populations may be regulated by a single factor, in others two or more factors—for example, food supply and predation—may interact, or the effect of a density-dependent factor may be greatly influenced by that of a density-independent one.

Causes of seasonal fluctuations

Seasonal fluctuations in numbers. Nearly all animal populations show seasonal changes, often quite large. In small organisms, some changes of this type are linked with the time needed to complete one generation, or in marine organisms they may be related to tidal cycles; but by far the biggest fluctuations are connected with seasonal change in the environment—from winter to summer at high latitudes, and from the dry to the rainy season in tropical grassland and savanna. This is because nearly all the animals in such places breed at one time of the year, usually in late spring or summer at high latitudes and in the rainy season (or just after it) at low latitudes. They increase greatly in numbers toward the end of the breeding season and thereafter decline until breeding commences the following year. Hence in studying statistics of animal populations based on annual censuses, it should be remembered that between the annual counts there is each year a big and rapid increase in numbers, followed by a corresponding but more gradual decline, and that these seasonal fluctuations in numbers may be much bigger than those between one year and the next. Figure 5 shows the amplitude of these fluctuations in one species of bird.

In a few natural habitats, notably evergreen forests, tropical lakes and oceans, and the deep sea bottom, seasonal changes are small or almost absent. Even in a rain forest, however, there are small regular differences from month to month in rainfall, vegetation, and the availability of insects; these seem to affect the breeding seasons of many birds. On the other hand, many tropical seabirds that feed on marine organisms on or near the sea surface may breed in any month of the year, which suggests great uniformity in the numbers of their prey and hence also in the food of their prey.

From D. Lack, *Natural Regulation of Animal Numbers*, published by Clarendon Press, Oxford

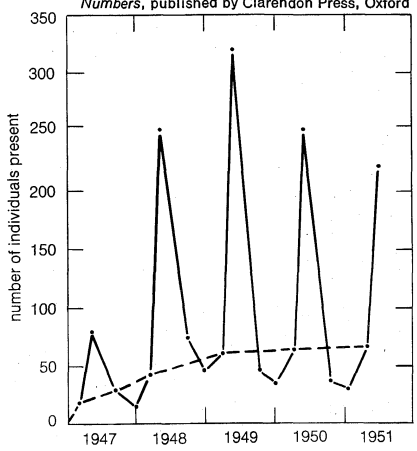


Figure 5: Seasonal changes in population of great tit *(Parus major)* in a wood near Oxford. Solid line represents total population, including juveniles; dashed line, number of breeding adults.

Importance of the climatic factor

Irregular annual fluctuations. All populations fluctuate in numbers to some extent, most fluctuations being irregular. (The few regular cycles are described later.) Large annual fluctuations are usually connected with climatic factors that affect natality or the survival of the young; they are most marked in animals or plants that live for only one year or at least have a high adult mortality, since in such species the young of one year form all, or at least most, of the adult population of

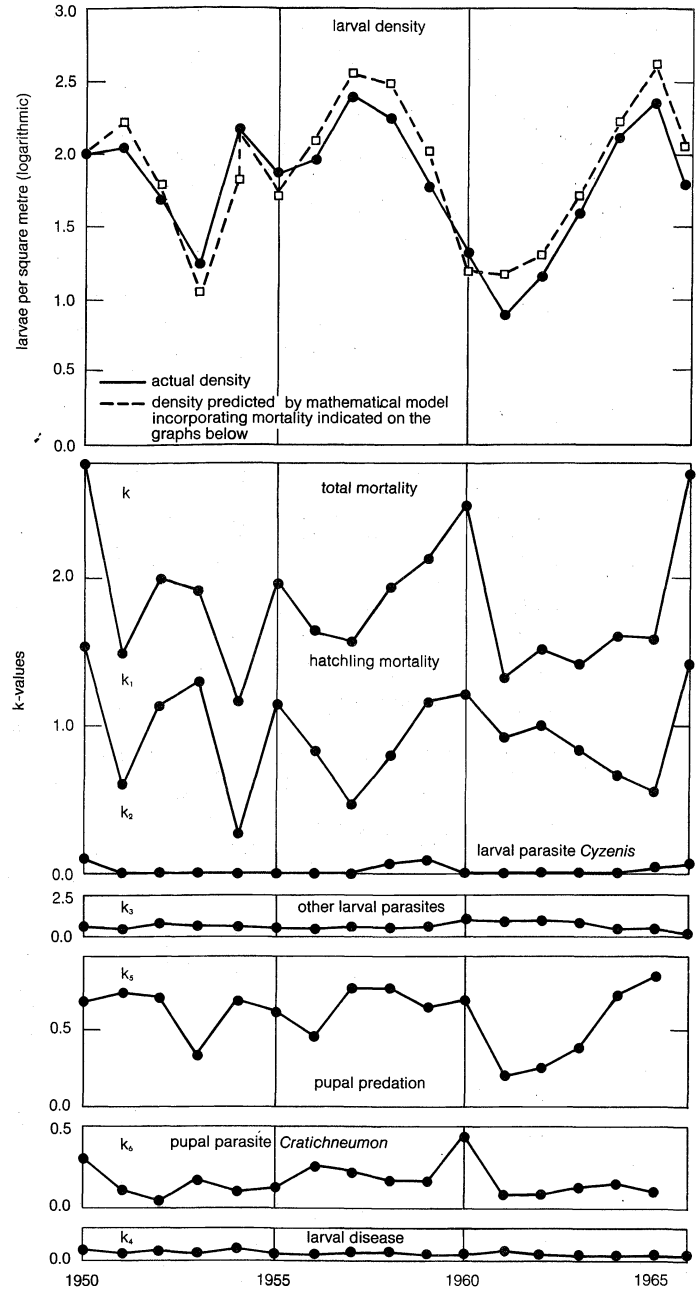


Figure 6: Study of winter moth *(Operophtera brumata)* over 16 years in Wytham Woods, Oxford, showing (top) total population and various sources of mortality (plotted on logarithmic scale). The k-value is the logarithm of the ratio of populations before and after the mortality in question has acted. Note that hatchling mortality k_1 is the only one that parallels annual pattern of total mortality.

From T.R.E. Southwood (ed.), *Insect Abundance;* Blackwell Scientific Publications, Ltd., Oxford

the following year. Large annual fluctuations in the production of young (or of seeds) may, however, cause little change in the numbers of a species in which the adults have reached the limit set by space or food supply, especially if the adults have a low mortality. Also, not all annual fluctuations result from variations in the birth and survival of young; in a species living near the northern limit of its range the adult mortality linked with winter cold may vary greatly from year to year.

Irregular fluctuations in numbers have been of special interest to entomologists, particularly in respect to species that remain comparatively scarce for many years and then have a huge outbreak. Examples are the spruce budworm moth, the larvae of which occasionally devastate the forests of balsam fir in northeastern America, and several other species of moths, the caterpillars of which occasionally strip German pine forests of their

needles. It is thought that in most years the numbers of such species are held down by insect parasitoids or predators, but that if climatic factors lead to an exceptionally high production of larvae, especially in two or more consecutive years, the natural enemies do not have time to increase proportionately; thus, the caterpillars then increase rapidly until they are so numerous that they defoliate the trees and eventually die in large numbers of starvation.

Such occasional huge outbreaks are rare. A more usual picture is presented by the annual fluctuations of the winter moth (*Operophtera brumata*) in a wood near Oxford, shown by the continuous line in Figure 6 (plotted on a logarithmic scale). Figure 6 shows on the top line the mortality of the same species at each stage of its life history; the mortality assignable to each parasite, predator, and disease organism has been plotted separately. Only the hatchling mortality (that is, the loss of larvae between hatching and settling on the young leaves) fluctuates parallel to the overall mortality; it therefore must be reckoned as the key factor in the annual fluctuations in numbers. The numbers of various other species of caterpillars on the same oak leaves fluctuate parallel to those of the winter moth, suggesting that a climatic factor is responsible. This graphic method of key-factor analysis is important in entomological research.

The great tit breeding near Oxford has also fluctuated irregularly in numbers, varying in one wood between 20 and 86 breeding pairs. In this species, the key factor has proved to be the mortality of the juveniles between the time they leave the nest in midsummer and the following November; in that month the juvenile to adult ratio has varied in different years between 0.2 to 1 and 2.6 to 1, parallel to the subsequent size of the breeding population. The cause of these juvenile losses is not known, but it is probably linked with a climatic factor since the number of breeding pairs has fluctuated more or less in parallel for 20 years in broadleaved woods in England and The Netherlands and also in pine woods (a suboptimal habitat) in both countries. The related blue tit in the same woods has fluctuated in a like manner.

Another species with marked annual fluctuations in numbers is the red grouse. The adult grouse eat mainly heather. The average clutch is slightly higher and the proportion of surviving young much higher in summers when the heather has a high nutritive value. The actual mechanism of the link between the state of the heather and juvenile survival is not known. In the years when many young survive, the size of the autumn territories is smaller (see below *Density-dependent natality*), and the breeding population of the following year is larger. These fluctuations in numbers are proportionately greater in areas of low, rather than high, average density of grouse.

The population of the gray heron in England fluctuates as a result of differences in adult mortality; it decreases markedly after each severe winter, presumably because of starvation, recovers quickly in the next one or two years, and thereafter stays more or less stable until the next hard winter. The speed with which numbers return to their earlier level shows the potentiality for increase discussed earlier. That this potentiality is not realized except after a setback in numbers shows that some other factor holds it in check.

Reproduction rates and population density

Density-dependent natality. Density-dependent effects may be of two types: direct, in which increasing population exerts a negative effect on factors causing population growth; or indirect, in which increasing population exerts a positive effect, tending to promote further population increase. The latter would not, however, regulate numbers. The natality of many animals is known to be influenced by food supplies, and since food may be scarcer when the animals are more numerous, it is at least theoretically possible that natality varies inversely with population density. An example is the African elephant (*Loxodonta africana*); as shown in Table 3, both the age of maturity and the average interval between successive calves vary with population density. This is perhaps because, at higher densities, each elephant obtains less food, although the point has not been studied. Another probable example is provided by the southern elephant seal (*Mirounga leonina*). On the island of South Georgia, where adults are killed in large numbers, cows first have pups at the age of three; bulls first arrive on the breeding ground at the age of four and first hold harems at the age of seven. In contrast, on Macquarie Island, where the animals are protected, cows first have pups at a mean age of six; bulls first arrive at the age of six to eight and breed much later (one marked bull at 13 or 14). The age of first breeding depends on body size, and young seals grow much faster on South Georgia than on Macquarie, presumably because the removal of many adults on South Georgia leaves more food for the young. In the North American mule deer (*Odocoileus hemionus*), the average age of first breeding is also known to vary somewhat with population density. The same is doubtless true of other mammals.

Table 3: Reproduction of African Elephant Loxodonta Africana in Relation to Population Density

locality	density per sq mi	mean age at puberty (years)		mean interval between calves (years)
		female	male	
Mkomasi (very dry)	3	11	...	3.1
Tsavo (dry)	3	12½	14½	6.8
Murchison North (less dry)	4	14	15	7.0
Murchison South (less dry)	6	18	19½	8.9
Budongo (forest)	6–7	20	...	...

Source: Studies by R.M. Laws.

In the great tit at Oxford (in which the fluctuations in the number of breeding pairs have already been described) clutch size has been observed to vary inversely to some extent with population density, falling by an average of two eggs when the population doubles. This is, as yet, the only such example demonstrated in birds. In fish subjected to heavy fishing, the natality of the remainder rises, because with fewer numbers more food is available and the fish grow faster; both the age of first breeding and the number of eggs vary with body size. In the aphid *Drepanosiphum platanoides*, and doubtless other species, crowding inhibits breeding; this may reflect reduced food supplies, since breeding is also inhibited by food shortage. The number of eggs produced per female is also density dependent in the pine beauty moth, because when the larvae are crowded each forms a smaller pupa, and hence, in the following year a smaller adult, which lays fewer eggs than a larger adult. Again, the number of eggs laid by an insect parasitoid is equal to the number of unparasitized hosts it finds, and the latter number depends in part on the population density of the parasitoid species. Natality is also density dependent in triclad planarians: when two-thirds of the adults of a *Polycelis* species were experimentally removed from a pond, the natality of the remainder was 12 times as high as before, because the remaining adults grew much larger, a higher proportion of them bred, and the average number of eggs per cocoon increased; this effect was presumably the result of an increase in the food supply, because a similar effect was produced in another experiment simply by adding food to a pond, leaving the number of adults untouched.

Density-dependent mortality. A simple example of density-dependent mortality was cited earlier for nestling European swifts in which the mortality from starvation is much higher in broods of four than of three. Broods of young are not, however, independent populations. Three examples from such populations follow.

Density-dependent mortality in sockeye salmon

Salmon fry. In British Columbia, the adults of sockeye salmon come up the rivers to spawn in freshwaters between July and November. The proportion of eggs that fail to hatch rises markedly with an increasing density of adults. Latecomers may expose eggs laid by earlier females; many individuals are forced to lay in suboptimal places susceptible to either drying out or freezing; and

crowded eggs may die from fungal infection or shortage of oxygen. These factors are not sufficient to regulate the size of the population, however, since the total number of young that hatch continues to rise, although at a slower rate.

The small fry suffer heavy predation, chiefly from fishes of other species, on the way from where they hatch to the lake nurseries where they grow. Probably this predation is proportionately less when population density is high. This is evidently because, in years when prey is sparse, the numbers of the predators are held down by the numbers of their prey, and hence in a year with far more fry than usual there are proportionately fewer predators (as their numbers have not yet increased). In such years, small fry have a greater chance of survival than usual, which is probably why, as mentioned earlier, one of the four successive sockeye populations is much larger than the others.

After the fry reach the lake nurseries there is further heavy mortality from predatory fishes, especially in the first 100 days. This is directly related to population density and therefore affects their numbers. While the immediate cause of their mortality is predation, it is closely linked with food shortage, because when the fry are at higher densities their food is sparser; they grow more slowly than at other times, and because they are smaller they are more vulnerable to predation.

The growth rate of the fry is inversely related to their population density, and so is the average weight of the smolts when they set off for the sea. Moreover, the proportion of smolts returning later to breed is strongly correlated with their weight at departure and hence with their density as fry in the lakes. The population density of the fry is therefore inversely correlated with their subsequent survival at several stages of their life history; its overall effect on the production of adult salmon is summarized in Figure 7. The density-dependent mortality of

From *Mitteilungen der Internationalen Vereinigung fur Theoretische und Angewandte Limnologie* (1965)

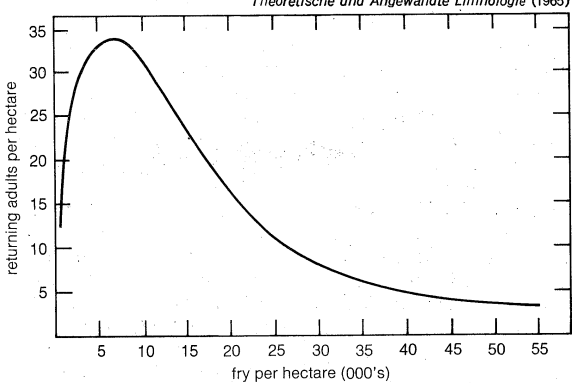


Figure 7: Density-dependent relationships in sockeye salmon *(Oncorhynchus nerka)*: relation between density of fry in lakes and that of returning adults.

the salmon fry regulates their numbers, bringing them down when they are high and allowing them to increase when they are low. The causal factors are predation acting in combination with food shortage.

Density-dependent mortality in flatworms

Adult flatworms. Food shortage also limits the population density of freshwater flatworms (genera *Polycelis* and *Dugesia*), though in a complex way because these relatively simple animals have the power, when food is short, of feeding from their own tissues. A *Polycelis tenuis* eight to nine millimetres (about 0.33 inches) long can survive without food for eight months, during which it shrinks to 1.5 millimetres (0.06 inch). Only the large individuals breed. Three species studied in North Wales feed on oligochaete worms, one of them also on pond snails. When their food supplies start to increase in spring, the surviving large adults breed. But their numbers are sufficiently close to the food limit to mean that the appearance of many small young produces a shortage, as a result of which many large adults become small and many young fail to grow. A little breeding continues until autumn, at which time food is sparse and hardly any large individuals remain; the heavy mortality of small individuals from starvation allows the remaining adults to grow larger. As mentioned previously, the experimental removal of two-thirds of the adults from one pond, and the provision of much extra food in another, caused a large increase in natality. In addition, in both ponds the mortality fell and the proportion of big adults increased. In an experiment in which a population was kept without food, the large individuals shrank to medium size, the medium sized became small, and the small died.

Plant seedlings. Poppies (*Papaver*) have small seeds with small food reserves. When experimental plots are sown with seeds at different densities, the mortality of the seedlings becomes higher at the higher densities, until it is so high that the addition of further seeds does not produce any further increase in the number of surviving plants and may even cause a decrease. The mortality among densely planted seedlings in this and other species has a number of causes, including higher humidity and hence greater fungal infection, more blanching because the plants partly screen each other from the light, and a great attraction of predators such as wireworms.

In similar experiments involving plants with large seeds, such as the corn cockle (*Agrostemma githago*), each seedling has more food reserves and mortality does not rise with increasing density, so that the greater the density of seeds, the greater the density of plants produced. At the higher densities, however, the rate of growth of each plant is greatly reduced, and above a certain density there is no further increase in the total weight of plant matter per unit of area, or in seed production per unit of area. Corn or maize (*Zea*) is another large-seeded plant in which the rate of growth decreases markedly with increasing density of seedlings; above a certain density, the number of seeds on the plants falls, not merely per individual plant but per unit area, although the total weight of green shoots continues to rise. These findings are of importance in determining the most economic density at which to sow crop seeds.

Predator–prey oscillations. In the examples so far considered, mortality was directly dependent on density, in the sense that the higher the density, the greater the proportion of individuals dying. But in some cases the effect of population density on mortality is delayed, notably when a prey species and its predator (or an insect parasite) each limit the numbers of the other. If, for example, both species are scarce, the prey can increase rapidly through reproduction; but the predator can then also increase in numbers, and the process continues until the predator is so numerous that it begins to reduce the numbers of its prey; this in turn brings down the numbers of the predator, and the cycle starts again. In such a cycle, the relationship between prey numbers and prey mortality is not simple and direct. Instead, the prey mortality is low throughout the phase of increase and high throughout the phase of decrease. A particular density of prey is followed by an increase in numbers at one phase of the cycle and by a decrease at the other. This is because the increase of the predator lags behind the increase of the prey.

Early mathematical analyses suggested that, in a predator–prey interaction of this type, the numbers of both species should oscillate regularly and that the oscillations might increase in amplitude with time. Efforts to test this idea in the laboratory ran into difficulties in achieving ecological balance: the predator tended to increase, exterminate its prey, and then die out. Regular oscillations

From *Hilgardia* (1958)

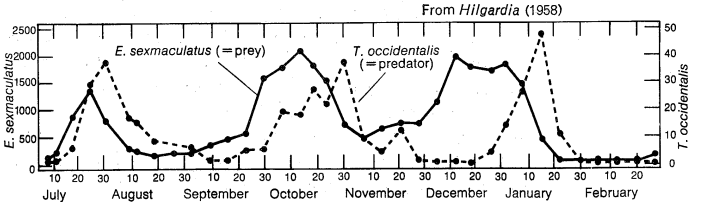


Figure 8: Laboratory study of oscillations in population density of a predator and its prey. The predatory mite *(Typhlodromus occidentalis)* preys upon the orange-feeding, six-spotted mite *(Eotetranychus sexmaculatus)*.

were finally produced in a closed laboratory population (Figure 8), but only when the situation was complicated by the addition of launching posts to assist the prey to reach undiscovered food sources within the culture and of partial barriers to reduce the movements of the predator between these sources.

Permanent reduction in numbers

Few natural populations of predators and prey oscillate regularly in numbers. In some cases in which a predator has been successfully introduced to control a pest, the predator has at first increased greatly; as a result the prey has decreased greatly, and after it the predator, following which both species have remained extremely scarce. This happened when the ladybird beetle was introduced and successfully controlled the cottony-cushion scale insect *Icerya purchasi* in the citrus groves of California. The same result followed when the moth *Cactoblastis cactorum* was introduced and successfully controlled the prickly pear *Opuntia* in Australia. As of 1971 there were over 60 cases in which an insect pest had been greatly reduced through the introduction of a predator or insect parasite and a number of others in which a harmful plant had been similarly controlled by a phytophagous (plant-feeding) insect. These large-scale field experiments show that a single predator is capable of holding the numbers of its prey to a level far below that which prevailed before its arrival. It is generally supposed that the prey continues to survive and to multiply in little pockets, but before it has multiplied far, the pockets are discovered and eliminated by the predator, by which time, however, further pockets have appeared elsewhere. This capacity for movement of both prey and predator can only with difficulty be reproduced in the laboratory.

A phytophagous mite that damaged strawberry plants was maintained at a low enough level by a predator so as to cause minimal harm. The reverse can occur, however, in populations that lose their predators. An example is the mule deer (*Odocoileus hemionus*) on the Kaibab Plateau in Arizona, following the removal of carnivorous mammals and human hunters. A population of about 4,000 deer in 1906 increased to about 100,000 in 1924 but then fell dramatically to about 20,000 in 1931 and to less than 10,000 in 1939, probably because overgrazing prevented adequate regeneration of the food plants (a delayed density-dependent interaction). Similar effects occurred in other parts of the United States following the removal of the natural enemies of the deer. In parts of Alaska where wolves are present, mule deer have a light winter mortality from starvation and the range remains in fair condition; but in parts where wolves are absent, the deer have a heavy winter mortality from starvation and there is marked deterioration in the condition of the range. Thus the interaction between the wolves and the deer greatly modifies the interaction between the deer and their plant food.

A similar situation is found on Isle Royale in Lake Superior, to which a few moose (*Alces americana*) crossed early in the present century. They increased to about 3,000 by 1930, when their food supplies were badly damaged, decreased through starvation to 200 in 1935, rose again to 800 in 1948, and decreased to 500 in 1950. Shortly before 1950, 20 wolves (*Canis lupus*) arrived; after that, up to the last published count in 1961, moose numbers stabilized around 600. The wolves tended to take calves or debilitated animals.

Regular fluctuations in numbers

Regular, large, and continuing cycles occur in the numbers of some wild mammals. If a cycle is regular, one may be reasonably certain that its cause is simple. The best attested example is that of the lynx (*Felis lynx*), which the records of the Hudson's Bay Company, covering over a century, show to have a remarkably uniform periodicity of nine to ten years. The years of peak numbers are similar in its prey, the varying hare (*Lepus americanus*), and clearly its numbers are determined by those of the hare. This is probably not a simple predator–prey interaction, for there is no evidence that the numbers of the hare are brought down primarily through predation by the lynx. The regularity in the successive declines of the hare suggests a simple interaction with some other factor, and the continuation of each decline for several successive years is characteristic of a delayed density-dependent factor; interaction with plant foods is perhaps involved, but this has not been investigated.

The best known regular fluctuations in numbers are the four-year cycles of lemmings. The only population so far studied for a long period is that of the brown lemming (*Lemmus trimucronatus*) at Barrow, Alaska. Lemmings are vegetarian, and at the peak in numbers, about every fourth year, there is heavy overgrazing followed by starvation. Grazing transfers most of the nutrient salts from the vegetation to the lemmings, but the nutrients eventually return to the soil and thus, after delay, become available again to the plants, which regenerate only slowly in the cold soil. Once the plants have regenerated, the lemmings increase again, and the cycle is repeated. Removal of the plant cover by the lemmings also exposes them to predatory birds and mammals, but the latter are not the primary cause of their decline.

Several alternative theories of the lemming cycle have been proposed. One is that lemmings regulate their own numbers by emigration. Emigration is particularly characteristic of the European species *L. lemmus*, which moves out of areas where food has become sparse to others where food is present. But when all suitable ground has been colonized and overgrazed, emigration is of no further benefit; the lemmings then descend in large numbers from the tundra and may even cross rivers or fjords. Another theory is that physiological stress caused by overcrowding (not starvation) induces metabolic deficiencies in the lemmings, especially in the endocrine system, that reduce fecundity or kill them. A third theory advanced to explain the continuation of the declines after numbers have fallen is that genetic changes occur in the population at different stages of the cycle.

The mammals and birds that prey on lemmings, like the predators of the varying hare, fluctuate in numbers with their prey. The goshawk (*Accipiter gentilis*) and the American red fox (*Vulpes fulva*) are of special interest in this connection because in some areas they prey on lemmings and have a four-year cycle, and in other areas they prey on varying hares and have a ten-year cycle. The numbers of grouse and ptarmigan living in the same areas also fluctuate in parallel, for unknown reasons, but presumably because their numbers are linked with those of the rodents through common foods or common predators.

Irregular fluctuations in numbers

The examples discussed so far concern simple, or apparently simple, interactions; but most animals, especially those living in natural habitats outside the Arctic, eat diverse foods and have diverse enemies. Because a diversity of factors is involved, their fluctuations are likely to be irregular. The population of the winter moth at Oxford has already been described in relation to key factors (see Figure 6). This species has a single food, the leaves of the oak, but its diverse enemies include larval and pupal parasites, a disease organism, and several pupal predators. Detailed analysis of the mortality attributable to each of these causes has shown that the main density-dependent factor is predation on pupae in the soil by staphylinid beetles, moles, and shrews, and that the mortality from this cause is sufficient to regulate numbers. A ten-fold increase in the density of pupae, for example, is accompanied by an increase in mortality equivalent to 55 percent.

In 1949 the winter moth was found to have become established as an introduced species in Nova Scotia, where it caused severe damage to trees through defoliation; its numbers were limited periodically by starvation. When the tachinid parasite *Cyzenis* was introduced from Europe, the numbers of the winter moth fell to one-thousandth of what they had been, and defoliation ceased. Although *Cyzenis* is present at Oxford, it is unimportant in controlling numbers there. This illustrates a point made earlier, that the numbers of a species may be regulated in different ways in different parts of its range. To give another example, much of the mortality of young barnacles (*Balanus*) on the rocks in Scotland is the result of intraspecific crowding, but in San Juan Island in the state of Washington, it is the result of predation by

snails (*Thais*); these very different factors govern both the distribution and abundance of the barnacles in the areas concerned.

MOVEMENTS

Movements often have an important effect on the size of a population, especially in mitigating temporary food shortage or overcrowding and are of four main types: migration, emigration, dispersion, and dispersal.

Migration. As commonly understood, migration is a regular, seasonal, long-distance, large-scale movement of populations in set directions between two areas, one of which (rarely both) is used for breeding. The term is also used for short movements, in any compass direction, to a special habitat for breeding, such as are found in many types of land, freshwater, and marine animals.

Migratory species

Many birds breed at high latitudes in summer and move for the winter to low latitudes or to coasts or to the sea. Many others have extensive movements between a breeding and a nonbreeding area within the tropics. A very few, notably the redpoll (*Carduelis flammea*) in northern Europe, the quail (*Coturnix coturnix*) around the Mediterranean, and the quelea (*Quelea quelea*) in tropical Africa, almost certainly breed in two areas in succession, but this is not yet fully proved. In mammals, also, certain bats and the caribou (*Rangifer tarandus*) migrate to breed at high latitudes in summer; various whales feed and form fat stores at high latitudes in summer and breed at low latitudes in winter. Various antelope migrate in the grasslands of tropical Africa. Some fishes and at least one squid migrate regularly like birds. In various eels and salmon, each individual moves once in its life from a breeding to a feeding area, returning after several years to the breeding area and then dying. Various northern butterflies breed in the north in summer; their offspring migrate south and breed; and their offspring in turn fly north the following spring, so that while each individual migrates only once, there is a twice-yearly migration of the population. The monarch butterfly (*Danaus plexippus*), on the other hand, migrates from a northern breeding area to hibernate in the south, the same individuals returning to the northern breeding area the following summer. Other butterflies migrate within the tropics.

The migratory habit enables a species to breed in one area and to survive the winter in another, or to grow large in one area and to breed in another, or even to breed in two areas at different seasons. Nonmigratory animals may have to survive cold or drought in a resting stage as eggs or pupae, or, if adult, in hibernation or aestivation.

Certain northern birds are partial migrants, some of the population leaving and the rest staying for the winter in the breeding area. In such species, the advantages of leaving or staying must be almost equal; this view is supported by the fact that the proportion migrating increases toward the north of the range, where the winters are more severe. Like other adaptations, migratory movements are anticipatory. If food will become scarce, it is advantageous for an animal to move away before this occurs, and similarly it is advantageous to leave a wintering area so as to arrive on time in the breeding area. Many migrant species have evolved special abilities that enable them to sense time, accumulate fat reserves, and navigate over long distances—strong circumstantial evidence that migration is advantageous to the species concerned.

Animals that emigrate

Emigration. Emigration differs from migration in that it occurs at irregular intervals, in a variety of directions, without a set return movement, and in direct response to food shortage or overcrowding. The European red squirrel (*Sciurus vulgaris*) and the Siberian nutcracker emigrate from the taiga in years when they are numerous and coniferous seed crops fail; the waxwings *Bombycilla garrulus* and some other northern birds emigrate when particular berry crops fail; the lemmings (*Lemmus*) move from the tundra when they have overeaten their vegetable foods; the hawks and owls that prey on lemmings also move when the latter decline; water birds breeding beside shallow lakes in the steppes leave in the occasional summers when the lakes dry out; locusts move when overcrowded, producing a special migratory form; and various aphids, after multiplying greatly on a food plant, produce winged forms that drift away (the last response has been shown in *Drepanosiphum platanoides* to be density dependent at the fourth instar—*i.e.*, before the adult stage). Some emigrations are combined with migration, in that they may occur at set seasons and in set directions with later return movements, but the numbers participating are much greater in occasional years than in others.

Emigration is advantageous for any individual that has a greater chance of survival if it moves than if it stays, and hence the habit is explicable in terms of natural selection. It may be specially advantageous in cases of local crowding or local food shortage, when small and inconspicuous movements suffice to take the emigrants to suitable new areas. Even in occasional catastrophic years, when many emigrants die, the chance of survival may be higher for an individual if it moves than if it stays; recoveries of marked birds have shown that some of the emigrants, in fact, survive and return later to their breeding grounds. There is no positive evidence for the alternative idea that the emigrants constitute a doomed surplus that moves out to die in order to save a remnant of the species from starvation; no means are known by which a species can evolve behaviour that benefits unrelated members of the species but is harmful to the participating individuals.

Dispersion. The small and usually inconspicuous movements by which various animals space themselves out, especially for breeding, lead to what is called dispersion. Birds that are solitary breeders tend to be spaced out fairly evenly but with a tendency to be denser where food and other conditions are more favourable; in colonial species, both the spacing and the size of the breeding colonies vary somewhat with the availability of resources. Since adult birds usually breed from year to year in the same place, dispersion is brought about mainly by the younger individuals that are settling for the first time; they seek the preferred habitat of their species, but if it is densely oocupied, they may settle in suboptimal habitats. The latter may give them a better chance of raising young than would otherwise suitable but crowded areas, and therefore suboptimal habitats are occupied chiefly in years of high numbers. Similarly, the young adults of colonial species tend to settle in established breeding colonies but may move elsewhere if these are crowded in relation to resources. In winter, birds join feeding flocks if this helps them to obtain food and leave them if it does not. Similar principles apply to many other kinds of animals.

Dispersal. Dispersal is often used as a general term for animal movements but also in a special sense for the movements of young animals or plant seeds from their birthplace to where they settle down. It has little effect on numbers, although it is an important means by which an increasing species gradually extends its range. In various sedentary or sessile marine animals, however, and in many land plants, specialized larvae or seeds have adaptations for active or passive transport that enable them to settle at long distances from their birthplace.

IV. Interactions of populations

INTERACTIONS BETWEEN MEMBERS OF A POPULATION

Members of the same population may cooperate or compete with each other. Both forms of behaviour are the products of natural selection, in that they benefit the individuals concerned or their offspring.

Examples of cooperation

Cooperative behaviour. Much the most common example of cooperation is parental behaviour, found in simple form in various insects, fishes, and reptiles that protect their young. All mammals, many birds, and solitary wasps feed their young. So do social insects such as ants and termites, but in a specialized way. Cooperative breeding on a small scale occurs in a few birds. In one Australian songbird, a young male of the previous year's brood helps its parents to feed the new brood; in another, several pairs, all offspring of an initial pair, lay eggs in the same nest and feed a combined brood; and in yet another bird, a flightless rail, two males (usually broth-

ers) mate with one female and raise a joint family. In these ways the birds raise more young than they would otherwise, and since the extra helpers are close relatives, the habit can be evolved through natural selection. In certain mammals, family groups of young born in different years also stay together.

Social behaviour between unrelated individuals can be evolved through natural selection only if each participant is likely to benefit. The synchronized swarming for mating of various insects and marine invertebrates probably reduces the chance that any individual will be taken by predators; in polychaete worms it may also increase the chance of the eggs being fertilized. Similarly, the aggregation for mating of certain promiscuous male birds probably helps females to find them and increases the chance that an approaching predator will be noticed. The latter holds true also in the many species of birds, herbivorous mammals, and fish that feed in flocks, herds, or shoals. It has been shown that a solitary wood pigeon spends more time looking for possible predators, and therefore less time looking for food, than does one in a flock. If a predator is seen by one of a group, that member's alarm behaviour alerts the others. Aggregation may also increase feeding efficiency. Birds that search for localized foods, such as seeds, fruits, or shoals of fish, often locate them when they see feeding members of their own species. Certain pelicans and cormorants even help each other to catch prey by driving fish in groups. Lions, wolves, and hunting dogs cooperate in hunting herbivorous mammals.

Competitive behaviour. Competition occurs when resources are sparse in relation to the number of individuals requiring them and when the activities of one individual significantly reduce the resources available for another. In animals or plants of the same species, competition is mainly for food (nutrients); sometimes it is for space—in a sessile animal or a flowering plant for space in which to settle, and in various higher animals for a feeding or breeding territory. There may also be competition for mates.

Methods of competition

Most competition for food between members of the same species happens simply because each individual consumes whatever prey or nutrients it finds. However, in various flock-feeding birds that search for discrete items, such as seeds, one individual may drive another from what it has found. In a first encounter there may be an actual fight, but in later ones the individual defeated on the first occasion usually retreats to avoid a fight. In this way an order of precedence ("pecking order") may be developed throughout a whole flock.

The influence of both flocking and peck order on numbers has been studied in the wood pigeon in rural England. The number of birds that survive the winter each year is closely correlated with the annual variations in the availability of their main food, the leaves of clover in pasture. In the winter flocks, some individuals consistently displace others from food items, and these dominant birds are better nourished and heavier. The rate of disappearance of marked individuals during the winter, presumably from starvation, was found to be greater among subordinate than among dominant individuals, among lighter than among heavier individuals, and among juveniles than among adults. Further, the size of each flock was found to be correlated with the density of clover leaves in the area: when the flock becomes too large for the available clover, the subordinate individuals suffer disproportionately and tend to leave; but if they feed solitarily, they find even less food because of the time spent in looking for predators, and hence they starve. These individuals may also show symptoms of physiological stress such as enlarged adrenal glands, but this is the result of starvation, not crowding. A consequence of the pecking order is that, at any one time, only a few individual wood pigeons—those lowest in the hierarchy—will be starving, while the rest will be well nourished.

Competition often involves the possession and defense of territory. A few bird species that are solitary feeders, especially those that pounce on their prey from perches, defend individual feeding territories. Many more species of birds defend breeding territories, some of them solely for obtaining mates, others solely for nesting, but some for both breeding and collecting food. A few species, like the raven (*Corvus corax*) and the oyster catcher (*Haematopus ostralegus*), maintain an almost constant number of breeding pairs from year to year, presumably as a result of their territorial behaviour. The tawny owl (*Strix aluco*) also appears to maintain stable numbers through territorial behaviour. Its food supply, woodland rodents, is extremely variable; indeed, in some years the birds cannot breed, but they remain on their territories, perhaps because they hunt most efficiently when they know an area well. In contrast, the territories of the nomadic short-eared owl (*Asio flammeus*) are much smaller in years when, and places where, their rodent prey is abundant. Territory size also varies greatly in the same place in different years in the great tit (but here it is unrelated to food supplies) and red grouse; if owners are shot, other individuals take their place—coming, in the great tit, from suboptimal habitats, and in the red grouse, from individuals without territories that will die unless they obtain them. The pressure to obtain territories is evidently great, but the general significance of bird territory in relation to population density and food supply remains unknown.

Many rodents, ungulates, and carnivorous mammals fight each other at times. The significance of such fighting in nature, and its possible role with respect to regulating numbers or food supplies, is not known. Much fighting, of course, is connected with mating, especially in polygynous mammals, but this is not of concern here. Where fighting is related to population density, one may guess that it stems from a shortage of food or from some other factor that constitutes the primary means of population regulation. Territorial and other fighting also occurs in various lizards and fishes, and in sticklebacks it appears to limit the breeding density.

Dragonflies and ant colonies are also territorial. So are a number of marine invertebrates, the food of which is uniformly distributed, notably the grazers and those that eat detritus. The marine amphipod crustacean *Erichthonus braziliensis* defends its feeding territory from wandering individuals of its species, and, as it grows larger and feeds over a larger area, it evicts smaller individuals that have settled within this area. Barnacles also compete, in this case for space in which to settle and later to grow, and they destroy neighbours in the process.

INTERACTIONS BETWEEN POPULATIONS OF DIFFERENT SPECIES

The principle of competitive exclusion

Competition among species. It is generally agreed that two species with the same ecology—that is, which relate in the same way to the same environment—cannot persist in the same area. Where resources are limited, this is an inevitable consequence of natural selection, since it is unlikely that both species will be equally well adapted. If they meet, either one eliminates the other or each replaces the other in parts of the same range or habitat. Where two species coexist, each depends on different foods or other resources. This principle of competitive exclusion is fundamental to the understanding of animal and plant communities.

The main evidence in nature for competitive exclusion is that, in fact, no two species living in the same area have the same ecology; and in many cases of apparent overlap, important differences have been found by careful observation. Instances are known among birds in which one species has replaced another. The habitat of suburban Singapore was at one time colonized by a native munia, *Munia majah*, which was later replaced by the introduced *Munia punctulata*. It was also colonized by a native tailorbird, *Orthotomus sericeus*, but as forests were cleared to the north a related Indian species, *O. sutorius*, spread south and replaced it, though the native species still holds its own in the original coastal habitat. The fields and towns of the Canary Islands were formerly inhabited by the rock sparrow (*Petronia petronia*); but the Spanish sparrow (*Passer hispaniolensis*) later established itself in the towns, and the rock sparrow is now confined to the

countryside. Replacement has also been observed in wholly natural conditions. The brambling (*Fringilla montifringilla*) has replaced the chaffinch (*F. coelebs*) in the extreme north of Europe, and in recent warmer summers the boundary between the two species has shifted north, presumably because of competitive exclusion. Similarly the more southern curlew (*Numenius arquata*) has spread north at the expense of the related whimbrel (*N. phaeopus*).

In birds, no two species are known to coexist and eat the same food supplies, except where food is temporarily superabundant. But animals that are not limited in numbers by food can persist in the same place, even though they eat the same types of food. That is why several species of caterpillars can be found on the leaves of the same tree; they are limited in numbers by insect parasitoids, and each caterpillar species is parasitized by a different species or set of species. It is less easy to see how plants can coexist, since they depend on the same basic resources of light, nutrients, and space; but their roots descend to different depths, they grow at different seasons or to different heights, and they respond differently to the insects that eat them. Laboratory experiments have shown that two species of grasses, or a species of grass and of a clover, can coexist in a stable manner. Other experiments involving microscopic plants, protozoans, flour beetles, and other small animals have demonstrated competitive exclusion among them; while the principle itself scarcely needs such demonstration, the experiments have provided quantitative analyses of the ways in which such competition works under different conditions and of the effects of the differing adaptations of the competing organisms.

Animals that live upon one another

Negative direct interactions. Other species of animals interact with each other because one lives upon another. A predator is usually, though not always, larger than its main prey; a parasite is much smaller; while insect parasitoids, the larvae of which eat the living larvae or pupae of another species, are intermediate in various respects between predators and true parasites. Predators and parasitoids destroy their prey. The young of brood parasites, such as cuckoos (*Cuculus*) and cuckoo bees (*Psithyrus*), are raised by other birds and by social insects, respectively; in some cases they destroy the host young. Many internal parasites do little harm. Some disease organisms kill their prey and others seriously weaken them, but many have evolved less harmful effects, since it is usually not advantageous to them to destroy their home. Most blood-sucking and other ectoparasites do little harm, except as carriers of disease organisms. There are also bird "pirates," such as skuas or jaegers (*Stercorarius*), which merely rob other birds of their prey.

Most predator species take a variety of prey species. Many take different main prey in different areas or at different seasons in the same area, and if their normal main prey becomes scarce, they often turn to other species. Insect parasitoids, in contrast, often have specific hosts; their time of emergence coincides with that of the larvae or pupae that they parasitize, so their effect on prey numbers is somewhat different from that of predators. Both predators and parasitoids, however, may hold plant-eating prey species below the level at which they seriously damage the vegetation.

The role of disease in regulating numbers is uncertain. In various mammals, such as the mule deer (*Odocoileus hemionus*), in which undernourished animals often die of disease, food shortage is the critical factor. Diseases caught from introduced birds have eliminated various native birds from Hawaii and probably New Zealand, but these are examples of elimination rather than regulation of numbers. Since disease organisms spread more readily in crowded than in sparse populations, their influence is potentially density dependent; but this may be offset by their tendency to evolve less deleterious effects with time. Nearly all the attempts to control insect pests by disease organisms have failed. In Australia, however, huge numbers of the rabbit *Oryctolagus cuniculus* were killed by a viral disease (myxomatosis) in the years immediately after it was introduced, and two decades later, their numbers were well below what they had been. In laboratory populations of the flour beetle *Tribolium castaneum*, the level at which numbers are maintained with a constant food supply is 2½ times as high in the absence of the sporozoan parasite *Adelina* than when the parasite is present. In experiments on competitive exclusion, *Tribolium castaneum* usually eliminated the related species *T. confusum* in the absence of the parasite, but when it was present in both species, *T. confusum* was the more successful.

Cooperation among species

Positive direct interactions. Brief mention may be made of the ways in which different species may help each other, although these do not have a direct influence on populations except in allowing a species to fill an ecological niche it otherwise would not fill, or to obtain food more easily, or to obtain greater defense from predators. In some such associations, each species is entirely dependent on the other, as for instance in the fungus and green alga that comprise a lichen, or in the mycorrhizal fungi on the roots of various trees and orchids, or in the intestinal micro-organisms that break up cellulose for herbivorous mammals or termites. In other cases, although both species benefit, each can exist without the other—for example, in the associations of crabs and sea anemones, in which the crab gains protection through the anemone's tentacles, and the anemone gains food fragments from the crab. Again, the greater honey guide (*Indicator indicator*) finds wild bees' nests, to which it guides the ratel or honey badger (*Mellivora capensis*); the latter breaks them open, after which both feed on the contents. In many other associations, only one of the two species benefits. Various birds follow, or ride upon, large mammals, catching the insects they disturb. Small birds nest alongside the homes of stinging hymenopterans or large raptors, thus gaining protection from predators. There are even looser associations, such as the mixed-species herds of mammals in which all the individuals probably benefit from the greater awareness of the approach of a predator.

BIBLIOGRAPHY

General works: W.C. ALLEE *et al., Principles of Animal Ecology* (1949), includes an excellent discusison of life tables, mortality, and laboratory studies of population; H.G. ANDREWARTHA and L.C. BIRCH, *The Distribution and Abundance of Animals* (1954), with special reference to insects, takes the view that animal populations are not regulated by density-dependent factors; L.R. CLARK *et al., The Ecology of Insect Populations in Theory and Practice* (1967), covers the diverse theories of population control as they relate to insects; D.L. LACK, *The Natural Regulation of Animal Numbers* (1954), adheres to the idea of density-dependent regulation; M.E. SOLOMON, *Population Dynamics* (1969), a short introduction to the principles of density-dependent regulation, primarily in insects; V.C. WYNNE-EDWARDS, *Animal Dispersion in Relation to Social Behaviour* (1962), holds to the view that animals regulate their own numbers by various types of behaviour evolved through group selection.

Specialized works: P. DEBACH and E.I. SCHLINGER, (eds.), *Biological Control of Insect Pests and Weeds* (1964), includes excellent examples of insect pests controlled by other insects and plant pests controlled by insects; G.F. GAUSE, *The Struggle for Existence* (1934, reprinted 1969), a classic discussion of the first laboratory experiments with protozoa in the study of predator–prey interactions and competitive exclusion; L.B. KEITH, *Wildlife's Ten-Year Cycle* (1963), a discussion of the regular 10-year cycle of various northern mammals and birds; D.L. LACK, *Population Studies of Birds* (1966), summarizes and relates many recent studies by a number of authors; T.T. MACAN (ed.), "Factors That Regulate the Sizes of Natural Populations in Fresh Water," *Communs. Int. Assn. Theor. Appl. Limnol.*, No. 13 (1965), includes papers on density-dependent limitation of sockeye salmon and on factors limiting flatworm populations; E.S. RUSSELL, *The Overfishing Problem* (1942), a short introduction to the study of fish populations in relation to commercial fishing; R.M.F.S. SADLEIR, *The Ecology of Reproduction in Wild and Domestic Mammals* (1969), includes information on litter-size and age of breeding; T.R.E. SOUTHWOOD (ed.), *Insect Abundance* (1968), includes the study of the winter moth; A. WATSON (ed.), *Animal Populations in Relation to Their Food Resources* (1970), recent reviews of the influence of food on numbers, in many types of animals.

(D.L.L.)

Populations, Human

Human populations can be studied on the basis of nation, language, religion, degree of economic or political complexity, or number of inhabitants per square mile, but the most basic approach involves analysis of population differences in biological composition. Usually designated as race, the biological variation among human populations is one aspect of man's relation to the rest of the animal kingdom. Wherever its members may live now, each race was once a geographical unit, a collection of genetically related local populations, differing from other geographical units of the habitable parts of the earth in heritable biological characteristics, including visible ones, such as skin colour, hair form, facial features, and body proportions, and invisible ones, such as the numerous blood groups now known. Unlike the visible characteristics, which can be expressed by a series of measurements, the latter cannot be expressed statistically in terms of overall biometric statistics, by ranges, means, standard deviations, standard errors, and the like; and thus they are not morphological (having to do with form or structure). They are rather expressed in terms of calculated gene frequencies, something that cannot yet be done with polygenic criteria (which are based on the interplay of two or more genes), such as the crown diameters of the teeth. To this uncertainty may be added individual differences influenced by growth, aging, and use, such as the breadth of the human face. Current limitations on the ability to calculate gene frequencies, however, force anthropologists to use morphological techniques in order to study human variation.

Race as a geographical unit

Regardless of the methods used to study them, however, races have long existed in time and space. Without a knowledge of human variations in time, the concept of race becomes two-dimensional, lacking depth and impossible to understand clearly. Because virtually all of the known remains of ancient humans lack soft parts, it is not possible to do blood-typing or other tests to determine the genetic composition of prehistoric populations. The specimens that must be used to reconstruct the evolution of living races are fossils consisting of bones and teeth, from which only morphological information can be gained. Modern population genetics is a welcome addition to the store of knowledge and methodology used in the study of human populations, but it cannot completely replace morphology because of the impossibility of carrying out genetic studies of human populations of times past.

Evolution and genetic variability

Evolution involves the interaction of the forces of mutation, variation, and selection. At any given moment its end product is a change in the frequencies of genes in a given population, and this change constitutes adaptation to either or both of the twin aspects of environment, the physical and the behavioural. For human beings, behavioural selection is more or less synonymous with cultural selection.

In man, the evolutionary result of behavioural patterns (behavioural selection) primarily concerns success in reproduction, which may vary greatly between individuals, particularly in societies in which men may marry several wives. This was probably a common custom in the period when modern human races were being formed, about 10,000 years ago, when all men were still hunters and gatherers, living in bands of between 20 and 100 persons, probably averaging close to 40. Population densities in such eras were less than one person per square mile. This projection is made by studying the ways of life of hunting peoples who survived into the 19th and 20th centuries in scattered parts of the world. Among such hunters each band seeks its food within its own territory.

Although the mating procedures of human groups 10,000 years ago probably varied, it is probable that most or all forbade sexual relations between parents and children and between brothers and sisters. By extension of this kinship relation, first cousins were sometimes also sexually out-of-bounds, either on both sides of the family, if descent were traced through both father and mother, or on one side only, if descent were traced through a single parent. The application of these incest rules plus the small sizes of the bands usually made exogamous marriages (marriages outside the group) necessary. The usual breeding unit was thus a number of neighbouring bands that habitually exchanged marriage partners and occupied one more or less unified geographical area containing several food-gathering territories.

The members of such a cluster of contiguous bands spoke their own language, observed their local customs, and guarded their sources of food from trespass by others. Strong cultural ties developed and bound them together and set them apart from outsiders. Within such a unit everyone was, from a modern point of view, quite closely related to everyone else. Under these circumstances local genetic combinations could vary considerably over short geographical distances.

Culture, kinship, and human variability

Behavioural selection also took place on an individual basis. A poor hunter, unable to bring in much game, might not be able to obtain a wife, and the genes that made him an economic liability would thus be reduced in the local gene pool. An outstanding hunter and leader in counsel and war might acquire several wives and beget more than his share of children. His favourable genetic characteristics and those of the women who shared his favours would be disproportionately high in the next generation and might thus contribute to an overall genetic change, including the incidental proliferation of other genes unrelated to procreative success (fitness).

THE PREHISTORIC BASIS OF HUMAN RACES

The primary breeding grounds of early man. During the late Pleistocene (the last geological epoch before the present one, lasting from about 2,000,000 years ago to about 11,000 years ago) many bands and groups of related bands of hunters occupied the habitable parts of the world's primary geographical regions, known to zoologists as the Ethiopian, Palearctic, Oriental, Australian, Nearctic, and Neotropical regions. When modern human races were first evolving, toward the end of the Pleistocene, only the first three regions were inhabited by human beings. Those of the New World (Nearctic and Neotropical) and the Australian were separated from the continental landmasses of the Old World and its islands by Wallace's Line (the observed dividing line between Asian and Australian faunal distribution) in the East Indies, by the cold landmass of Eurasia north of the winter frost line, and by the Bering Strait (Figure 1).

From *The Living Races of Man* by Carleton S. Coon, copyright © 1965 by Carleton S. Coon; reprinted by permission of Alfred A. Knopf, Inc.

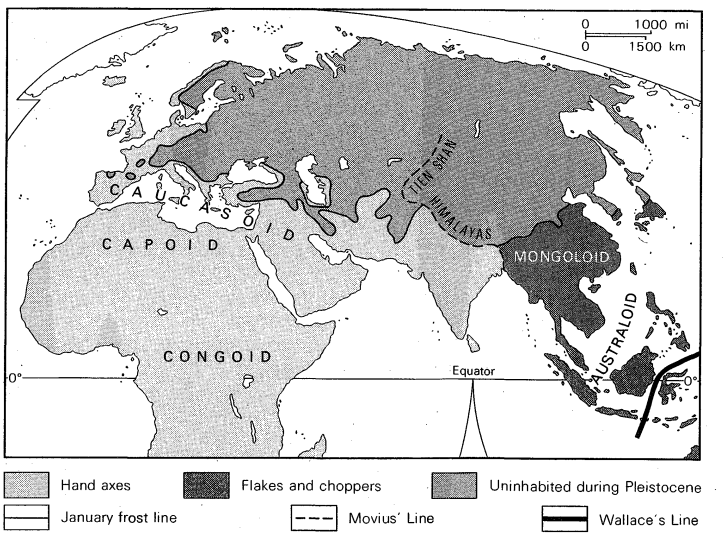


Figure 1: Distribution of human groups through habitable regions of the world in the late Pleistocene.

The geographical range of early man

For ancient peoples still lacking adequate environmental protection, the area of Europe habitable in winter was a stretch of land along the Atlantic, Mediterranean, and Black Sea coasts, tempered by the Gulf Stream and by large inland bodies of water. Nearly all of central and eastern Europe and of central and northern Asia were

too cold to support human life and so acted as a barrier to migration. Pleistocene people lived in parts of Anatolia, all of Syria, Palestine, and Arabia, and parts of Iran and Afghanistan. Nearly all of Pakistan and India were habitable, as were Southeast Asia, most of Indonesia, and much of eastern China. The entire continent of Africa was warm enough to be habitable, but water created certain barriers. Before the Nile had broken through to the Mediterranean, a band of shallow lakes stretched across the southern border of the Sahara (then a fertile plain). The Congo River Basin was also flooded until after the end of the Pleistocene.

The Palearctic, African, and Oriental faunal regions are thus narrowed down into six major breeding grounds of early man, each more or less isolated. One was in Europe and western Asia, a second in Africa north of the sub-Saharan water barrier, a third south of it and excluding the Congo River Basin, a fourth on the Indian subcontinent, a fifth in Southeast Asia and Indonesia, and a sixth in eastern China. Paleontologists have discovered early fossil remains in all of them except India, where geological conditions are generally unfavourable for the preservation of bone. Fossil remains in the other five indicate that each area contained a morphologically distinct human population of its own. Some may have contained more than one such population, and what kind or kinds of men lived in India is not yet known. Science may have one clue in stone tools. In all of the inhabited regions, west of Movius' Line (see Figure 1), where hand axes were made, stone-tool cultures were similar. In the eastern regions a different series of stone-tool cultures is found, based on flakes and choppers rather than hand axes. In India the two overlap.

The braincases, facial bones, and teeth of each of the five known major geographical populations differed in varying degree from the others, and these differences are reflected to a certain extent in the corresponding bones and teeth of peoples still living in the regions or in more recently settled regions inhabited by migrants from the originally inhabited areas. Available evidence thus indicates that some of the basic morphological differences among living human races (which are greater than those between geographical races of most other mammalian species) are very old and that the living races of man may not, as some anthropologists believe, have arisen from a single parental population some 30,000 years ago and then have become rapidly differentiated into the living races.

The major divisions of mankind during the Pleistocene. The Pleistocene human populations of the six major breeding grounds of the Old World are represented by a number of individual skeletons or parts thereof, not all of which have been adequately described in published articles. Most are cranial fragments and mandibles, with or without teeth. The number of fossil remains are approximately as follows: Europe, 200 individuals; western Asia, 40; North Africa and the Sahara, 15; sub-Saharan Africa, four; India, none; Southeast Asia and Indonesia (mostly from Java), 30; Australia, three; North America, three.

Grade and line variation

These specimens vary in two ways, known to paleontologists as grade and line variations. The former concerns the stage of evolutionary development of the individual as seen in progressive increases in brain size and progressive decreases in bony struts on the braincase generally related to jaw musculature. Browridges and nuchal crests (bony ridges at the back of the skull that anchors neck musculature) and, to a lesser extent, teeth grow smaller, and a chin emerges in specimens of more advanced grade. Since all races have gone through such changes, they are not really racial criteria, except in the limited sense that some groups show these "advanced" characteristics in more extreme form than others.

Most authorities postulate two sequential grades, called *Homo erectus* and *Homo sapiens* (*qq.v.*). Others favour the insertion between them of an intermediate stage, that of *Homo neanderthalensis*, though the label is a misnomer in that it tends to confuse the European and western Asian Neanderthals, a local population or race, with populations in other parts of the world, who were probably unrelated to the real Neanderthals, whose morphological features were derived in part from the fact that they had rickets.

Differences in line are racial. While living races differ in many cranial and dental details, the easiest to trace in the available fragmentary fossil material are the degree of facial flatness found in the region of the eyes and nose and the relative size of the front teeth compared to that of the cheek teeth (incisors and canines versus premolars and molars).

The faces of the European and western Asian skulls are the least flattened and have the most prominent nasal skeletons. The Chinese and, apparently, the North African skulls have the flattest faces. In the Chinese specimens the front teeth are large compared to the cheek teeth, and in the Javanese specimens the reverse is true. Chinese incisor teeth are particularly noted for their concave lingual (back) surfaces (shovel-shaped incisors). These differences are seen also in living populations from the same region, which thus seem to be descended, at least in part, from these fossil types of an earlier evolutionary grade.

Fossil evidence of racial distinctions

Because of these apparent continuities, fossil specimens from Europe and western Asia have been called Caucasoid; those from Java, Borneo, and Australia itself have been considered Australoid; those from eastern Asia and North America are termed Mongoloid; those from North Africa and the Sahara may be called Capoid (a term applied to the living Bushmen and Hottentots, who resemble the Pleistocene North African fossils in skull morphology); and those from southern Africa from Tanzania southward may be classed as Congoid. In southern Africa the Congoid peoples apparently preceded the Bushmen. The term Negroid (which usually includes Capoid and Congoid) is avoided for Negroes proper because it fails to include the Pygmies, who also belong in it, and Negroid also had been commonly used to imply such features as spiral hair and black skins, shared by Africans with Melanesians, Negritos, and other Asiatic and Oceanic peoples genetically unrelated to African Negroes.

As far as anthropologists can tell, the Pleistocene distribution of these five major geographical races of man were as indicated in Figure 1.

Prehistoric racial movements. During the latter part of the Würm, or Wisconsian (last), glaciation (from about 30,000 years ago to about 10,000) continental ice caps immobilized enough of the earth's water to lower the oceans more than 300 feet below their present levels. At that time only a narrow strait at Wallace's Line separated the Sunda and the Sahul shelves, facilitating human migration from Southeast Asia and Indonesia into Australia, New Guinea, and Tasmania. The Bering Strait was a flat plain several hundred miles wide, over which the ancestors of the earliest American Indians were able to enter North and South America, hitherto empty of human habitation. Meanwhile, the Nile had broken a channel through what are now its cataracts to empty into the Mediterranean Sea, thus draining most of the sub-Saharan lakes and swamps that had formed a water barrier between the central and western Sahara and tropical Africa. Considerably later, the Congo River eroded its way through Livingstone Falls to the Atlantic, draining its basin, into which the Pygmies later migrated, presumably from the swamps to the north.

Expansion of human populations with the introduction of agriculture

At the end of the Pleistocene, while the ice caps were shrinking and the seas rising, the previously glaciated regions of Europe, Asia, and North America were first occupied, by Caucasoids, Asiatic Mongoloids, and American Indians, respectively. By that time or soon afterward, some peoples in both hemispheres had begun to grow crops and to domesticate animals, allowing an increase in population density and causing expansion and migration. Mongoloid peoples moved down from China into Southeast Asia and Indonesia, displacing some of the indigenous Australoids, who took refuge in the mountains and islands. Capoids travelled through the eastern African highlands from the Sahara to South Africa, while European and western Asian hunters invaded North Africa,

to become the ancestors of the Berbers. Dravidian-speaking Caucasoids carried Western agriculture to India, while Munda-speaking peoples, part Australoid and part Mongoloid, introduced Southeast Asian plants to eastern parts of India. Agricultural people had settled numerous islands of the Pacific and Indian Oceans, notably Polynesia, Micronesia, Melanesia, and Madagascar.

By the time that Columbus and Magellan had completed their intercontinental voyages, the racial composition of the world's human population was already much more complex than it had been during the Pleistocene but less so than it has become since. The colonizations that followed the Age of Discovery gave birth to the intricate racial mosaics that now exist, and this distribution would have been difficult to understand without some idea of racial prehistory and early history.

THE LIVING RACES

Now that most of the world's populations have been described by human biologists, the five original geographical races, plus the American Indians, can be divided into several living components. Within its own extended region, each geographical race shows variations in body size, body proportions, head and face form, pigmentation, hair form, and body-hair distribution. These variations within a single geographical race are comparable in magnitude to those found within entire species of other animals across their geographical ranges. Among such geographical variations are differences in gross body size in response to temperature gradients also seen in other animals. According to Bergmann's Rule, in warm regions body size, in terms of weight, tends toward a minimum; it increases as temperature falls to a certain critical threshold and decreases again in regions of greater cold. According to Allen's Rule, body proportions also vary climatically. Lean trunks and long, slender extremities may be observed in arid regions, where greater skin-surface area in proportion to body weight may facilitate heat loss; stockier builds with relatively short extremities characterize the inhabitants of damp and cold places.

Extremes of variation in body size and hair form

Also, the Australoids and Congoids are divided between full-sized and pygmy populations. Among other animals, dwarf races, such as the pygmy hippopotamus and the pygmy chimpanzee, are often considered by zoologists to be separate species. The Australoids are further divided into populations having wavy hair and those having spiral hair. The extreme variation in hair form in man, with the coarse, straight hair of the Mongoloid at one extreme and the finely coiled, peppercorn hair of the Bushmen at the other, also exceeds the range found in nonhuman mammals, including the rest of the primates.

Races also differ physiologically and anatomically from one another. These differences involve the blood (and blood vessels), the endocrine hormones, the nervous system, and, to a certain extent, behaviour. Blood supplies the body with nutrients and oxygen, prevents certain infectious diseases, and helps maintain the heat balance of the body. It also humidifies and warms air entering the lungs through the nose. In deserts, mountainous regions, and the Arctic, a long, narrow nose serves as a better humidifier than a short, broad one, which is adequate in damp tropics. Since all of these functions are closely concerned with environmental protection, and each geographical race occupies a range of environments, certain parallels may be seen between those populations of most of the human races that occupy similar environments.

Physiological and cultural adaptations

Physiological adaptation to damp heat comes mainly from dense skin pigment, which absorbs solar heat and activates the sweat glands, cooling the air immediately around the skin. Adaptation to moderate cold comes either from a high basal metabolism with a heavy consumption of fat or from reduced circulation in the extremities in order to preserve heat needed for the more sensitive organs of the abdomen and for the brain. Adaptation to extreme cold is mostly cultural, through the use of warm clothing and housing, plus a heavy fat consumption. Adaptation to high altitude requires a large volume of blood and large lungs, supplying extra oxygen to compensate for the thinness of the atmosphere.

A yellowish or light-brown skin reflects light most efficiently in regions of intense sunlight, such as deserts and snowfields. Persons with fair skin may suffer from an overdose of vitamin D_2, particularly in deserts, while those with black skins may become overheated. In less brightly lit regions the latter may absorb too little vitamin D_2 to prevent rickets without dietary compensation.

Cultural behaviour is again responsible for differences within segments of populations living in urban civilizations. There a surplus of food, combined with the need for the skilled services of full-time specialists, has allowed for the rise of metalworkers, scribes, clergy, and the like. Craftsmen, scholars, and administrators tend to marry within their own groups, and through the generations their offspring are selected for their special capacities. On the other hand, a relaxation of selection also occurs, which may transmit such traits as myopia and colour-blindness, which would be selected out as disadvantageous among hunters. Individuals who migrate from farms to cities also have been shown to be different in some biological respects from those who remained behind, and comparable differences have been observed between migrants to new continents and those who chose to stay at home. The flood of scientists and scholars from despotic regimes in Europe to the New World exemplifies this principle, as does the more recent "brain drain."

Blood and race. In addition to morphological data, there are now broad geographical data on blood groups and factors, except from Arabia, China, and Siberia, where little information has yet been obtained. These blood groups and factors include the so-called transfusion (ABO) group, the Rhesus (Rh) group, the MNSs group, and numerous other factors, including variant hemoglobins and haptoglobins (for discussion of the various human blood groups, see BLOOD GROUPS). These conform more or less to the geographical distribution of morphological races, although no genetic association has yet been demonstrated between any blood factor and any somatic criterion. The transfusion group blood types have been shown to have a relationship to certain diseases. For example, blood type A resists smallpox, and B resists various cancers and possibly plague, and both A and B alternately resist different strains of *Escherichia coli*, which may cause infant diarrhea. Both Rh and transfusion-group incompatibilities can cause spontaneous abortion as a result of a reaction to the fetal blood as foreign. Sickle-cell anemia and thalassemia, genetic diseases involving abnormal hemoglobins, help affected individuals resist malaria.

Distribution of blood-type frequencies

Caucasoids and Congoids have two A factors in the transfusion blood group, A_1 and A_2, and both have the Rh-negative blood type. Both of these are almost or completely lacking in other races. Blood type O is highest in the populations of lands first occupied during the late Pleistocene (*i.e.*, Australia and the Americas, except for the Eskimo), possibly because of negative selection due to the absence of those diseases against which blood types A and B offer protection. Blood type M in the MNSs system is at its peak in the Americas, type N in Australia. The Pygmies have a special haptoglobin factor and the Bushmen a special hemoglobin factor, both of which other Africans lack. On the whole, European, western Asian, and African populations present an overall constellation of serological factors, with many local variations; eastern Asia, Australia, Oceania, and the Americas present another. Like morphology, serology divides the native populations of the world into two broad geographical realms, western and eastern.

Fingerprints and race. Like blood groups, fingerprints vary racially in frequency of types and are inherited. They are probably not influenced by environment, since their development is complete by the 18th week of fetal life. There are three basic fingerprint patterns: arches, loops, and whorls (see Figure 2). In general, Caucasoid populations have between 1% and 10% of arches, Negroes between 3% and 12%, and Pygmies and Bushmen between 10% and 16%. All other populations studied, except the Ainu of Hokkaido, Japan, have 5% or less. As for loops, Europeans, Africans, and Ainu range from

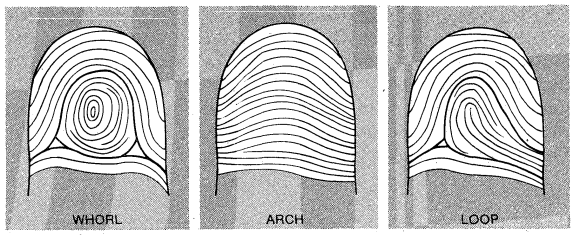


Figure 2: Basic types of fingerprints.
From H. Cummins and C. Midlo, *Fingerprints, Palms, and Soles: An Introduction to Dermatoglyphics* (1961); Dover Publications

52% and 76%, and all others range from 28% to 64%. In whorls, the Europeans, Africans, and Ainu range from 16% to 42%. Western populations exceed the eastern in arches and loops, and within the western group the Pygmies and Bushmen distinguish themselves from Negroes; eastern peoples, again excepting the Ainu, exceed westerners in percentage of whorls. The division of the world on this basis parallels that of the distribution of blood-group factors.

CAUCASOIDS

The Europeans. Modern Europeans are descended in part from the Pleistocene inhabitants of Europe and in part from migrants who have entered it from western and central Asia in more recent times. Most of them speak Indo-European languages, except the Finns, Lapps, Esths, Magyars, and others, who speak Finno-Ugric languages; and the Turks, Tatars, and (until recently) the Mongolian-speaking Kalmyks, who speak Altaic languages. Most Europeans are (or were) Christians, but there are Muslims in Turkey, Yugoslavia, and Albania; the Kalmyks were Buddhists; and there are people of Jewish faith in most European countries.

In 1899 the American economist W.Z. Ripley divided European populations into three main "races," the Nordic, Alpine, and Mediterranean. Later writers postulated more numerous divisions. As no rigid taxonomy of European populations on a subracial basis can be biologically justified, one can speak only in a broad sense of northwestern Europeans (Ripley's Nordics), southern Europeans (his Mediterraneans), central Europeans (his Alpines), and eastern Europeans (whom he included under Alpines).

Physical characteristics of Europeans

On the whole, the Europeans have the fairest skin in the world and the greatest frequencies of blond hair and light eyes. Depigmented skin incapable of tanning is concentrated around the North Sea and in Ireland. Blue eyes have the same general distribution, being most prevalent in Ireland. Blond hair has its centre farther to the east, around the Baltic, where it is commonly accompanied by gray eyes and fair skin capable of tanning. The darkest pigmentation of skin, hair, and eyes in Europe is found in Portugal, the countries bordering the Mediterranean, and in some of the Balkan countries, particularly Romania. Hair varies from curly to wavy to straight and is usually fine to medium-coarse. The curliest hair is found in the Mediterranean region. Beards and body hair (in males) are most abundant in central Europe.

The tallest European populations live near the winter frost line (32°F [0°C]), particularly in Scandinavia, Scotland, Yugoslavia, and Albania. The shortest are the Lapps, other Arctic Europeans, and some of the inhabitants of the Mediterranean regions, including Portugal. On the whole, Europeans have relatively long trunks in comparison to limb length, and the stockiest builds are found in central and eastern Europe. The Europeans also vary considerably in head form. Dolichocephaly (in which the head is long from back to front) is commonest in Scandinavia, the British Isles, Portugal, Spain, and southern Italy and its fringing islands. Brachycephaly (in which the head is comparatively short from back to front) is prevalent in central and eastern Europe. In some Balkan countries, notably Crna Gora (Montenegro) and Albania, an exaggerated degree of brachycephaly is produced by cradling infants in wooden cradles and thus flattening the back of the head (occiput).

European blood-type distribution

In the transfusion blood group, blood type O is particularly prevalent in western Ireland, northern Wales, the Scottish highlands, and Iceland. In this respect the Icelanders resemble the Irish more than they do the Norwegians. Blood type A exceeds type B in western Europe, and type B rises to exceed A in moving eastward from the Gulf of Bothnia to the Urals. In Europe, type A is divided between types A_1 and A_2, and A_2 reaches a world peak of 35 percent among the Lapps, three times higher than anywhere else. At a polar extreme in another respect are the Basques, who share with some Moroccan Berbers the world's highest frequency of the Rh-negative gene (Rh blood-group system), reaching over 50 percent. On these bases some serologists have postulated that the Lapps and Basques represent two fundamental racial components of the Europeans. Sardinia, too, has a unique population of Mediterranean morphological type, coupled with high frequencies of blood types O (transfusion system) and M (MNSs system) (75 percent each) and the lowest Rh-negative in Europe except for the Lapps. In parts of southern Europe and among European Jews, the so-called African Rhesus gene rises to 5 percent, higher than among other Europeans.

The western Asians. The peoples of western Asia may be regarded as an eastern extension of the European population, or the Europeans as a western extension of the western Asian Caucasoids, for there have been population movements in both directions. As in the Alpine zone of Europe, there are populations of stocky build, with abundant beard and body hair, including convergent eyebrows, in the mountains and plateaus of the north and lean, Mediterranean-like peoples in the deserts to the south and on the Yemen plateau. There is no northern tier of light-pigmented populations as in Europe, although blondism occurs sporadically among the Turks of the western shore of the Aegean, the Kurds, some of the Caucasian mountaineers (notably the Cherkess, or Circassians), northern Armenians, and the Kaffirs (Nuristanis) of northeastern Afghanistan and northwestern Pakistan. The Anatolian Turks show little physical evidence of their partly Mongoloid origin. The mountaineers who speak the Caucasian languages are closely similar in fingerprint patterns to the Basques, also linguistic relics.

Physical characteristics of western Asian peoples

The Arabs of the deserts and of the Yemen plateau are mostly lean in build and long headed, with narrow, often aquiline noses. They are less hairy than the mountain peoples to the north and characteristically brunet in colouring, although about 25 percent have mixed iris colours (brown with rays or zones of blue, gray, or green). There is a strong Negro element on the Yemen coast. The Persians of the Iranian plateau are mainly brunet and European in appearance, while the nomadic Baluchis of the south and the Pathans of Afghanistan are tall, lean, and often aquiline. The Tadzhik of northern Afghanistan are Persians physically as well as linguistically, while the Ḥazāra of the northern mountains are Mongols, remnants of an invasion, like the Kalmyks of southeastern Russia.

Serologically, most of the western Asian peoples have been incompletely studied or not studied at all. In the mountain areas, blood type B exceeds A in the transfusion system, as in eastern Europe, and the Rh-negative blood type is as common as in most of Europe. Both the Arabs and Persians are high in blood type O. On the whole, the blood-type distribution appears to consist of variations on a European theme.

North African Caucasoids. Four different kinds of peoples are native to Africa: Caucasoids, Pygmies, Capoids (mostly Bushmen and Hottentots), and Negroes, as well as many mixed populations. The Caucasoids apparently entered Africa from both Europe and western Asia, toward the end of the Pleistocene and later. In northwest Africa and the Sahara live the Caucasoid Berbers, now geographically separated into discrete groups by Arabs whose ancestors came from Arabia in two principal waves, during the 7th and 11th centuries AD. The first wave consisted largely of urban and agricultural peoples, many from Yemen. They built cities and absorbed some

of the Berbers, Romanized settlers, and Jews, including Judaized Berbers. The members of the second wave were Bedouins from the Syrian Desert (Bādiyat Shām), who continued their ancestral way of life, denuding the North African landscape and mixing little with the earlier inhabitants, who were mostly Berbers.

Berbers

Most of the Berbers, like the southern Europeans, could be classed as Mediterraneans, but some groups in the Moroccan Rif, middle Atlas Mountains, and the Jebel Nefusa of northwestern Libya include many individuals with fair skin, blue eyes, and light hair, particularly in their beards. The blood types of these Berbers are like those of northwestern Europeans, but in some tribes of the Moroccan middle Atlas the Rh-negative gene reaches a frequency elsewhere found only among Basques: 55 percent in one tribe. The same peoples are very high in blood type O.

The nomadic Arabs differ in no essential way from their kinsmen in the Syrian Desert. In Libya, which is mostly desert and oasis, there is a visible Negroid element in the sedentary populations, and the same is true of the Fellahin of Egypt, whether Copt or Muslim. Osteological studies have shown that the Negroid element was stronger in predynastic times than at present, reflecting an early movement northward along the banks of the Nile, which were then heavily forested.

CONGOIDS AND CAPOIDS

Congoids: the Pygmies. The Pygmies, who number about 150,000, now live in various enclaves in the tropical rain forest of Africa, between latitudes 5°N and 5°S, from Gabon to Lac Kivu and the slopes of the Ruwenzori Mountains; most of them, however, live in the Congo River Basin. This could hardly have been their original home, because it was underwater until some 7,000 years ago. There is historical evidence that Pygmies also once inhabited the swamps of the White Nile as well as the forests of West Africa. Living Pygmies, who are hunters, live in enclaves separated by agricultural Negroes, who entered the forest in recent times.

Physical characteristics of Pygmies

Most of the male Pygmies average five feet (150 centimetres) or less in height and about 70 pounds (32 kilograms) in weight; the women are shorter and lighter. Their legs are relatively short, their arms long, and their limb joints mobile. Running on fallen logs and grasping lianas, they can glide through dense forest more easily than larger and heavier people. Their babies are of normal size at birth, but experience no growth spurt, as seen in full-sized races, during adolescence. Their small size is not due to dietary deficiency, because they eat much protein, both meat and insects. No deficiency in growth hormone has yet been established. They are said to have the highest basal metabolic rate in the world.

At birth their skin is light, but it darkens later. Some in the west are yellowish as adults; those in the Ituri Forest are mahogany coloured; while those on the shores of Lac Kivu are nearly black and are taller and heavier than those living on lower ground. All of them have tightly coiled hair, the men are bearded, and some have heavy body hair. Their genitalia are not dwarfed but of normal size. Their eyes tend to bulge; their noses are short, concave, and broad; they have large teeth and alveolar prognathism (protrusion of the upper jaw).

Where Pygmies and Negroes mix, it is almost always in one direction: Negroes take Pygmy wives, but not the opposite. The Pygmies, thus, are essentially unmixed, though there is a Pygmy element in the Negro gene pool. Serologically, the Pygmies are high in both blood types A and B, the African Rh gene, and in the gene for sickle-cell anemia. They share these traits with Negroes. One blood factor has been found in Pygmies alone, Pyg-2.

The Pygmies vary among local populations in blood-type frequencies as in other respects. The Binga Pygmies of the west are closer to the Negroes, who have absorbed Pygmy genes, and the Mbuti of the Ituri Forest are closer to the Bushmen than the Binga and Mbuti are to each other. In fingerprints, too, the Binga resemble Negroes, while the Lake Kivu Pygmies and the Mbuti are high in arches and low in whorls, the Mbuti reaching a world extreme in the latter respect. These and other items of evidence suggest that at least two and possibly three kinds of Pygmies reflect separate episodes of dwarfing and that the western Pygmies have contributed the most to the Forest Negro gene pool, while the eastern Pygmies are in some way related to the Capoids.

Congoids: the Negroes. The Negroes are now the principal inhabitants of Africa, but their origins are little known. The only Pleistocene skeletal fossil material comes from Tanzania southward to the Cape of Good Hope, in which regions the Negroes apparently preceded the Bushmen. Remnants of old hunting-gathering Negro populations survive in southwest Africa (the Bergdama) and in southwestern Angola (the Kwáde). A clue to the origins of Negroes farther north may be found in the remains of a people who lived along the shores of Lake Rudolf (Kenya) as long as 9,000 years ago and harpooned hippopotamuses and large fish. Their characteristic implements have also been found in the Nile Valley as far north as Egypt and along the former rivers and lakes of the southern Sahara and Sudan and as far west as Sierra Leone. As early as 4000 BC, Negroes living in the Sudan began to cultivate local plants, and they later moved down the stretch of savanna that extends southward to Dahomey and Togo. The introduction of Indonesian plants—taro, yams, and plantains—about the beginning of the first millenium AD facilitated forest cultivation in former Pygmy country, and the later introduction of American cultivated plants, including maize (corn) and American beans, allowed further expansion into central Africa and then East Africa, where the ancestors of the living Bantu speakers of East and South Africa acquired cattle, multiplied rapidly, and thence expanded into South Africa, arriving there mostly after the European colonists.

Physical characteristics of Negroes

Though the physical characteristics of Negroes vary from region to region, they generally are spiral haired, have beards but little body hair, and are dark skinned. Stature varies regionally, the tallest being in the White Nile country and the shortest those in the Congo Basin who have mixed with Pygmies recently. Both arms and legs are usually long in proportion to trunk length, and the muscles of their extremities have long tendons and short bodies. The soles of their feet usually have a thick pad of fibrous fat, which causes the erroneous impression that their heel bones are elongated. Facial features are variable, but everted lips are common, as is prognathism, and their teeth are characteristically large. Anatomical as well as archaeological evidence suggests some early mixture with Caucasoids, particularly in East Africa among the ancestors of the Amharas, Tigre, and Tigrinya and those of the Gallas, Somalis, and other Hamitic-speaking peoples from the west, long before the invasions of the Horn of Africa by pre-Islamic south Arabians.

Serologically, Negroes resemble Europeans in having high frequencies of blood types A_2 and Rh-negative (the latter, 20–25 percent), but they differ in having about 60 percent of the African Rh blood type (*cDe*), the world's highest percentage of blood type P_1 in the P blood-group system (over 80 percent), and high frequencies of the abnormal hemoglobin that gives rise to sickle-cell anemia (hemoglobin S) in regions of chronic malaria, particularly where the Negroes have mixed most with Pygmies. While they are a distinct race, the Negroes have also derived genetic elements from the three other races of Africa: Caucasoids, Pygmies, and Capoids.

Capoids: the Bushmen and Hottentots. Aside from two relic groups in northern Tanzania, the Hadza and Sandawe, the living Capoids consist of the Bushmen hunters and Hottentot pastoralists of southern Africa. Bushmen today number some 55,000 in Botswana, southwest Africa, and southwestern Angola; and there are about 35,000 Hottentots, some mixed, in the northern Cape Province of South Africa and in southwest Africa. In 1500 the two peoples shared Africa south of the Zambezi and must have numbered in the millions.

Early Capoid skeletons from the Sahara, Sudan, and East Africa are those of full-sized people; the living Bushman males are about five feet tall and the Hottentots

some four inches taller. The Bushmen have peppercorn hair, growing in tufts on their heads; beards and body hair are absent or minimal. Their skins are yellowish, deeply wrinkled in old age; their ears are characteristically small, with straight upper borders, and lobeless. An epicanthic fold (Mongoloid eyefold) over the upper eyelid is common. Their limb proportions are not Negroid, and their arm and leg muscles have long bellies and short tendons. Their hands and feet are small. Fat-padded buttocks (steatopygia) are particularly developed among the females, who also have elongated vaginal lips, partly genetic in origin and partly developed by manipulation. Morphologically, the Hottentots resemble the Bushmen in most respects, but the peculiarities of the Bushman buttocks and genitalia are even more pronounced in the Hottentots, and many Hottentot men even lack axillary (armpit) and pubic hair.

Bushmen

The Bushmen have low frequencies of both A and B blood-type genes, lack the Rh-negative type, and have more of the African Rh gene (*cDe*; 89 percent) than nearly all Negroes. They also have a 40 percent frequency of the gamma globulin gene HL-A, elsewhere found only among Asiatic Mongoloids and Ainu. The Hottentots also have two cytogenetic peculiarities not yet studied in the Bushmen. Their Y-chromosomes are unusually long, like those of the Japanese and unlike those of the Caucasoids and Negroes. Twenty-four percent of a sample studied have fewer and 23 percent have more than 46 chromosomes, the standard human number.

AUSTRALOID AND OCEANIC PEOPLES

Australoids: The Tasmanians and Australian Aborigines. At the time of European discovery, Australia had about 300,000 native inhabitants, Tasmania about 4,000. In Australia the Aborigines were concentrated near the coast, particularly in the east and southeast and in the Murray River Valley. In Tasmania the inland bands used to visit the coast periodically for shellfish.

The Tasmanians, who became extinct except in mixed forms in 1876, were black skinned, woolly haired, and prognathous (with projecting jaws), and most of them had receding chins, deep nasal notches, and broad noses. The majority were of medium to short stature, though a few chiefs stood six feet tall. There is no direct information about their blood types available. They had probably been isolated from the Australian continent since about 7000 BC by the sinking of the land bridge across Bass Strait.

Australian Aborigines

The Australian Aborigines resemble the Tasmanians physically in many ways, but most of them are taller, the men averaging about five feet seven inches (170 centimetres) and the women about four inches less. As a rule they have straight or wavy hair, but woolly hair is found among some individuals, particularly along the coast. The men have full beard growth, and some have heavy body hair, and their hair turns gray early in adult life. On the densely wooded Atherton Plateau in northern Queensland, a local population has short, almost Pygmy-like stature and woolly hair. In other respects the Aborigines also vary regionally to a certain extent. The northerners, the tallest, are relatively lean; those in the southwest tend to be stockier and hairier. The desert people are lean, and some have convex nasal profiles. Blond hair, along with light-brown skin, is quite common.

Except in the north, where Indonesian contact has existed for several centuries, Australians have no blood type B. Type A averages 17 percent and is highest in the south. With figures reaching over 70 percent, their M blood type (MNSs blood group) is the most numerous in the world (it reaches similar frequencies in New Guinea), and the percentage of whorls in their fingerprints reaches a global maximum.

Archaeologists now believe that Australia has been inhabited since at least 26,000 BC, and the few fossil skulls yet found are all more massive than those of most living Aborigines, many of whom still have heavy browridges. Either this is representative of evolutionary processes in Australia or there have been multiple invasions by different peoples or both. In support of multiple invasions are the late appearance of the dingo (which could only have been introduced by man) no earlier than 5000 to 7000 BC, and the marginal distribution of woolly hair, possibly a characteristic of the first wave of migration. The ancestors of the Tasmanians and Australians could only have arrived there from Indonesia and Southeast Asia, where the Australoid racial line initially evolved.

Australoids: the Papuans. The Papuans proper are the indigenous peoples of New Guinea, living mostly in the interior but reaching the coast in the west and south; the rest of the coastline is occupied by Melanesian-speaking invaders. The Papuans speak many languages of their own, and comparable languages are spoken on Sumbawa, Halmahera, and Timor in Indonesia and in small enclaves on the Admiralty Islands, the Bismarck Archipelago, the Louisades, Solomon Islands, and New Hebrides. There may be 700 to 1,000 such languages, mostly in New Guinea.

The Papuans of New Guinea are Australoid, and some of them resemble Australian Aborigines in appearance, but all have woolly hair, brown skins (some of which are quite light), and in many cases high-bridged, aquiline noses. Like the Tasmanians and Australians, they have thick but mostly uneverted lips. Since the highlands of New Guinea are chilly and proteins are scarce, the natives are short, some falling to a mean male stature of five feet or lower, but they are not Pygmies; children from this region brought up on the coast reach normal stature. In the highland, blond individuals are not uncommon. As for their blood types, A is below 30 percent and B runs from 10 percent to 15 percent. Rh genes *C* and *e* both exceed 90 percent, and the commonest genotype is *CDe*.

Australoids: the Negritos. Small enclaves of Negritos, dark-skinned, spiral-haired peoples of reduced stature, comparable but genetically unrelated to African Pygmies, still survive in the Indonesian and Asiatic homelands of the Australoids. They include one element in the Kadar of the Cardamon Hills in Kerala State, India; the Andaman Islanders; the Semang of the northern mountains of the Malay Peninsula; the Aeta of the Philippines; and various groups in the Papuan-speaking parts of the islands of Sumbawa and Timor.

Andaman Islanders

Andaman Islanders include the Great Andamanese (now nearly extinct); the Jarawa, who live in the dense, hilly forests of South Andaman Island; and the Önge of Little Andaman. They are black skinned, frizzly haired, beardless, and steatopygous (having fat deposits on the buttocks), craniologically similar to African Negroes but with smaller teeth. The Önge, who alone have been studied serologically in adequate numbers, have over 50 percent of blood type A, and more B than O. Like the Papuans, they have 92 percent of Rh type *CDe*. Despite the prevalence of malaria, there is no sickle-cell anemia or related genetic abnormality. On the whole, the Andamanese may be genetically unique, descended from a few canoeloads of castaways landing on previously uninhabited islands.

The Semang of Malaysia

The Semang of the Malay Peninsula are also black skinned and spiral haired, but they are bearded, with a male mean stature of slightly over five feet. Their general morphology is Australoid. They are high in blood type O, with 15 percent of type A and 9 percent of type B and, uniquely for this part of the world, 27 percent of *cDe*, the African Rh type. The Philippine Negritos, remnants of the islands' pre-Indonesian population, live today in seven enclaves on the island of Luzon and one each on Palawan and Mindanao. Their skin colour varies from almost black to yellowish brown, their hair is tightly curled, and their facial features suggest those of some Australian Aborigines but usually lack excessive browridges. A few individuals are considerably taller than the others and more Australian or Tasmanian in appearance.

On the whole, the Negritos described above seem to represent dwarfed variants of the original Australoid population from which came the Tasmanians, the Papuans, and the first postulated wave of ancestral Australian Aborigines; the Andamanese are the most deviant.

Melanesians, Micronesians, and Polynesians. Melanesians inhabit much of coastal New Guinea, particularly

in the north, and on most of the islands stretching from the Bismarcks to New Caledonia and Fiji. According to a conventional classification, their languages belong to the Austronesian family, along with Indonesian, Micronesian, and Polynesian.

The Melanesians of the New Guinea coastal areas are darker skinned than the Papuans; many are actually black. Their facial features are more Negroid, and they usually lack the Papuan aquiline nose. On the Melanesian islands proper there is much local physical variation. The Solomon Islanders are the blackest, the New Caledonians brown-skinned, thick-set, and heavy-boned. Some have heavy browridges and jaws. Some have blond hair, partly from bleaching in seawater but probably also partly genetic. Individuals with pinkish or reddish skins are found on some of the islands, notably on Fergusson Island in the D'Entrecasteaux group, and in the interior of Malaita, where light skins may be combined with wavy or lightly curled hair. The origin of these wide variations in Melanesia is unknown.

The Micronesians contain a strong Melanesian element mixed with Mongoloid. Their hair ranges from frizzly to straight, and they are relatively light skinned but lack the blond element. Linguistically, Micronesian is closest to Melanesian, and the physical differences between the two peoples may perhaps be partly attributed to geographical selection related to the presence of malaria in Melanesia and its relative absence in Micronesia.

The Polynesians differ physically to a certain extent from island to island in stature, skin colour, head length, face form, and body build. On the whole they are taller than the Micronesians, more heavily built, and fatter. Some of the noblemen and royalty are of great size. Despite their wide geographical range, from Hawaii to Easter Island to New Zealand, they are physically more homogeneous than either the Melanesians or Micronesians.

Oceanic blood-type frequencies

In blood factors the Melanesians and Micronesians are very similar, with blood type A found in between 20 percent and 30 percent of the population and type B in between 10 percent and 20 percent. Blood type N exceeds M in the MNSs system, and in Micronesia Nc reaches 65 percent. The Rh *C* gene has an 80 percent to 90 percent frequency, most commonly in the *CDe* blood type. Most of the blood-group data in Polynesia come from Hawaii and New Zealand. Type A is higher than B, which is completely lacking on some islands; in general, A occurs in between 35 percent and 40 percent of the population but falls to 22 percent on Easter Island; type B ranges from zero to 5 percent. Among the New Zealand Maori the Rh blood type *CDe*, commonest elsewhere in Oceania, falls to below 50 percent, and *cDe* is elevated. This situation is paralleled only among the Indians of the American northwest coast and is partly responsible for the theory that the Polynesians may be of American Indian origin.

PEOPLES OF EAST ASIA

Northern Mongoloids. The Mongoloid peoples, not including the American Indians, account for over 71 percent of the world's population of 3,500,000,000. For present purposes they may be divided into two types, northern and southern, comprising about 64 percent and 7 percent, respectively, of the total 71 percent of the world's population. The northern group includes the Chinese, totalling about 2,000,000,000 people (not including Chinese overseas); the Koreans; the Tibetans; the Mongols; the Turkish-speaking peoples of central Asia; the Japanese; the natives of central and eastern Siberia; and the Eskimo.

Northern Mongoloid physical traits

The principal morphological traits of the Mongoloids include coarse, straight, black head hair; little beard and body hair; graying delayed until old age; yellowish skins ranging in hue from the equivalent of brunet white to dark brown; trunks relatively long compared to limb lengths; hands and feet usually small, with curved nails; flattish faces, with prominent cheekbones; noses either low bridged or aquiline; brown eyes, with epicanthic eyefolds present in varying percentages in different populations; large incisor teeth with some degree of shovelling (tongue side of tooth curves inward at the sides); and broad lower jaws. While their blood types are locally variable, A_2 is usually lacking, *CDe* is the most common Rh type, and Rh-negatives are mostly confined to China and Japan. Mongoloid fingerprints are high in whorls.

In China, statures range from quite tall in the north, including Manchuria, to short in the south, and the northern Chinese are relatively long headed, long faced and light skinned. Some skulls of Caucasoid appearance have been found in the royal burial pits of the Shang dynasty at An-yang, dated at about 1400 BC. There may be, thus, a Caucasoid element in the north Chinese, which shows up in their ABO frequencies, and especially in the "European" type, A_2.

Among Tibetans, aquiline noses are common, and skin colour is dark. The Mongols and their Siberian relatives, the Buryats, are mostly of medium stature and stocky build, with large, round, low-vaulted heads and medium-dark skin. The Turkish-speaking Kirgiz-Kazaks of central Asia are similar to the Mongols but taller, and many have aquiline profiles, while the Turkmen who live along the borders between the U.S.S.R., Afghanistan, and Iran are partly Caucasoid, usually tall and of rangy build and with variable facial features. Many are long faced and thin nosed; some are heavily bearded. They are apparently the mixed prototype of the Turkish invaders of Anatolia, whose descendants became essentially Caucasoid through mixture. The Koreans, whose language is related to that of the Tungus and, more remotely, to Turkish, are mostly medium to tall in height and robustly built, with brachycephalic heads (short from back to front) and long to medium faces.

The Japanese are characteristically of short stature (although the younger generation, whose diet includes meat and milk products, are taller), and the men can grow beards. Their facial features vary from broad, with concave nasal profiles, to long and narrow, with convex profiles, as seen among the nobility and geishas. The Siberian natives include the Yakuts, who are forest-dwelling Turks; the Evenki or Tungus (related to the Manchus); the Paleosiberian-speaking Yukaghirs, Koryaks, and Chukchi; and the Asiatic Eskimo, whose kinsmen live along the coast of North America as far as Labrador and Greenland. These people usually have long faces, narrow noses, and epicanthic folds. Their blood is high in type O, and B exceeds A, except in the American Eskimo, among whom A exceeds B, as it does in the Chinese, Koreans, and Japanese. The presence of type B differentiates the Eskimo from the American Indians, who have none.

Southern Mongoloids. The southern Mongoloids, as designated earlier, inhabit Vietnam, Laos, Cambodia, Thailand, Burma, Malaya, Indonesia, Brunei, and the Philippines and were the aboriginal inhabitants of Taiwan. The Malagasy Republic (formerly Madagascar) also is partly Indonesian racially.

The mountains of Southeast Asia run from north to south, and the mountain peoples are splintered into many ethnic and linguistic groups, while the lowland populations are more nearly uniform. The Lomwe, Miao-Yao, and Li live mostly in the north. Their ancestors came down from the mountains of southern China, where many of these tribesmen still live. The Thais, Laotians, and Mon-Khmer speakers (Cambodians) include both lowland and mountain groups, such as the Montagnards.

Physical characteristics of southern Mongoloids

The Southeast Asians are of short stature, and many of them are darker skinned than most Chinese; they also have a lower incidence of epicanthic eyefolds. The mountain peoples are taller than the plainsmen and have a higher incidence of convex nasal profiles. The Khmer of Cambodia are partly Australoid, with darker skin, broader noses, and more wavy or curly hair than most of the others. The Malay have mixed to a certain extent with Arabs from southern Arabia, the Javanese with invaders from India, and the Filipinos with Spaniards. There is also a Negrito strain in the inhabitants of the Outer Sundas and possibly an Ainu-like strain among the Taiwan Atayals. Among most of these peoples, blood type B exceeds A, except in the Sea Dayaks of Borneo and the northern Taiwan aborigines. The *CDe* blood type is the

commonest Rh combination, and Rh-negatives are absent or extremely rare.

Madagascar, the westernmost outlier of Indonesia in an ethnic sense, was first settled about the time of Christ by Indonesians who made use of monsoon winds to propel their boats across the Indian Ocean. The languages of Madagascar seem to be most closely related to languages spoken in Borneo, which was more densely populated before the epidemics of the 16th century. The Indonesians brought Negroes from eastern Africa to the western and least mountainous part of the island, and the result is a general mixture, with the Indonesian racial element most evident in the highlands and the African element most evident in the west and south, possibly because the Negroes are more resistant to malaria and less resistant to high altitude than the Mongoloids. Serologists have estimated that the inhabitants of Madagascar as a whole are about two-thirds African and one-third Indonesian in origin.

The Ainu. One of the great mysteries of physical anthropology is the racial affiliation of the Ainu, an apparently exotic Caucasoid people living by hunting and fishing on the outskirts of Mongoloid territory. Remnants of them still live on Hokkaido (Japan) and southern Sakhalin (U.S.S.R.). Until the 12th century AD some Ainu were still to be found on northern Honshū and, until 1873, on the Kuril Islands. Their long, wavy hair and full beards, body hair as abundant as on Armenians, light-brown (and in a few individuals greenish) eyes, brunet white skin colour, lack of an epicanthic fold except in mixture, broad faces, widely spaced eye sockets, and in most cases snubbed noses, have led to the anthropological cliché that Leo Tolstoy could have passed for an Ainu. The Eta, an outcaste class in Japan, resemble Ainu, at least in hairiness, and an Ainu-like element has been noted in the Ryukyu island population as well as among the Atayal aborigines of northern Taiwan.

Caucasoid features of the Ainu

Serologically, the Ainu communities are locally variable, but, on the whole, the gene for blood type A reaches 23 percent and that for B 28 percent, and there is no A_2. The Ainu resemble the Lapps in having 4 percent Rh-negative blood type and the world's highest incidence of the NS combination (in the MNSs system). The Ainu's fingerprints are Caucasoid, with loops outnumbering whorls by two to one.

Three theories have been postulated as to the origin of the Ainu: they are related to European or western Asian Caucasoids; they are related to the Australian Aborigines; or they are a race of their own as yet otherwise unrepresented in the fossil record. The Australian resemblance may be more a matter of grade than of line, and the third possibility, although remote, is probably the most likely.

THE PEOPLES OF THE INDIAN SUBCONTINENT

Aborigines. The Indian subcontinent includes the independent nations of India, Pakistan, Bangladesh, Nepal, Bhutan, the Indian protectorate of Sikkim, and Sri Lanka (Ceylon), which was a part of the subcontinent during the Pleistocene and is ethnically a part of it. Greater India contains members of all Old World races except the African ones. Although the subcontinent currently holds some 675,000,000 persons, parts of it are still thinly populated, particularly the forested, mountainous regions of southern India; the Thar Desert (shared by India and Pakistan); and the Terai Forest, lying mostly in southern Nepal.

The earliest inhabitants are represented by numerous scattered tribes of hunters and gatherers in the southern peninsula and by the few remaining Veddas of Ceylon. The Birhors of Chotanāgpur, the Chenchu of the Amradabad Plateau of Andhra Pradesh, and the Veddas are short, dark-skinned people, with hair mostly in ringlet form and facial features intermediate between those of Caucasoids and Australoids, the Australoid ingredient seen particularly in a depressed nasal root and, in some cases, wide nostrils. Exceptional are the Kadar of the Cardamon Hills of Kerala. Except for individuals recently mixed with Hindus from the plains, the Kadar contain two racial elements, a frizzly-haired Negrito and straight or wavy-haired Australoid. The Oraons of Bihār State and the Gonds of central India (both now agricultural) are mostly Australoid and mostly straight haired. Many other aboriginal peoples have been absorbed into the outcaste and lower caste populations of the Dravidian and Aryan speakers among whom they live.

Australoid populations of India

Dravidians. The Dravidians are the agricultural and urban peoples of southern India who speak Kannada, Telugu, Tulu, Malayalam, and Tamil. Tamil is also spoken by the inhabitants of the northern part of Ceylon; some of their ancestors came from southern India centuries ago, others more recently after the rise of the tea plantations. The Brahui of Pakistan, who live among the Baluchis, are also Dravidian speakers. It is generally believed that the Dravidians of southern India migrated to their present homes from the Indus Valley region before and during the Aryan invasions.

These Dravidians are divided into the usual Hindu castes. They are dark-skinned Caucasoids in all essential characteristics, but they differ among themselves principally in skin shade and nose form, the higher castes having a tendency to lighter skins and narrower noses. Prevalent among Dravidians, particularly the Tamils of Ceylon, is an excess growth of hair on the ears. This is a genetic trait also noted among southern Europeans. The Brahuis are lighter skinned than the other Dravidians and quite hairy.

Munda and Austro-Asiatic speakers. Scattered along the northeastern edge of the Deccan Plateau live about 5,000,000 casteless agricultural people who speak languages of the Munda group, a branch of Austronesian. Their ancestors probably migrated from Southeast Asia in pre-Aryan times, introducing the cultivated food plants of that region. They are mostly Australoid with a Mongoloid element more visible in some individuals than in others and are racially akin to the Cambodians. North of the Ganges, the Khasi inhabit the Khasi Plateau of Assam and have their capital at Gauhāti. In their territory is Cherrapunji, which has the highest recorded annual rainfall in the world. Their language belongs to the Mon-Khmer branch of the Austronesian family, and their ancestors, like those of the Munda speakers, probably migrated from Southeast Asia. They are mostly Mongoloid, with light skins, as may be expected in their cloudy, dimly-lit environment.

Mongoloid elements in Indian populations

Another isolated, mostly Mongoloid people are the Newars of the Kāthmāndu Valley of Nepal, who lived there before the arrival of the now-dominant Aryan-speaking Nepalese. They are mostly urban. Their language is Tibeto-Burman. They are light-yellowish skinned, and some of them resemble Chinese. The same is true of the Lepchās, the natives of Sikkim, now outnumbered by Nepalese immigrants. Another Mongoloid people are the Nāgas of the Nāga Hills of northeastern Assam; they are darker skinned, and many have convex nasal profiles.

Aryans. The peoples of India, Pakistan, Nepal, and Ceylon who speak Sanskrit-derived languages form the bulk of the subcontinent's population. They are mostly descended from Bronze-Age invaders from the northwest who crossed the Hindu Kush Mountains during the second millenium BC, replacing the probably Dravidian-speaking inhabitants of the Indus Valley and spreading out over the Indo-Gangetic plain wherever they could pass with their ox-drawn wagons. They introduced the major features of the Hindu religion to India and the framework of their present, highly elaborate caste system, with its basic fourfold division into priests (Brahmins), warriors (Kṣatriya), tradesmen and cultivators (Vaiśya), and inferior craftsmen (Śudra). By practicing endogamy (marriage within the group) and observing strict dietary restrictions, they have preserved genetic continuity with their Aryan ancestors to a considerable extent, particularly in the upper castes, while the lower castes physically suggest, in varying extents, the absorption of earlier Dravidians, particularly in skin colour, the fairest skins being found in northwestern India and West Pakistan.

Upper-caste Aryans are mostly dark-haired and brown-eyed Caucasoids, resembling Europeans. Blue eyes are

occasionally seen among some of them, particularly Sikhs and Kashmiris, and some of the previously mentioned Kaffirs of northwest Pakistan are blond, with Nordic features. Undoubtedly, climatic selection over the millenia has been partly responsible for the reduction in blondism among most of the upper-caste Aryans. In Nepal, Aryan invaders overran much of the kingdom, and most of the Nepalese are partly Caucasoid. The Mongoloid element increases past the upper-altitude limit of malaria in the foothills of the Himalayas. The Sherpas of the high Himalayas are Mongoloid Tibetans, as are the Bhutanese.

Sinhalese

The principal Indo-European-speakers of Sri Lanka (Ceylon) are the Sinhalese, whose ancestors invaded the island by sea from Bengal in the 6th century BC. Except for Tamil tea-plantation workers, the Sinhalese are still the principal inhabitants of the southern part of the island and the mountainous interior. The Sinhalese are mostly dark-skinned Causasoids with a minor Mongoloid element, probably brought from Assam. This element is most apparent in the highlanders, from Kandy to Nuwara Eliya, and some of the statues of ancient kings show Mongoloid features.

Blood-type distributions in India. The Hindus of caste, whether Aryan or Dravidian, have a blood-group distribution like that of the eastern Mediterranean region of Europe and west Asia, with type B more frequent than A, some A_2, and about 24 percent of Rh-negative. The Brahmins tend to have more of type A than B, and, on the whole, B surpasses A more often in the east than in the west. Casteless Hindus and aboriginal tribesmen are high in B, as are the food-gathering groups of the south, who lack Rh-negative. Among Hindus of caste, blood type M exceeds N in the MNSs system, and M is usually accompanied by S; but, among the aboriginal Chenchu, M and N are about equal, and S is tied to N, an Australoid linkage. The data covered here indicate what morphology also shows, that the Dravidians and Aryans are both of Caucasoid origin, while the aborigines contain an Australoid ancestral element.

THE WESTERN HEMISPHERE AND THE COLONIAL POPULATIONS

The American Indians. The American Indians, except for the Eskimo and Aleuts, were the sole inhabitants of the New World before its discovery by Europeans, and, as far as is known, all of the Indians' ancestors entered North America over the Bering Strait when it was above sea level during the last Ice Age. Owing to the distribution of Late Pleistocene glaciers, their initial routes southward would have been in the western part of the continent. The northeast, which became ice-free more recently, was first inhabited no earlier than was the tip end of South America.

It is not known how the ancestors of the American Indians got to the Bering Strait, but the most feasible route, considering the climate during the Ice Age, would have been along the eastern Asian coast, from lands occupied by Mongoloids and possibly by ancestors of the Ainu. Nor is it known what the Indians' ancestors looked like, because there is as yet no pertinent skeletal material from northeast Asia. The best is from three skeletons found in the Upper Cave at Chou-k'ou-tien, near Peking; they could have been ancestral Indians or Tibetans or Nāgas.

Physical characteristics of the American Indians

Whether their origin was single or multiple, the American Indians are more nearly homogeneous than the inhabitants of any other large landmass. Like the Asiatic Mongoloids, most of them have shovel-shaped incisor teeth; a fingerprint pattern high in whorls; brown eyes, occasionally with epicanthic folds; coarse, straight, black head hair that grays late in life; and scanty beards and body hair. But some of them, such as the Indians of the northwest coast and northern California and the Alakaluf of southern Chile, can grow full beards. Skin pigmentation ranges from almost Caucasoid brunet white in northwestern Canada through various shades of reddish brown in Central America and the Andes to a pale yellowish in the shady atmosphere of the equatorial forests to almost white among the Alakalufs, who also have blue patches, indicating deeply underlying pigment, on their backs and genital organs.

In response to latitude and climate, their stature range is great. The tallest Indians live in the North American Great Plains, the South American pampas, and Patagonia. The shortest are found among the highland Maya of Guatemala, some Andean populations of Colombia, and in the tropical forests. In some of the highland groups this reduced stature may be attributed to malnutrition, inbreeding, or both.

Among the Indians as a whole, body build varies relatively little. They usually have long trunks, little or no spinal curvature, and small hands and feet. Facial features are variable but mostly within the Asiatic Mongoloid ranges. Aquiline noses are common in cool regions and in mountains, while concave profiles are frequent among the forest Indians of tropical South America. Some of these South American Indians also have moderately everted lips—their Old World counterparts may be found in Borneo.

Despite this essentially Mongoloid anatomy, many Indians have a Caucasoid appearance, especially in facial features and particularly among the New England Indians. Prehistoric Indian skulls from this region have narrow faces and noses that are hard to fit into the Mongoloid pattern, and the same has been found among the skulls of the Ona of Tierra del Fuego. Whether the Caucasoid morphological factor is of genetic origin or due to parallelism is not known.

Blood-type distribution of the American Indians

In terms of blood groups, the American Indians show at least as great regional diversity as they do in morphology, although they are uniform in the absence of blood types A_2, B, and Rh-negative. In Mexico, Central America, and South America, only blood type O is found in the ABO blood group among full-blooded Indians. In North America, A ranges from less than 1 percent in the south to 35 percent in the north, with a pocket of 59 percent among the Bloods and Blackfoot. This world's highest frequency of blood type A may be no accident but the focal point of a cline. It is known that in 1837 smallpox was introduced from a fur-company steamer at Fort Kenzie, Montana, and thence spread to Canada, killing off two-thirds of the Bloods and Blackfoot and many Piegans. If these Indians had originally possessed, say, 35 percent of A, the partial immunity against smallpox attributed to the gene for blood type A could explain the rise of the blood type to its present elevated figure in the next century.

The Rh gene *C* falls between 50% and 60% in most American Indian populations. It rises to over 60% in eastern Brazil and southern California and dips to 30% to 40% on the northwest coast. A comparable regional variation may be seen with *E*, which is over 50% on the northwest coast and down to 20% to 30% in Mexico and below 40% in eastern Brazil. Among most Indian groups the frequency of Rh blood types *CDe* and *cDe* is about 50% each, but among the northwest coast Indians *cDe* exceeds *CDe* by about 20%, as it does among the New Zealand Maoris. The gene for blood type M is elevated among all American Indians but hits a high of 90% to 95% in a belt centred in Mexico and Central America and reaching north into the United States and south into South America.

The rare Diego-positive (Di +) blood type is also characteristic of the American Indians but occurs mostly in South America, where it reaches 34 percent among tribes on the Colombian–Venezuelan border. Diego-positive has also turned up sporadically among Japanese, southern Chinese, and the Sea Dayaks of Borneo and among a handful of European Caucasoids lacking known Mongoloid ancestry.

The present distribution of human populations. In AD 1500 the population of the world was probably about 500,000,000, a small percentage of what it is today. This significant increase has not been equally shared by all races. The Caucasoids and Asiatic Mongoloids have grown the most in numbers, while the Capoids have suffered the greatest decline. Especially among the Cauca-

soids, a large part of the increase is due to their settlement of previously thinly populated regions in the Americas, Australia, and New Zealand, where the distribution of races has consequently changed the most.

The United States and southern Canada have become extensions of Europe, as have southern Brazil, Uruguay, and Argentina. About 11 percent of the population of the United States is Negro. These North American Negroes have about 20 percent of European genes, while the percentage of Negro genes in the white population is negligible. In the West Indies, the percentage of the population that is Negro, with or without mixture, ranges from 100 percent in Haiti to about 25 percent each in Cuba and Puerto Rico. In French Guiana the Negro element is 91 percent; in Brazil, 11 percent, as in the United States.

From Mexico to Uruguay and Argentina the percentages of unmixed Caucasoids ranges from 2 percent in Honduras to 30 percent in Chile, except for Costa Rica, with 80 percent, and Brazil, with 62 percent. Uruguay's population is 90 percent unmixed Caucasoid and Argentina's 97 percent. Indians comprise 29 percent of the population in Mexico, 60 percent in Ecuador, 46 percent in Peru, and 53 percent in Bolivia. Elsewhere in Latin America the population is 11 percent Indian or less. Mestizos, who are European–Indian mixtures, outnumber Caucasoids in these countries, except in Costa Rica, Brazil, Uruguay, and Argentina, where there are relatively few Indians. The Mestizos also outnumber the Indians everywhere but in Ecuador, Peru, and Bolivia, where the thin air of the Andean altiplano has inhibited the reproduction of Caucasoids and Negroes. In South Africa, out of a little over 16,000,000 persons, 3,000,000 are listed as White, 11,000,000 as Bantu, and 1,500,000 as Coloured, as well as 500,000 Asiatics, mostly from India. All of these peoples settled South Africa after AD 1500, except some ancestors of the Coloureds, who were Hottentots.

In other countries, where Caucasoids have colonized lands previously occupied by other Caucasoids, the fate of the settlers has been different. The French quit Algeria after 150 years of colonization, and the Israelis have had much trouble with the Arabs in Palestine. What will happen to the races of man, even in the near future, is hard to predict, except that a general mixture of races (panmixia) will not follow, and that some of the marginal populations of the world may stage a comeback, such as the American Indians, Siberian tribesmen, and even the Australian Aborigines. The Bushmen, at the bottom of the racial population ladder, may hold their own or even increase. Chances are at least fair that the children and grandchildren of today's adults have the privilege of seeing all the major races of man—at least on television.

BIBLIOGRAPHY. C.S. COON, *The Races of Europe* (1939), a detailed geographical study of the subracial characteristics of Caucasoids in Europe, Asia, and Africa; *The Origin of Races* (1962), the work in which Coon presented his theory of the parallel transition of the principal human subspecies from the *Homo erectus* grade to that of *Homo sapiens*; and with E.E. HUNT, JR., *The Living Races of Man* (1965), a sequel to *The Origin of Races*, giving the distribution of modern races, covering the morphology, physiology, serology, and adaptation to environment, including disease; H. CUMMINS and C. MIDLO, *Fingerprints, Palms, and Soles: An Introduction to Dermatoglyphics* (1961), the definitive work on dermatoglyphics; S.M. GARN, *Human Races* (1965), a brief, lucid account of the process of race formation with a study of microevolution into small, subracial populations; F.S. HULSE, *The Human Species* (1963), a general survey of human evolution and racial distribution; W.W. HOWELLS, *Mankind in the Making*, 2nd ed. (1967), mostly concerned with human evolution, but also presents original ideas about modern races and their geographical distribution; A.E. MOURANT, *The Distribution of the Human Blood Groups* (1954), the classic work on human blood groups, their distribution, and racial significance; W.Z. RIPLEY, *The Races of Europe* (1899), Ripley's original and often-quoted division of the Caucasoids into Nordics, Alpines, and Mediterraneans.

(C.S.C.)

Porifera

The sponges, which comprise the phylum Porifera, are among the most primitive multicellular animals. They number approximately 5,000 described species and inhabit all seas, where they occur attached to surfaces from the intertidal zone to depths of 8,500 metres or more. One family, chiefly, the Spongillidae, is found in fresh water. Adult sponges lack a definite nervous system and do not show conspicuous movements of body parts.

Early naturalists regarded the sponges as plants because of their frequent branching form and their lack of obvious movement. The animal nature of sponges, first described in 1755, was confirmed in 1765 after observations of their water currents and the movement of their openings. In structure, function, and development, sponges are distinct from other animals. Many zoologists regard sponges as occupying an isolated position in the animal kingdom and classify them in the subkingdom Parazoa; on the other hand, sponges are sometimes considered relatives of another invertebrate animal group, the coelenterates.

The phylum Porifera may be divided into three classes on the basis of the composition of the skeletal elements —the Calcispongiae and Hyalospongiae include about 20 percent of the known species of sponges; the remainder are placed in the class Demospongiae (see Figure 1).

From (*Spongia*) T.J. Parker and W.A. Haswell, *A Textbook of Zoology*, St. Martins Press, Inc., Macmillan & Co., Ltd.; (others) G.C. Pimentel, *Natural History*

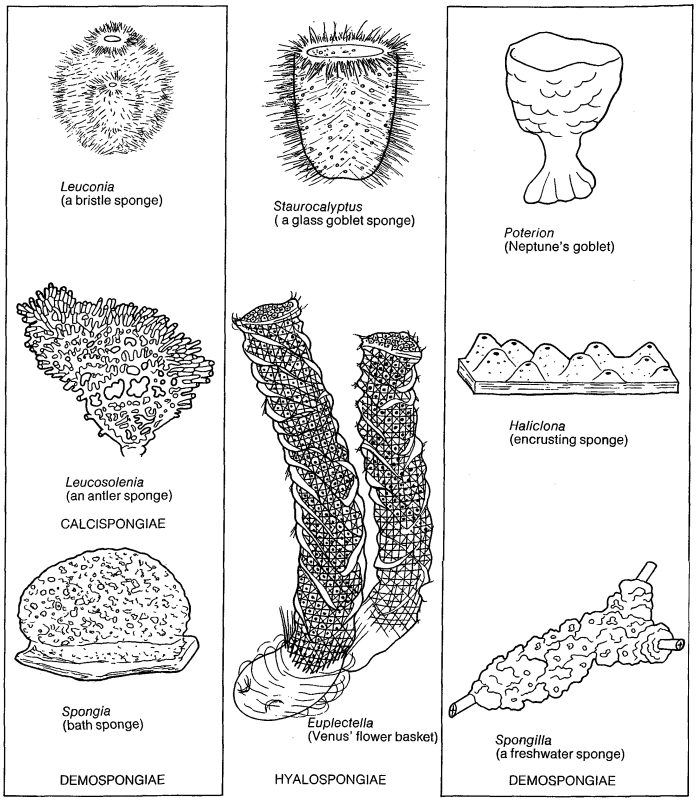


Figure 1: Representative sponges.

GENERAL FEATURES

Importance. Sponges derived from certain species of the Demospongiae—*e.g.*, *Spongia officinalis, Hippospongia communis, Spongia zimocca, Spongia graminea*—have been known since ancient times. Familiar household items in ancient Greece and Rome, sponges were used to apply paint, as mops, and as substitutes for drinking vessels. During the Middle Ages, burned sponge was reputed to have therapeutic value in the treatment of various diseases. Natural sponges now are used mostly in arts and crafts such as pottery and jewelry making, painting and decorating, and in surgical medicine. Synthetic sponges have largely replaced natural ones for household use.

Commercial value

The living sponge is a mass of cells and fibres, its interior permeated by an intricate system of canals that open as holes of various sizes through the tough dark brown or black skin, which may be hairy from fibre ends that pierce it. Only after it has been completely cleaned of its

millions of living cells does a sponge resemble the sponge of commerce; *i.e.*, a soft and elastic spongin skeletal framework. Commercially valuable sponges, which may be found from tidal level to a depth of about 200 feet, usually are harvested by hooking or harpooning in shallow waters, by skin diving or by deepwater fishing. Although the most valuable sponges are found in the eastern Mediterranean area, they also are harvested off the west coast of Florida and the Florida Keys, in the West Indies, off Mexico and British Honduras, and, to a limited extent, off the Philippine Islands. Because they have the ability to regenerate lost parts, sponges can be cultivated from small fragments.

Sponges are valuable from a scientific point of view because of their unusual cellular organization (the cells do not form tissues such as those found in other animals), their ability to regenerate lost parts, and their biochemical features (they have many compounds not known in other animals). Sponges comprise an important part of the organisms that live in the depths of the sea (benthos) and may be associated with other organisms; *e.g.*, many types of animals live in sponges.

Size range and diversity of structure. Most sponges are a few centimetres in size, but some urn-shaped or shapeless ones are less than a centimetre; others, shaped like vases, tubes, or branches, may be one to two metres tall, and broad rounded masses may be one to two metres in diameter. Size within a species may vary with age, environmental conditions, and food supply.

The external appearance of sponges is variable. They may appear bushy or treelike and have fingerlike projections. Some, particularly in the class Demospongiae, are shapeless, or amorphous, masses that form thin encrustations on objects or are cushion shaped. A few Demospongiae have well-defined spherical shapes as in *Tethya aurantium*, the sea orange; others may be cup- or fan-shaped. Calcispongiae of the genus *Scypha* are shaped like tubular sacs, with an opening (osculum) at the tip. Members of the Hyalospongiae are erect or cylindrical, with a stalklike base.

Range of colour. Colour among sponges is variable. Deep-water sponges usually show a neutral colour, drab or brownish; shallow-water sponges, frequently brightly coloured, range from red, yellow, and orange to violet and occasionally black. Most Calcispongiae are white.

Effects of symbionts

Some sponges (*e.g.*, the Spongillidae) are often greenish because green algae live in a symbiotic relationship within them; others are violet or pinkish, because they harbour symbiotic blue-green algae. These symbionts endow the sponges with colour as long as light is available; the sponges become white in the dark when no photosynthesis occurs and the algal pigments utilized in photosynthesis are no longer produced. Another variable character in sponges is consistency, which may range from the soft and viscous state of some encrusting species to the hard stonelike quality of the genus *Petrosia*. In addition, the surface of a sponge may be smooth, velvety, rough with protruding skeletal elements called spicules, or conulose (*i.e.*, provided with conical protrusions called conuli).

Distribution and abundance. The Porifera are present at all water depths, from the tidal zone to the deepest regions (abyss). They occur at all latitudes and are particularly abundant in Antarctic waters. Calcispongiae and Demospongiae are found mainly on the rocky bottoms of the continental shelf; the Hyalospongiae are characteristic of the deepest muddy bottoms of oceans and seas. In some environments, sponges are the dominating organisms; sometimes they cover wide areas, especially on rocky overhangs and in the caves of the littoral, or shore, zone. A restricted number of species are adapted to brackish waters; and members of the family Spongillidae (class Demospongiae) populate the fresh waters of rivers and lakes.

NATURAL HISTORY

Life cycle. Most sponges reproduce sexually, although asexual reproduction may occur. Sponges are generally hermaphroditic (male and female germ cells in one animal); sometimes sex inversion, in which male and female germ cells develop at different times in the same animals, occurs.

Sexual reproduction. The fertilization of an egg by a spermatozoan is peculiar in sponges in that a spermatozoan, after its release from a sponge, is carried by the water current until it is captured by a specialized cell called a choanocyte, or collar cell, in another sponge; the choanocyte transforms into an amoeboid-shaped cell called a carrier cell, which gives up the spermatozoan to an egg, lying near a chamber formed by choanocytes and containing long lashlike appendages called flagella.

Types of larvae

Development of the embryo may occur in one of several ways characteristic of the different groups; as a result, more than one type of larva is found. The characteristic larva of the Calcispongiae and of some Demospongiae (*e.g.*, *Oscarella*), called an amphiblastula, is oval in shape and has a cavity in the middle; the front half of the larva consists of cylindrical, flagellated cells, the other half of round cells without flagella. The larva swims with the flagellated portion forward. The amphiblastula is preceded by a stage (stomoblastula) in which the central cavity of a hollow mass of cells (blastula) opens outward and is surrounded by round granular cells (macromers), which are distinguished from other cells with flagella (micromers). The most common larval form among the Demospongiae is called a parenchymula; it is solid and compact, with an outer layer of flagellated cells and an inner mass of cells.

A larva swims for a period of time that may vary from a few hours to a few days before it descends to find a surface suitable for attachment. After attachment, the larva undergoes developmental changes (metamorphosis) becoming a young individual called an olynthus in the Calcispongiae, a rhagon in the Demospongiae; the young individual gradually develops into an adult sponge. The metamorphosis following larval attachment involves changes in the relative positions and functions of larval cells. In one larval type (parenchymula), the flagellated outer cells become the collar cells (choanocytes) of the interior of the adult sponge; the interior cells of the larva give rise, in the adult, to the cell layer (pinacoderm) and the different cells (*e.g.*, archaeocytes, collencytes) found in the amorphous substance (mesohyl) that fills the sponge. In the amphiblastula, the choanocytes are derived from the forward flagellated region; the other cells and the mesohyl are derived from the posterior half.

The sexual maturation of sponges is connected with the temperature of the water in which they live. In the temperate regions, maturation occurs mainly from spring to autumn; sometimes two distinct periods of reproduction occur, one in spring, the other in autumn. Some sponges mature at any time of the year; *e.g.*, *Scypha*, formerly called *Sycon*. Tropical sponges also apparently mature any time throughout the year. Sponges for the most part bear living young (*i.e.*, are viviparous); the larvae are released through the canals of the excurrent (outgoing) water system and an opening (osculum) also involved in that system. A few sponges (*e.g.*, *Cliona* and *Tethya*) lay eggs (*i.e.*, are oviparous).

The life-span of sponges is not well known; the small encrusting forms probably live about a year, disappearing during a season unfavourable to their survival; small fragments of an individual, however, may persist and reproduce new individuals in the next season. The large species have a much longer life-span; bath sponges (*Spongia, Hippospongia*), for example, attain a commercially desirable size after seven years and may live as long as 20 years.

Asexual reproduction. Asexual reproduction also occurs in sponges in various ways; the best known method is called gemmulation. Gemmulation begins when aggregates of cells, mostly archaeocytes, which, when they become laden with reserve food granules, are called thesocytes, become isolated at the surface of a sponge and are then called gemmules. These are expelled from the adult sponge and, in some marine species, serve as a normal reproductive process or, sometimes, as a means

Gemmules

to carry the sponges over periods of unfavourable conditions when the adults degenerate; *e.g.*, drought, temperature extremes.

The freshwater Spongillidae undergo a slightly different gemmulation. Gemmules consist of aggregates of archaeocytes laden with reserve granules; in addition, however, they are surrounded by protective membranes formed by the archaeocytes. The protective covering is generally reinforced by spicules, which vary in shape according to the species and are useful in classification. Freshwater sponge gemmules allow a species to survive unfavourable conditions in a state in which vital activities are almost completely suspended. In cold regions, gemmulation occurs in winter, and the inactive gemmules are said to hibernate; in warm regions, gemmulation occurs in summer, and the gemmules are said to estivate. In spring or autumn when favourable conditions return, the gemmules germinate, their archaeocytes emerge through an opening (micropyle), the various cellular types differentiate, and a new sponge grows. Other methods of asexual reproduction include formation of stolons (rootlike extensions) and fragmentation of individuals.

Regeneration. The extraordinary capacity of sponges to regenerate is manifested not only by restoration of damaged or lost parts but also by complete regeneration of an adult from fragments or even single cells. Sponge cells may be separated by mechanical methods (*e.g.*, squeezing a piece of sponge through fine silk cloth) or by chemical methods (*e.g.*, elimination of calcium and magnesium from seawater). Dissociated cells settle, migrate, and form active aggregates in which the archaeocytes play an important role. In order for small aggregates of cells to form larger aggregates, the cells must generally become attached to a surface, where they flatten and develop an envelope of special cells (pinacocytes); this is called the diamorph stage. Reconstitution of the choanocyte chambers and of the canal system follow soon afterward, resulting in a young sponge that is functional and able to grow. It is generally believed that the reconstitution process, even if it involves cell division, is not comparable with embryonic development, because the various types of dissociated cells participate in the formation of the new sponge by sorting and rearranging themselves, rather than by differentiating from primitive cell types. Regeneration in sponges is of theoretical interest in connection with certain cell properties; *e.g.*, adhesion, sorting out, and movement.

During unfavourable conditions, sponges are reduced to small fragments that may consist only of masses of archaeocytes covered by layers of pinacocytes. A complete sponge forms from these fragments when favourable conditions return.

Individuality of sponges

The regenerative abilities of sponges, their lack of a central coordinating organ (brain), and the peculiar migratory ability of cells within the organisms combine to make it somewhat difficult to define the individuality of a sponge. Various attempts to establish whether sponges are individuals or colonies have failed to solve the problem, and, thus far, no completely satisfactory criterion for defining a sponge has been found. Zoologists involved in the study of sponges empirically define a sponge individual as a mass that is enveloped by a common ectoderm, *i.e.*, by a common cellular layer.

Ecology. Most Porifera, very sensitive to a wide range of ecological factors, are difficult to raise under laboratory conditions. Few species (*e.g.*, *Hymeniacidon sanguinea*) are able to tolerate long periods of emersion and variations in such physical factors as light, temperature, and salinity.

Habitats. Light is a limiting factor for the survival of most sponges in a given habitat. Littoral-dwelling sponges generally develop in caves, on shadowed walls, or under small shelters such as those provided by crevices. Some species, mainly in the tropics, however, are covered by a metre or less of water and thus are almost completely exposed to irradiation from the sun. Symbiotic relationships between algae and sponges usually occur in strongly illuminated zones; the algae may act as a protective device because they deposit pigments in the superficial cell layers of the sponge. In some sponges (*e.g.*, *Petrosia ficiformis*), colour is related to the number of symbionts; in a cave, for example, sponges gradually change from intensely coloured specimens to light-coloured, sometimes white, ones in the depth of the cave where the number of algae decreases.

Porifera of the family Clionidae (class Demospongiae) live in galleries they excavate in shells of mollusks, in corals, in limestone, and in other calcareous materials. The boring activities of clionids are accomplished by the excavation, possibly involving both chemical and mechanical action, of numerous, small chips of calcium carbonate. Cytoplasmic projections and films put out by sponge cells in contact with a calcareous surface apparently come into intimate contact with the calcium carbonate, resulting in the removal of particles of relatively uniform size. Clionid sponges weaken limestone breakwaters and coral reefs, making them more easily subject to further abrasion by waves. In addition, they weaken oystershells.

Attachment to a surface

Although most sponges settle and grow on hard or rocky surfaces, some anchor to a firm object on soft bottoms, on sand, on mud, or on debris. Unattached sponges are rare. Different species may compete for a surface, and superposition of one species on another sometimes occurs; the presence of a rich population of different species on the same surface may help them to survive by the modifications each contributes to the environmental microclimate surrounding them, thereby providing protection against extreme fluctuations of physical factors such as temperature and light.

Associations with other organisms. The Porifera often grow on or near other organisms, sometimes killing those they cover; the sessile (attached) barnacle *Balanus*, for example, may be killed in this way. In other cases, associations may provide advantages to both organisms, particularly those between sponges and crustaceans. Some crustaceans, mainly crabs, use sponges for camouflage by removing a piece of a living sponge and holding it against their carapace (shell); the best known example of this type of mutualistic association is that of the sponge *Suberites domuncula* and hermit crabs, which live in the shells of gastropod mollusks. The advantage to the sponge is that it is carried by the mollusk; the hermit crab gains protection not only by living in the shell of the mollusk but also through the disagreeable smell and taste of the sponge, which discourages attack by fishes and other enemies.

Various plants and animals may live on the surface of a sponge or inside its canals and cavities. In some cases the associations are specific; *e.g.*, the coral *Parazoanthus axinellae* grows on the sponge *Axinella*. The organisms that live in the cavities of sponges include crustaceans, nematode and polychaete worms, ophiuroid echinoderms (brittle stars), and bivalve mollusks; some inhabit a sponge for occasional shelter or nourishment, others establish more intimate associations as parasites or predators. Young shrimps of the genus *Spongicola* penetrate certain sponges of the class Calcispongiae, live in them in pairs, and presumably are trapped for life in the rigid skeleton of the sponge; the Japanese consider these shrimps a symbol of matrimonial faithfulness. The number of organisms that live within a single sponge may be very high; thousands of organisms of various species, for example, may be found in *Spheciospongia vesparia*, a Caribbean sponge.

Some organisms that live on (called epibionts) and in (called endobionts) sponges act as parasites. Cyclopoid copepods are the most important parasites of marine sponges; in fact, some genera of these crustaceans have become modified as a consequence of their parasitic existence. Freshwater sponges also are attacked by parasites such as rotifers and mites, which lay eggs in them; larvae of the neuropteran insect family Sisyridae (spongillaflies) live in, and feed upon, freshwater sponges. In general, sponges are protected from predators by their disagreeable taste and smell and by their hard skeletal elements (spicules). In some cases, however, sponges are eaten by other organisms; *e.g.*, mollusks—gastropods such as

snails and nudibranch slugs, prosobranchs such as *Patella* and *Littorina*, and chitons—some crustaceans, and some fishes (especially on coral reefs).

Algal associations

The most important symbiotic associations of sponges occur with single-celled and multicellular algae. The algae may live in the surface layers of the sponge, inside the cells, or among them. The sponge protects the algae from enemies, from unfavourable environmental conditions, and from their own metabolic waste products; the sponge uses the algae as a source of oxygen, as a mechanism for eliminating its products of metabolism, as a screen against sunlight, and as a food source (consuming both algal waste products and dying algae). Sponges of the freshwater Spongillidae and various species of marine littoral sponges consume dying green and blue-green algae respectively. The algae, which provide the Spongillidae with their characteristic green colour, may be transmitted through the gemmules. In some boring clionid sponges (*Cliona viridis*) of the class Demospongiae, some single-celled brown algae are constantly present. The marine sponges may also harbour multicellular blue-green algae (*e.g., Oscillatoria*), red (Rhodophyceae) and green (Chlorphyceae) algae. Red and green algae sometimes provide skeletal support for certain sponges.

Diseases of sponges. Sponges may be attacked by diseases of epidemic character, the agents of which are not well known. The commercial sponges of the West Indies once were nearly completely destroyed by a fungus-like micro-organism; other sponges were not damaged.

FORM AND FUNCTION

The Porifera are unusual animals in that they lack definite organs to carry out their various functions. The most important structure is the system of canals and chambers, called a water current system, through which water circulates to bring food to the sponge.

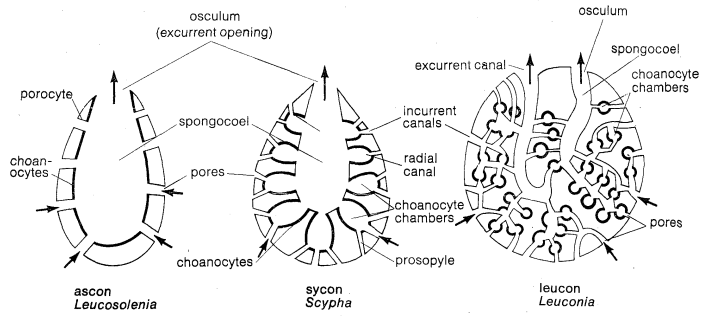


Figure 2: Water current systems in calcareous sponges. Arrows indicate the direction of water flow.

Water current system. The essential elements of the water current system include the pores, or ostia, through which water enters the sponge (incurrent system); the choanocytes, or collar cells, which are flagellated cells that capture food; and the oscula, openings through which water is expelled (excurrent system). Three types of water current systems of increasingly complex structure may be distinguished by the arrangement of choanocytes and the development of canals—ascon, sycon, and leucon (Figure 2). The simplest, or ascon, type, found only in certain primitive genera of Calcispongiae (*e.g., Leucosolenia*), is characterized by an arrangement of choanocytes around a central cavity that directly communicates with the osculum. The walls of these sponges are thin, lack canals, and are perforated by pores, which actually are openings through cells (porocytes). The sycon type of water current system, found also in Calcispongiae is at first characterized by choanocytes that surround fingerlike projections called radial canals of the sponge wall. Water enters the radial canals directly through pores, makes its way into the central cavity, or spongocoel, and leaves by way of an osculum. In most syconoid sponges (*e.g., Scypha*) the radial canals are bordered by incurrent canals through which passes the water entering the pores; other openings (prosopyles) allow water into the choanocytes, from which it passes directly into the internal cavity and out of it through the osculum. In the leucon type, which is found in the more advanced Calcispongiae and in the other classes (Demospongiae and Hyalospongiae), the radial canals are replaced by numerous small flagellated chambers in which the choanocytes are localized; the chambers, scattered throughout the body of the sponge, have pores through which water passes into a complex system of incurrent canals, then into a spongocoel (internal cavity) by way of excurrent canals. Water enters very small pores found among the cells (pinacocytes), which line the outer surface of the sponge. After passing through a system of incurrent canals and cavities, also lined with pinacocytes, the water reaches the flagellated chambers, enters them through openings (prosopyles), and leaves through other openings (apopyles). The water is expelled through the osculum after passing through a system of excurrent canals and cavities lined with pinacocytes. During the development of many sponges, a simpler water current system (rhagon) precedes the leucon type. The rhagon type is characterized by reduced excurrent canals and by a large central cavity. In some Demospongiae the body is organized in two parts, an external ectosome without choanocytes, and an internal choanosome with choanocytes.

Leucon water current system

Cell types. The sponges lack a well-defined organization of tissues. Single layers of cells line the outer surface of the body and the internal cavities; other cells, both motile and fixed, and fibres occur in an amorphous substance (mesohyl), gelatinous in nature. It has not been possible thus far to identify with certainty similarities of origin (homologies) between the various types of sponge cells and those of higher animals. Each type of sponge cell performs particular functions; the cells either may gather in certain areas of the sponge or form layers and membranes. They are easily modified, both in form and function, during larval development and during adult life. Furthermore, they have a remarkable ability to migrate and to transform from one cell type to another, although the mechanisms involved are not known. Three principal types of cells may be distinguished—choanocytes, archaeocytes, and pinacocytes-collencytes (Figure 3).

Choanocytes and archaeocytes. The choanocytes are provided with a flagellum, which is surrounded by a collar composed of cytoplasm. The main function of the flagellum apparently is to produce the water current, that of the collar is to capture food particles.

The archaeocytes, which are scattered in the mesohyl, have remarkable potentialities for transformation into various cell types, especially in the Demospongiae. Some persist and reproduce during the life of the sponge without specializing, thus forming an embryonic reserve from which other cellular types may be derived; others become specialized to carry out particular functions. Archeocytes, often called amoebocytes, are amoeboid cells (*i.e.*, they have the ability to move); their cytoplasm contains large quantities of ribonucleic acid (RNA), and their large nuclei contain small bodies known as nucleoli. Amoebocytes function in regeneration and in transportation of food particles acquired at the choanocytes throughout the body of the sponge. During this distribution process, they carry granules consisting of various substances. Depending on the nature of the substance, amoebocytes may be known by several names; *e.g.*, granulocytes, thesocytes. Amoebocytes laden with various pigments (carotenoids and melanin, sometimes contained in algal symbionts) confer various colours to the sponge.

The archaeocytes may be important in sexual reproduction, if, as is postulated, male and female reproductive cells are derived from them. This role is disputed, however, since in some cases, mainly in the Calcispongiae, reproductive cells, particularly those of the male, are derived from choanocytes.

Pinacocytes, collencytes, and other cell types. Pinacocytes (Figure 3) form the pinacoderm, a single cell layer found on the body surface and lining the canals. Various types of pinacocytes occur—basipinacocytes are in contact with the surface to which the sponge is attached,

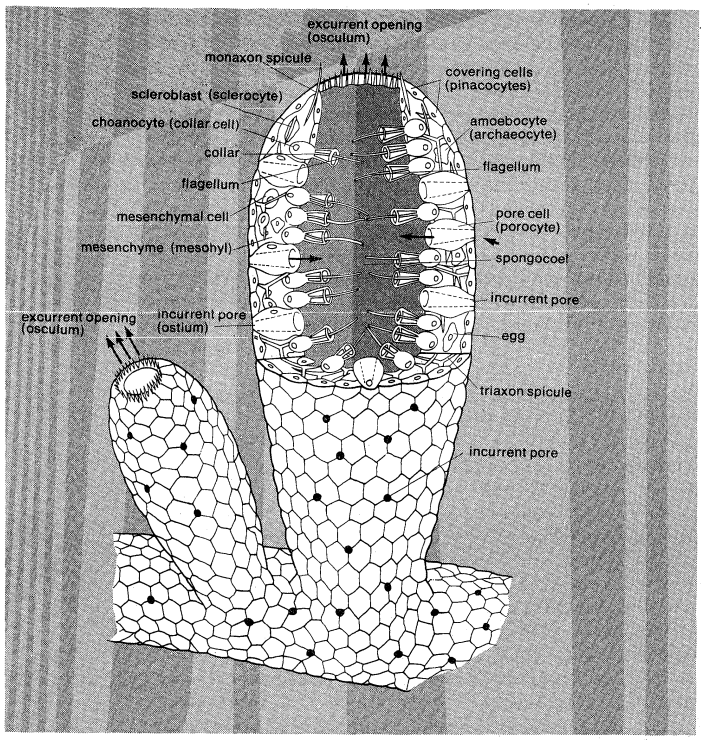


Figure 3: Part of a simple sponge with the upper part cut away to show cell structure.
From R. Buchsbaum, *Animals Without Backbones* (1948); University of Chicago Press

exopinacocytes are found on the surface of the sponge, and endopinacocytes line the canals. Pinacocytes are flattened cells containing many granules; capable of contracting, pinacocytes may cause a reduction in the volume of the sponge if it is disturbed. In the Calcispongiae, the outer surface of the body also contains flattened granular cells called porocytes because they contain the pores needed to allow water into the sponge. The porocytes can contract, thus closing the pores during unfavourable environmental conditions.

Lophocytes, sclerocytes, and myocytes

The collencytes, found in the mesohyl, secrete fibres and often form a net in the cytoplasm. The mesohyl of sponges contains other types of cells (lophocytes, sclerocytes, myocytes) believed to be derived from archaeocytes. Lophocytes, similar to but larger than collencytes, have long cytoplasmic processes at one end, giving them the appearance of a comet; they apparently secrete fibres (spongin) that form skeletal material. The sclerocytes, or scleroblasts, which also produce skeletal material, are classified according to the chemical nature of the spicules; calcoblasts secrete calcareous spicules, silicoblasts siliceous (glasslike) ones. The myocytes are elongated, contractile cells, particularly abundant near the oscula, where they control their expansion and contraction. The presence of nerve cells in sponges is a matter of dispute; the general opinion, however, is that none exist, not even in a primitive form.

Skeleton. The skeleton of sponges is of great taxonomic significance. It may be mineral in nature (calcareous or siliceous) or composed of protein and other components (spongin, see Figure 4). The mineral skeleton is formed for the most part by units called spicules, either scattered throughout the sponge or united to form fibres; spicules are classified as megascleres, which function in support, and microscleres, which function in protection and also aid in support.

Spicule types

Mineral skeletons. Calcareous spicules, characteristic of the Calcispongiae, are composed chiefly of calcium carbonate in crystalline forms; *e.g.*, calcite, aragonite. Most calcareous spicules have one axis (monoaxon), which is usually pointed at both ends; these spicules are called oxeas. Triaxons have three rays and are called triacts; tetraxons have four rays and are called tetracts. Figure 4 shows spicule types.

Siliceous spicules, found in the Demospongiae and in the Hyalospongiae, are made essentially of silicic acid; they also contain some water, a small quantity of other compounds containing sodium, potassium, iron, and chlorine, and a small quantity of organic matter, called spiculin, which forms an axial fibre. The spicules of the Hyalospongiae are variable in form and often have remarkable dimensions. Characteristic spicules of the Hyalospongiae are triaxon forms with three orthogonal axes (that is, six rays); this form is called hexactine, and the class is also known as the Hexactinellida. The spicules are connected in a continuous network, and after the death of the sponge and the loss of its soft parts, the skeleton that remains has a delicate glass texture; *e.g.*, the Venus basket, *Euplectella*. Bundles of large spicules form stalks that allow members of the Hyalospongiae to attach to the muddy bottoms of the deeper parts of the ocean in which they generally live. In the genus *Monoraphis*, the stalk is one enormous spicule that may attain a length of two or three metres and a thickness of approximately one centimetre.

The siliceous spicules, consisting of both megascleres and microscleres, of the Demospongiae have an enormous variety of forms. The megascleres may be monaxons with both ends pointed (oxeas), with one end pointed and the other rounded (styles), or with both ends rounded (strongyles). If one end is swollen styles are called tylostyles and strongyles tylostrongyles; the spicules with both ends swollen are called tylotes. If the surface of the spicules is spiny instead of smooth, the spicules are called, respectively, acanthoxeas, acanthostyles, and acantho-

From (top right, dictyonine framework) R. Buchsbaum, *Animals Without Backbones* (1948), University of Chicago Press; (others) C. Villee *et al.*, *General Zoology*, 3rd ed. (1968), W.B. Saunders Company, Philadelphia

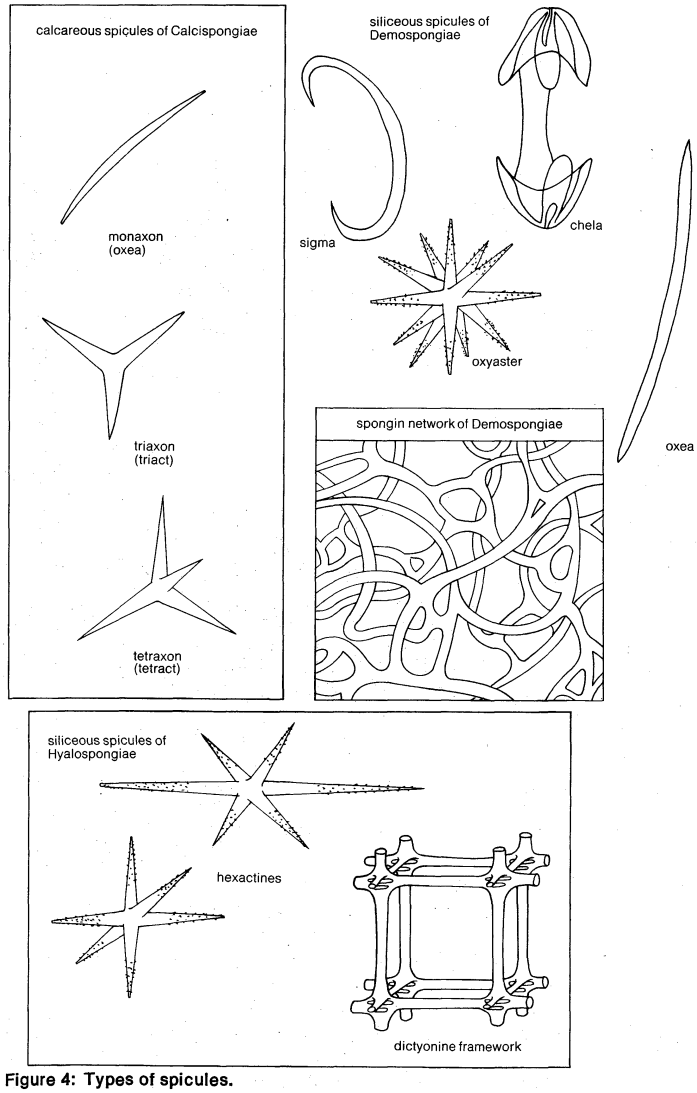


Figure 4: Types of spicules.

strongyles. The megascleres also include triacts with three rays and tetracts, called calthrops, with four rays. Tetracts with one axis (rhabdome) longer than the other three (collectively the cladome) are called protriaenes, plagiotriaenes, anatriaenes, or dichotriaenes, depending on the way the rays of the cladome are directed.

All of the microscleres apparently are derived from a spherical type with many axons (polyaxon); the result is a series of star-shaped spicules, or asters, with various numbers of rays. Spicules with rays missing or reduced (called spheres, sterrasters, and discasters) often form a protective armour around the sponge. More specialized types of microscleres include sigmas, toxas, chelas, and anchoras; the last two have plates or teeth at each end and may be distinguished as isochelas and isanchoras or anisochelas and anisanchoras, depending on whether the ends are equal or unequal.

Other types. A few Demospongiae (*e.g., Oscarella, Halisarca,* and *Chondrosia*) lack skeletons. One group (Ceractinomorpha) has a type of spongin, which, in certain orders (Axinellida, Poecilosclerida, and Haplosclerida), cements the spicules in bundles or meshes, thereby increasing the elastic nature of the skeleton. In another group of Demospongiae (Keratosa), spongin fibres constitute the entire skeleton; the spongin fibres may be branched (order Dendroceratida), netlike (order Dictyoceratida), without inclusions (commercial sponges, which are therefore soft and elastic), or with inclusions (*e.g.*, grains of sand, fragments of spicules). In the genus *Ircinia,* the fibres are accompanied by thin spongin filaments that fill the mesohyl.

Spongin

Specialized types of skeletons in two groups of great paleontological importance are now represented by only a limited number of species, in the Calcispongiae and Demospongiae. Those in the Calcispongiae (Pharetronida) have skeletons formed by an amorphous mass of calcium carbonate, with which few spicules are associated. Those in the Demospongiae (Lithistida) form a heterogeneous group in which irregularly branched spicules (desmas) form a compact skeleton. Some Demospongiae, found mainly on the coral reefs, possess a compact calcareous skeleton, which incorporates both siliceous spicules and organic fibres.

Functional features. *Feeding and digestion.* The Porifera are filter feeders, utilizing food particles suspended in the water and captured by the choanocytes. Food particles consist essentially of bacteria, other micro-organisms, and particles of organic debris; sponges also probably absorb dissolved organic substances. Experiments with starch grains, bacteria, and particles of carmine pigment show that the particles ingested by the choanocytes are transferred by thesocytes (specialized amoebocytes) throughout the sponge. Intracellular digestion occurs in both choanocytes and thesocytes. A large quantity of food is absorbed by a sponge; for example, a sponge ten centimetres tall with a diameter of one centimetre has about 2,250,000 flagellated chambers and utilizes about 22.5 litres (about 23 quarts) of water a day. The water current system, therefore, is efficient, in spite of the lack of a nervous coordinating centre. The water flow into and out of the sponge is maintained by continuous movement of the flagella of the choanocytes; flagellar movement pushes water into the choanocyte chambers, creating in them a pressure that extends to the excurrent canals, as water passing into the pores is drawn along the system of incurrent canals. Water flow, which is also regulated by contraction and expansion of the pores and of the oscula, is slow at the pores, fast at the oscula; as a result, water expelled with force through the oscula cannot contaminate water entering the pores. Toxic or irritating substances cause the pores to close.

Oxygen uptake and excretion. Respiratory organs are lacking in sponges; oxygen is supplied by a direct exchange between the tissues and the surrounding water. Consequently, the consumption of oxygen is related to the velocity of the current. Excretion occurs through both the oscula and the surface of the sponge. Special amoebocytes disintegrate in the mesohyl, and their granules are expelled through the canals. The excretory products of the sponges—ammonia and other nitrogen-containing substances—account for their characteristic unpleasant odour and probably make them undesirable as food for other animals. Many sponges (*e.g.*, the tropical sponge *Tedania ignis*) exude large quantities of mucus, and some species produce toxic substances, which may cause inflammation and skin reactions in man.

Coordination. A sponge usually contracts if it is handled or removed from water. The oscula close for various reasons—if the sponge is exposed to air, if oxygen is not available, if harmful chemical compounds are present, or if temperatures are extremely high or low. The absence of a nervous system would seem to be confirmed by the lack of coordination in the behaviour of oscula, pores, and choanocytes and by the slow reaction of sponges to stimuli. A primitive system of coordination, probably chemical in nature, apparently exists. Contraction of myocytes occurs around the oscula and along the peripheral canals of Demospongiae. Some Calcispongiae of the ascon type reduce their volume through contraction of the superficial layer of pinacocytes.

Biochemical aspects. The sponges produce substances with antibiotic activity (*e.g.*, ectyonin), which possibly function during the selection of bacteria and other micro-organisms on which they feed. The Porifera contain a greater variety of fatty substances (*e.g.*, sterols) than do other animals. Some of these sterols (*e.g.*, clionasterol, poriferasterol) are found only in sponges; others (*e.g.*, cholesterol) are common in other animals. Numerous carotenoid pigments occur in sponges; as many as 15 different ones have been found in one sponge. Although carotenoid pigments are among the most important pigments found in sponges, melanin, chlorophyll, and phycoerythrin, which are derived from algal symbionts and from the diet, also occur. Sponges accumulate silicon, calcium, and considerable quantities of metals; *e.g.*, iron, manganese, nickel, and gold. The spongins are iodine or bromine-containing scleroproteins similar to the keratin found in skin, claws, hair, and feathers of other animals. The two types of spongin, known as A and B, differ in composition and structure.

Sterols and pigments

EVOLUTION AND PALEONTOLOGY

Sponges have evolved in their own way along a developmental line foreign to that of other animals. They probably arose from the Protozoa, most likely from some flagellated stock. The group of flagellates that gave rise to the sponges is uncertain. The choanocytes of sponges resemble a group of protozoans, the choanoflagellates; one genus, *Proterospongia,* was so named because its structure seemed to indicate it as a possible sponge ancestor. Choanoflagellate protozoan colonies, however, do not develop through embryological stages as do the sponges. The primitive structure of the Porifera does indicate affinities with certain types of protozoan colonies; both lack integration of parts, mouths, and digestive systems, and both have a type of skeletal formation in which single elements are produced by a single cell or by a small group of cells.

The Porifera appeared in the Early Cambrian Period of the Paleozoic Era (about 550,000,000 years ago). Of the classes known in the Middle Cambrian Period (Hexactinellida or Hyalospongiae, Heteractinellida, and Demospongiae), the Heteractinellida are extinct; the Calcispongiae appear in the Carboniferous Period (about 345,000,000 to 280,000,000 years ago). Sponges have undergone a very limited evolution; *i.e.,* living forms do not differ substantially from many groups represented in the Paleozoic Era.

The Porifera, in spite of their mineral skeleton, are not well preserved as fossils; usually, only the type of spicules known as megascleres are preserved. The only groups well represented as fossils are those with a compact skeleton; *e.g.*, pharetronids (Calcispongiae), lithistids (Demospongiae), and Dictyonina (Hyalospongiae). Since these are specialized groups, they cannot provide a precise picture of the history of the phylum.

The oldest sponge fossils, from the Lower Cambrian, are well represented in the Burgess Shale of the Middle

Cambrian in western Canada by *Protospongia*, a member of the Hyalospongiae, and by *Eiffelia*, of the extinct Heteractinellida. They resembled a pouch and had very thin walls; their spicules were arranged in a single layer, and their water current system was probably of an ascon type. Various genera of the order Lyssacina (Hyalospongiae) are known from the Paleozoic from spicules. In the Mesozoic Era (about 225,000,000 to 65,000,000 years ago), the Hyalospongiae gave rise to a group (order Dictyonina) with the compact skeleton commonly found in sediments. Beginning with the Late Cretaceous Period (about 100,000,000 years ago), the species of Hyalospongiae have affinities with recent ones.

Fossil Calcispongiae appear in the Carboniferous Period (about 345,000,000 to 280,000,000 years ago). The first Pharetronida (Calcispongiae) from the Permian Period (about 280,000,000 to 225,000,000 years ago), have both a compact calcareous skeleton and spicules; because calcareous spicules are easily dissolved during fossilization, Calcispongiae which lack a compact skeleton have left few traces.

According to some experts, the Demospongiae followed a line of evolution originating in Tetractinellida (with four rayed spicules) and advancing, after reduction of the tetraxon spicules into monoaxons, to Keratosa; in addition, the polyaxon microscleres (asters) evolved into specialized forms (sigmas and chelas), and spongin appeared to cement the spicules, finally developing into a skeleton formed exclusively of horny fibres. According to the recent opinion by some zoologists and paleontologists, however, the Demospongiae are a diphyletic class, which is separable into two large groups (Tetractinomorpha and Ceractinomorpha). The Tetractinomorpha are characterized by four-rayed megascleres, asters, and no spongin; on the other hand, the Ceractinomorpha have monaxon megascleres, no asters, and spongin. That these morphological differences are accompanied by important embryological differences is supported by the facts that the oldest Demospongiae, from the Cambrian, are monaxonid (with only oxeas) and that the four-rayed spicule originates later, in the Carboniferous Period. Well-represented since the Paleozoic, the Lithistida have a siliceous skeleton of netlike desmas and are considered polyphyletic; the lithistid modification of the skeleton appeared independently in different lines of the Demospongiae. Few lithistid genera are extant. Demospongiae lacking a compact skeleton are poorly represented in sediments.

The phylogenetic relationships among the classes are difficult to ascertain, although common spicular characters may be found among Calcispongiae, Hyalospongiae, and tetractinellid Demospongiae. Noteworthy is the existence of sponges with both calcareous and siliceous elements in their skeleton; these sponges show affinities with both the pharetronid Calcispongiae and the Demospongiae. The extinct Archaeocyatha, with a cup-shaped body fixed on a surface and a skeleton of porous walls, were aberrant forms that lived exclusively in the Cambrian and that are generally considered as a separate phylum, even though they do show some affinities with the Calcispongiae.

CLASSIFICATION

Distinguishing taxonomic features. The most important feature used to distinguish sponges is the skeleton. The general architecture of the skeleton is used to differentiate families, the particular combinations of spicular types to define genera, and the form and dimensions of single spicule types to differentiate species. Other morphological characters used to define the systematic categories include shape, colour, consistency, and surface (smooth, rough, or conulose) of the sponge and distribution and character of the oscula. These characters, best studied in living specimens, often have, as in the case of skeletal features, remarkable interspecies variability. Observations of living samples and estimates of the range of variability, however, may allow the identification of a species without microscopic examination of the skeleton. Cytological and embryological features are used as diagnostic characters in both general classification and species identification of the Demospongiae and Calcispongiae. Ecological and distributional characters are important in distinguishing species, particularly in groups (*e.g.*, haliclonids) in which skeletal and embryological characters are so uniform as to be of little taxonomic value. Although biochemical criteria—for instance, amino acid composition—have been used in some cases to evaluate general phylogenetic problems and separate orders, the results do not differ substantially from those obtained with other criteria.

Annotated classification. The classification below, which refers to the three living classes of sponges, is adapted from that of E. Topsent as modified by C. Levi on the basis of embryological data; it is accepted by most experts in the field.

PHYLUM PORIFERA (sponges)

Primitive aquatic invertebrate animals; about 5,000 species in all seas; attach to surfaces from intertidal zone to depths of 8,500 m or more; propel water containing food particles through a system of canals in body (filter feeding); composed of many cell types (archaeocytes of various kinds, pinacocytes, collencytes); structure of water current system variable (ascon, sycon, leucon); skeleton either mineral (calcium carbonate or silicic acid), spongin, sometimes lacking, or of mixed composition; skeletal units (spicules) variable in form; both sexual and asexual reproduction (gemmulation); ability to regenerate lost parts; contain many unique chemical compounds; form symbiotic relationships with many other types of organisms.

Class Calcispongiae

Skeleton of spicules of calcium carbonate; species either vase-shaped compact structures, loose networks of thin tubes, or irregular massive colonies; mostly small in size; inhabit shallow waters of all seas, from intertidal regions to depths of 200 m; a few species to 800 m; about 300 species.

Subclass Calcinea

Larva called parenchymula (solid, compact, with outer layer of flagellated cells, inner mass of cells); flagella of choanocytes (collar cells) arise independently of nucleus; some 3-rayed spicules in most species; water current system ascon, sycon, or leucon type; includes pharetronid sponges with rigid skeleton of fused spicules or of a calcareous network; genera include *Clathrina, Leucetta, Petrobiona* (a pharetronid).

Subclass Calcaronea

Larva called amphiblastula (oval in shape with front half of flagellated cells, rear half without flagellated cells); flagella of choanocytes arise directly from nucleus; spicules 3-rayed, with one ray characteristically longer than other two; water current system ascon, sycon, or leucon type; *Leucosolenia, Scypha* (formerly called *Sycon*), *Grantia, Lelapia* (with a rigid skeleton composed of bundles of modified rayed spicules).

Class Hyalospongiae (Hexactinellida)

Skeleton basically of hexactinal (6-rayed) siliceous spicules and lacking in spongin; exclusively marine, in deeper waters of all seas, depths from 25 to 8,500 m; commonly fixed firmly to a hard surface, some species anchored in soft bottom sediments; *Hexactinella, Aphrocallistes, Farrea, Dactylocalyx, Euplectella, Rhabdocalyptus*; about 500 species.

Class Demospongiae

Skeleton either of 1- or 4-rayed siliceous spicules, spongin fibres, or both; skeleton lacking in a few primitive genera; most abundant and widely distributed group of sponges (about 4,200 species); occur from intertidal regions to depths of about 5,500 m in seas; Spongillidae the chief freshwater sponge family; species vary greatly in form and size; range from thin encrustations several cm in diameter to huge cake-shaped species 2 m in diameter; many species with desmas (develop as a result of secondary deposits of silica around ordinary spicules); evolution independently among several orders of interlocking of adjacent desmas to form a stony skeleton (lithistid sponges).

Critical appraisal. The living orders of sponges are distributed in the Calcispongiae, Hyalospongiae, and Demospongiae. The Calcispongiae may be separated into two subclasses (Calcinea and Calcaronea), which represent two phyletic series separated mainly by cytological and embryological characters; this classification has replaced the traditional classification of the calcareous sponges into Homocoela and Heterocoela, which was based on the types of the water current system. The tradi-

tional classification of the Hyalospongiae into the orders Lyssacina and Dictyonina is still valid. The separation of the Demospongiae into orders is not well defined, because the phylogeny of the group is not well known. Two groups of orders may be distinguished within the Demospongiae. One, in which spongin is an important skeletal element (Ceractinomorpha), includes Poecilosclerida, Haplosclerida, Dendroceratida, and Dictyoceratida, the latter two derived from another subdivision, the Keratosa. The other group (Tetractinomorpha) includes Homosclerophorida, Tetractinellida, Epipolasida, Hadromerida, and Axinellida; four-rayed spicules are important only in the Homosclerophorida and in the Tetractinellida, but the other orders are clearly related to them. Hadromerida, Axinellida, and Epipolasida are placed together in a group called the Clavaxinellida. The Epipolasida, an ill-defined order, are nearest to the Tetractinellida. Many genera of Demospongiae, mainly in the Tetractinomorpha, are transitional, and their actual phylogenetic position is not clear.

Alternate subclassification systems

Other classifications of sponges may concern only differences in detail; *e.g.*, the rank of order or of suborder given to the Dictyoceratida and Dendroceratida, the inclusion of the Homosclerophorida in the Tetractinellida, or the abolition of the order Epipolasida. A classification for the Demospongiae, proposed in 1936, is similar to that stated above in its subdivision into orders but is revolutionary with regard to the placement of families and of genera. It has not been followed in general by the other systematists of the group.

Still another classification of the Porifera places them in a subkingdom of the Parazoa as two phyla (Nuda and Gelatinosa). The Nuda includes a single class and order, the Hexactinellida; the Gelatinosa is divided into two classes (Calcarea and Demospongia). The separation between Nuda and Gelatinosa is based only on the absence or presence of an intercellular matrix of fluids and fibrils around the net of cells. Although the proposal has not yet been evaluated critically, it seems justifiable to keep the Porifera in a single phylum because of the common structural basis regarding their many characters, (*e.g.*, anatomy, function, and embryology). Other proposals include establishment of a new order Pharetronida separated from the Calcinea and Calcaronea in the Calcispongiae, and the creation of a new class (Sclerospongiae) for the silicocalcareous or coralline sponges.

An important controversy in the systematics of the sponges concerns the value given to species described using traditional criteria. The use of species in classifying sponges is justified by the great variability of diagnostic characters, especially the spicules. Attempts to merge large groups of species into single systematic units have resulted in excessive lumping of species, which has been much criticized. In classifying sponges, therefore, one diagnostic feature (*e.g.*, characters of skeleton or spicules) is not sufficient; rather, all diagnostic characters should be considered.

BIBLIOGRAPHY

General: P.R. BERGQUIST and C.A. TIZARD, "Sponge Industry," in F.E. FIRTH (ed.), *Encyclopedia of Marine Resources*, pp. 665–670 (1969); W.G. FRY (ed.), "The Biology of the Porifera," *Symp. Zool. Soc. Lond.*, No. 25, (1970), a symposium covering the biology and paleontology of sponges; L.H. HYMAN, *The Invertebrates*, vol. 1, *Protozoa Through Ctenophora*, pp. 284–364 (1940), a thorough treatment of the Porifera; E.A. MINCHIN, "Porifera," in E. RAY LANKESTER, *Treatise on Zoology*, vol. 2 (1900), an ample, classical treatment of the Porifera; G.C.J. VOSMAER, *Bibliography of Sponges: 1551–1913* (1928), a complete bibliography to 1913.

Advanced: M. BURTON, *A Revision of the Classification of the Calcareous Sponges* (1963), a well-illustrated diagnosis and description of the calcareous sponges of the world; M. FLORKIN and B.T. SCHEER (eds.), "Porifera," in *Chemical Zoology*, vol. 2, pp. 1–76 (1968), reviews of various aspects of the biochemistry and physiology of sponges; W.D. HARTMAN, "A Reexamination of Bidder's Classification of the Calcarea," *Syst. Zool.*, 7:97–110 (1958), a critical analysis of the modern classification of Calcispongiae; M.W. DE LAUBENFELS, "Porifera," in *A Treatise on Invertebrate Paleontology*, pt. *E* (1955), a well-documented systematic account of fossil sponges; C. LEVI, "Etude des Halisarca de Roscoff: Embryologie et systématique des Démosponges," *Archs. Zool. Exp. Gén.*, 93:1–181 (1956), a fundamental work, in French, on the embryology of sponges and its connections with systematics; E. TOPSENT, "Spongiaires de l'Atlantique et de la Méditerranée, provenant des croisières du Prince Albert Ier de Monaco," *Résult. Camp. Scient. Prince Albert I*, 74:1–376 (1928), a well-illustrated monumental classic, in French; D.A. WEBB, "The Histology, Cytology and Embryology of Sponges," *Q. Jl. Microsc. Sci.*, 78:51–70 (1935), well-documented.

(M. Sa.)

Portugal

One of the oldest countries in Europe, with established frontiers since 1297, Portugal lies on the extreme western rim of the European Continent. It is bordered on the east and the north by Spain and on the west and the south by the Atlantic Ocean. A narrow rectangular country, occupying slightly more than 15 percent of the Iberian Peninsula. Portugal has a total area of 34,170 square miles (88,500 square kilometres). Despite its small area, the country displays a great diversity of geographical features and rises to 6,539 feet (1,993 metres) at its highest point on the Serra da Estrela. To the west and southwest lie the Atlantic islands of the Azores and Madeira, which are part of metropolitan Portugal. In addition, Portuguese overseas territories include the Cape Verde Islands, São Tomé e Príncipe, Portuguese Guinea, Angola, and Mozambique, all in Africa; Macau in Asia; and Timor in Oceania. The Portuguese consider these overseas provinces as integral parts of Portugal, a single national territory that is juridically one despite its geographical separation. The United Nations, however, defines the Portuguese overseas territories as non-self-governing territories.

One of the poorest countries in western Europe and still one of the most politically authoritarian, Portugal was showing signs of greater liberalization in the early 1970s. Some observers have noted that the war against the African liberation movements has placed a heavy strain on an otherwise potentially buoyant economy. As a member of the European Free Trade Association (EFTA) and the North Atlantic Treaty Organization (NATO) and by virtue of its overseas possessions, Portugal plays a much greater role in both European and world affairs than its small size would suggest.

This article covers only metropolitan Portugal; there are separate articles on the overseas territories. (See ANGOLA; CAPE VERDE ISLANDS; MACAU; MOZAMBIQUE; and PORTUGUESE GUINEA. For a related physical feature, see TAGUS RIVER. See also LISBON.)

THE LAND

Most of Portugal consists of lowlands, with only 11.6 percent of its area lying above 2,300 feet—and most of that north of the Tagus River (Rio Tejo), which cuts the country in two. Southern Portugal is a region of plains and low tablelands, of broad river basins, and of gently undulating country, with rare outcrops of hills and only one mountain range rising over 3,280 feet. In the north, however, over 90 percent of the area lies above 1,300 feet. Mountains more than 3,280 feet high rear up only 30 miles from the sea, and huge plateaus are cut by deep valleys.

Physical geography. *Minho.* In the northwest, the mountains of Minho province, surmounted by Mt. Larouco, 5,003 feet (1,525 metres) high, form an amphitheatre facing the Atlantic Ocean. The area is composed of metamorphic rocks (crystalline schists, slates, and quartzites) and intrusive or granitic rocks, marked by fault lines and hot springs. Parallel to the mountainous ridges, rivers such as the Lima and Cávado flow through deep gorges and flat-floored valleys. Along the coast a narrow beach rises gradually to meet the low hills, which lie just a few miles inland.

Trás-os-Montes e Alto Douro. Behind the mountains of Minho is the province of Trás-os-Montes e Alto Douro bordered on two sides by Spain. It is an extension of the Meseta, a high, partly fractured plateau (from 2,000 to 2,600 feet high) composed largely of ancient Precambrian rock, such as granites, schists, and slates, deeply gashed by rivers, which have carved canyons as deep as 1,600

MAP INDEX

Political subdivisions

Angra do Heroísmo 38·39n 27·13w
Aveiro 40·38n 8·39w
Beja 38·01n 7·52w
Braga 41·33n 8·26w
Bragança 41·49n 6·45w
Castelo Branco 39·49n 7·30w
Coimbra 40·12n 8·25w
Évora 38·34n 7·54w
Faro 37·01n 7·56w
Funchal 32·38n 16·54w
Guarda 40·32n 7·16w
Horta 38·32n 28·38w
Leiria 39·45n 8·48w
Lisbon (Lisboa) 38·48n 9·16w
Ponta Delgada 37·44n 25·40w
Portalegre 39·17n 7·26w
Porto 41·11n 8·36w
Santarém 38·50n 8·56w
Setúbal 38·37n 9·00w
Viana do Castelo 41·42n 8·50w
Vila Real 41·18n 7·45w
Viseu 40·39n 7·55w

Historical provinces

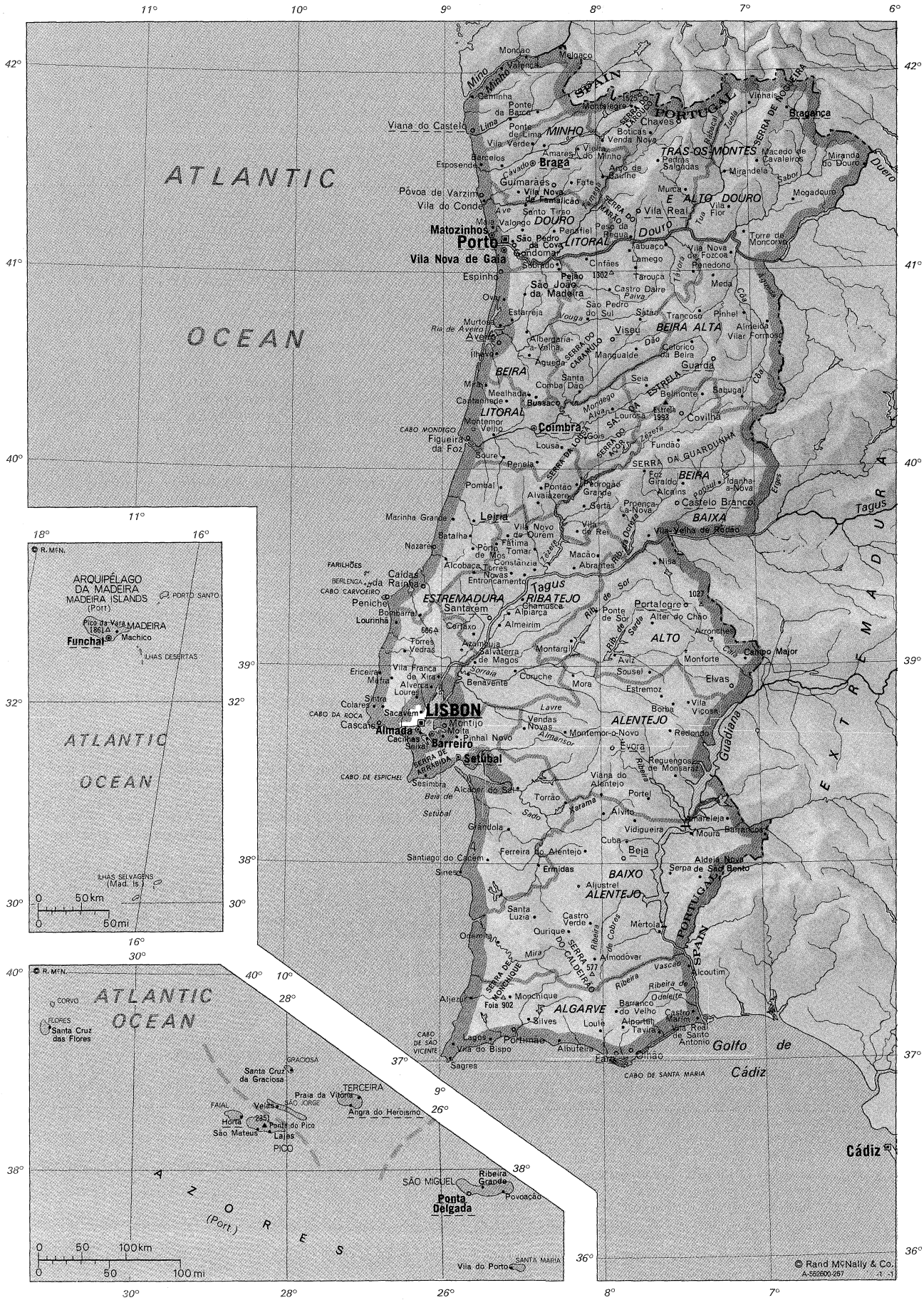
Algarve 37·10n 8·15w
Alto Alentejo 38·50n 7·40w
Baixo Alentejo 37·55n 8·10w
Beira Alta 40·40n 7·35w
Beira Baixa 39·45n 7·30w
Beira Litoral 40·15n 8·25w
Douro Litoral 41·05n 8·20w
Estremadura 39·15n 9·10w
Minho 41·40n 8·30w
Ribatejo 39·15n 8·30w
Trás-os-Montes e Alto Douro 41·40n 7·15w

Cities and towns

Abrantes 39·28n 8·12w
Águeda 40·34n 8·27w
Albergaria-a-Velha 40·42n 8·29w
Albufeira 37·05n 8·15w
Alcácer do Sal 38·22n 8·30w
Alcains 39·55n 7·27w
Alcobaça 39·33n 8·59w
Aldeia Nova de São Bento 37·55n 7·25w
Aljezur 37·19n 8·48w
Aljustrel 37·52n 8·10w
Almada 38·41n 9·09w
Almeida 40·43n 6·54w
Almeirim 39·12n 8·38w
Almodôvar 37·31n 8·04w
Alpiarça 39·15n 8·35w
Alportel 37·09n 7·53w
Alter do Chão 39·12n 7·40w
Alvaiázere 39·49n 8·23w
Alverca 38·54n 9·02w
Alvito 38·15n 7·59w
Amareleja 38·12n 7·14w
Amares 41·38n 8·21w
Angra do Heroísmo 38·39n 27·13w
Arco de Baúlhe 41·29n 7·58w
Arronches 39·07n 7·17w
Aveiro 40·38n 8·39w
Aviz 39·03n 7·53w
Azambuja 39·04n 8·52w
Barcelos 41·32n 8·37w
Barranco do Velho 37·14n 7·56w
Barrancos 38·08n 6·59w
Barreiro 38·40n 9·04w
Batalha 39·39n 8·50w
Beja 38·01n 7·52w
Belmonte 40·21n 7·21w
Benavente 38·59n 8·48w
Bombarral 39·16n 9·09w
Borba 38·48n 7·27w
Boticas 41·41n 7·40w
Braga 41·33n 8·26w
Bragança 41·49n 6·45w
Bussaco 40·23n 8·22w
Cacílhas 38·41n 9·09w
Caldas da Rainha 39·24n 9·08w
Caminha 41·52n 8·50w
Campo Major 39·01n 7·04w
Cantanhede 40·21n 8·36w
Cartaxo 39·09n 8·47w
Cascais 38·42n 9·25w
Castanheira de Pêra 40·00n 8·13w
Castelo Branco 39·49n 7·30w
Castro Daire 40·54n 7·56w
Castro Marim 37·13n 7·26w
Castro Verde 37·42n 8·05w
Celorico da Beira 40·38n 7·23w
Chamusca 39·21n 8·29w
Chaves 41·44n 7·28w
Cinfães 41·04n 8·05w
Coimbra 40·12n 8·25w
Colares 38·48n 9·27w
Constância 39·28n 8·20w
Coruche 38·57n 8·31w
Covilhã 40·17n 7·30w
Cuba 38·10n 7·53w
Elvas 38·53n 7·10w
Entroncamento 39·28n 8·28w
Ericeira 38·59n 9·25w
Ermidas 38·00n 8·23w
Espinho 41·00n 8·39w
Esposende 41·32n 8·47w
Estarreja 40·45n 8·34w
Estremoz 38·51n 7·35w
Évora 38·34n 7·54w
Fafe 41·27n 8·10w
Faro 37·01n 7·56w
Fátima 39·37n 8·39w
Ferreira do Alentejo 38·03n 8·07w
Figueira da Foz 40·09n 8·52w
Foz Giraldo 40·00n 7·43w
Funchal 32·38n 16·54w
Fundão 40·08n 7·30w
Góis 40·09n 8·07w
Gondomar 41·09n 8·32w
Grândola 38·10n 8·34w
Guarda 40·32n 7·16w
Guimarães 41·27n 8·18w
Horta 38·32n 28·38w
Idanha-a-Nova 39·55n 7·14w
Ilhavo 40·36n 8·40w
Lagos 37·06n 8·40w
Lajes 38·23n 28·16w
Lamego 41·06n 7·49w
Leiria 39·45n 8·48w
Lisbon (Lisboa) 38·43n 9·08w
Loulé 37·08n 8·02w
Loures 38·50n 9·10w
Lourinhã 39·14n 9·19w
Lourosa 40·19n 7·56w
Lousã 40·07n 8·15w
Macão 39·33n 8·00w
Macedo de Cavaleiros 41·32n 6·58w
Machico 32·42n 16·46w
Mafra 38·56n 9·20w
Maia 41·14n 8·37w
Mangualde 40·36n 7·46w
Marinha Grande 39·45n 8·56w
Matosinhos 41·11n 8·42w
Mealhada 40·22n 8·27w
Meda 40·58n 7·16w
Melgaço 42·07n 8·16w
Mértola 37·38n 7·40w
Mira 40·26n 8·44w
Miranda do Douro 41·30n 6·16w
Mirandela 41·29n 7·11w
Mogadouro 41·20n 6·39w
Moita 38·39n 9·00w
Moncão 42·05n 8·29w
Monchique 37·19n 8·33w
Monforte 39·03n 7·26w
Montalegre 41·49n 7·48w
Montargil 39·05n 8·10w
Montemor-o-Novo 38·39n 8·13w
Montemor o Velho 40·10n 8·41w
Montijo 38·42n 8·58w
Mora 38·56n 8·10w
Moura 38·08n 7·27w
Murça 41·24n 7·27w
Murtosa 40·44n 8·38w
Nazaré 39·36n 9·04w
Nisa 39·31n 7·39w
Odemira 37·36n 8·38w
Olhão 37·02n 8·50w
Oporto, see Porto
Ourique 37·39n 8·13w
Ovar 40·52n 8·38w
Pedras Salgadas 41·32n 7·36w
Pedrogão Grande 39·55n 8·09w
Pejão 41·00n 8·19w
Penafiel 41·12n 8·17w
Penedono 40·59n 7·24w
Penela 40·02n 8·23w
Peniche 39·21n 9·23w
Peso da Régua 41·10n 7·47w
Pinhal Novo 38·38n 8·55w
Pinhel 40·46n 7·04w
Pombal 39·55n 8·38w
Ponta Delgada 37·44n 25·40w
Pontão 39·55n 8·22w
Ponte da Barca 41·48n 8·25w
Ponte de Sôr 39·15n 8·01w
Ponte do Lima 41·46n 8·35w
Portalegre 39·17n 7·26w
Portel 38·18n 7·42w
Portimão 37·08n 8·32w
Porto (Oporto) 41·11n 8·36w
Pôrto de Mós 39·36n 8·39w
Povoação 37·45n 25·15w
Póvoa de Varzim 41·23n 8·46w
Praia da Vitória 38·44n 27·04w
Proença-a-Nova 39·45n 7·55w
Redondo 38·39n 7·33w
Reguengos de Monsaraz 38·25n 7·32w
Ribeira Grande 37·49n 25·31w
Sabugal 40·21n 7·05w
Sacavém 38·47n 9·06w
Sagres 37·00n 8·56w
Salvaterra de Magos 39·01n 8·48w
Santa Comba Dão 40·24n 8·08w
Santa Cruz da Graciosa 39·05n 28·01w
Santa Cruz das Flores 39·27n 31·07w
Santa Luzia 37·44n 8·24w
Santarém 39·14n 8·41w
Santiago do Cacém 38·01n 8·42w
Santo Tirso 41·21n 8·28w
São João da Madeira 40·54n 8·30w
São Mateus 38·26n 28·27w
São Pedro da Cova 41·09n 8·31w
São Pedro do Sul 40·45n 8·04w
Sátão 40·44n 7·44w
Seia 40·25n 7·42w
Seixal 38·38n 9·06w
Serpa 37·56n 7·36w
Sertã 39·48n 8·06w
Sesimbra 38·26n 9·06w
Setúbal 38·32n 8·54w
Silves 37·11n 8·26w
Sines 37·57n 8·52w
Sintra 38·48n 9·23w
Sobrado 41·02n 8·16w
Soure 40·03n 8·38w
Sousel 38·57n 7·40w
Tabuaço 41·07n 7·34w
Tarouca 41·00n 7·40w
Tavira 37·07n 7·39w
Tomar 39·36n 8·25w
Torrão 38·18n 8·13w
Torre de Moncorvo 41·10n 7·03w
Torres Novas 39·29n 8·32w
Torres Vedras 39·06n 9·16w
Trancoso 40·47n 7·21w
Valença 42·02n 8·38w
Valongo 41·11n 8·30w
Velas 38·41n 28·13w
Venda Nova 41·40n 7·58w
Vendas Novas 38·41n 8·28w
Viana do Alentejo 38·20n 8·00w
Viana do Castelo 41·42n 8·50w
Vidigueira 38·13n 7·48w
Vieira do Minho 41·39n 8·09w
Vila de Rei 39·40n 8·09w
Vila do Bispo 37·05n 8·55w
Vila do Conde 41·21n 8·45w
Vila do Porto 36·56n 25·09w
Vila Flor 41·18n 7·09w
Vila Franca de Xira 38·57n 8·59w
Vila Nova de Famalicão 41·25n 8·32w
Vila Nova de Fozcoa 41·05n 7·12w
Vila Nova de Gaia 41·08n 8·37w
Vila Novo de Ourém 39·39n 8·35w
Vila Real 41·18n 7·45w
Vila Real de Santo António 37·12n 7·25w
Vilar Formoso 40·37n 6·50w
Vila Velha de Ródão 39·38n 7·40w
Vila Verde 41·39n 8·26w
Vila Viçosa 38·47n 8·13w
Vinhais 41·50n 7·00w
Viseu 40·39n 7·55w

Physical features and points of interest

Açor, Serra do, *mountains* 40·13n 7·48w
Águeda, *river* 41·02n 6·56w
Almansor, *river* 38·56n 8·54w
Alva, *river* 40·18n 8·15w
Arrábida, Serra de, *mountains* 38·30n 9·00w
Atlantic Ocean 41·00n 10·00w
Ave, *river* 41·20n 8·45w
Aveiro, Ria de, *river mouth* 40·38n 8·44w
Azores, *islands* 38·30n 28·00w
Berlenga, Ilha, *island* 39·25n 9·30w
Cádiz, Golfo de, *gulf* 36·50n 7·00w
Caia, *river* 38·50n 7·05w
Caldeirão, Serra do, *mountains* 37·42n 8·21w
Caramulo, Serra do, *mountains* 40·34n 8·11w
Carvoeiro, Cabo, *cape* 39·21n 9·24w
Cávado, *river* 41·32n 8·48w
Côa, *river* 41·05n 7·06w
Cobres, Ribeira de, *river* 37·50n 7·52w
Corvo, *island* 39·42n 31·06w
Dão, *river* 40·20n 8·11w
Degebe, Ribeira de, *creek* 38·13n 7·29w
Desertas, Ilhas, *island* 32·30n 16·30w
Divôr, *river* 38·59n 8·29w
Douro, *river* 41·08n 8·40w
Espichel, Cabo de, *cape* 38·25n 9·13w
Estrela, Serra da, *mountains* 40·20n 7·38w
Faial, *island* 38·34n 28·42w
Farilhões, Ilhas, *islands* 39·28n 9·34w
Flores, *island* 39·26n 31·13w
Foia, *mountain* 37·19n 8·36w
Graciosa, *island* 39·04n 28·00w
Guadiana, *river* 37·14n 7·22w
Guardunha Serra da 40·05n 7·31w
Larouco, Serra do, *mountains* 41·55n 7·40w
Lavre, *river* 38·49n 8·40w
Lima, *river* 41·41n 8·50w
Lousã, Serra do, *mountains* 40·04n 8·13w
Madeira, *island* 32·44n 17·00w
Madeira Islands 32·40n 16·45w
Marão, Serra do, *mountains* 41·15n 7·55w
Minho, *river* 41·52n 8·51w
Mira, *river* 37·43n 8·47w
Monchique, Serra de, *mountains* 37·19n 8·36w
Mondego, *river* 40·09n 8·52w
Mondego, Cabo, *cape* 40·11n 8·55w
Montemuro, *mountain* 40·59n 7·55w
Nogueira, Serra de, *mountains* 41·42n 6·52w
Ocreza, Ribeira, *river* 39·32n 7·50w
Odeleite, Ribeira de, *river* 37·21n 7·27w
Paiva, *river* 41·04n 8·16w
Pico, *island* 38·28n 28·20w
Pico, Ponta do, *peak* 38·28n 28·25w
Ponsul, *river* 39·40n 7·31w
Porto Santo, *island* 33·04n 16·20w
Rabaçal, *river* 41·30n 7·12w
Roca, Cabo da, *cape* 38·47n 9·30w
Sabor, *river* 41·10n 7·07w
Sado, *river* 38·29n 8·55w
Santa Maria, *island* 36·58n 25·06w
Santa Maria, Cabo de, *cape* 36·58n 7·54w
São Jorge, *island* 38·38n 28·03w
São Miguel, *island* 37·47n 25·30w
São Vicente, Cabo de 37·01n 9·00w
Selvagens, Ilhas, *islands* 30·05n 15·55w
Serda, Ribeira de, *river* 38·56n 8·03w
Setúbal, Baía de, *bay* 38·27n 8·53w
Sor, Ribeira de, *river* 39·00n 8·17w
Sorraia, *river* 38·56n 8·53w
Tagus (Tejo), *river* 38·40n 9·24w
Tâmega, *river* 41·05n 8·21w
Távora, *river* 41·09n 7·35w
Tejo, Rio, see Tagus
Tera, *river* 38·56n 8·03w
Terceira, *island* 38·43n 24·13w
Tua, *river* 41·13n 7·26w
Tuela, *river* 41·30n 7·12w
Vara, Pico da, *peak* 32·45n 16·56w
Vascão, Ribeira do, *river* 37·31n 7·31w
Vouga, *river* 40·41n 8·40w
Xarama, *river* 38·14n 8·20w
Zêzere, *river* 39·28n 8·20w


ATLANTIC OCEAN
SPAIN
PORTUGAL
Viana do Castelo
Braga
Bragança
Vila Real
Matozinhos
Porto
Vila Nova de Gaia
Aveiro
Viseu
Coimbra
Figueira da Foz
Guarda
Castelo Branco
Leiria
Tagus
Nazaré
Caldas da Rainha
Peniche
Santarém
LISBON
Almada
Cascais
Sintra
Setúbal
Cabo da Roca
Cabo de Espichel
Évora
Beja
Portalegre
Elvas
Estremoz
Sousel
Montemor-o-Novo
Vendas Novas
Alcácer do Sal
Sesimbra
Serpa
Moura
Mértola
Castro Verde
Ourique
Almodôvar
Aljustrel
Odemira
Aljezur
Monchique
Foia 902
Silves
Loulé
Albufeira
Lagos
Portimão
Vila do Bispo
Sagres
Cabo de São Vicente
Faro
Olhão
Vila Real de Santo António
Golfo de Cádiz
Cádiz
Tavira
Castro Marim
Alcoutim
Vascão
Ribeira de Odelouca
Barranco do Velho
Alportel
Almansor
Sado
Lavre
Xarrama
Ribeira de Cobres
Ribeira de Mira
Ribeira de Aljezur
Ribeira de Seixe
Guadiana
Caia
Tejo
Sorraia
Ribeira de Sor
Ribeira de Sorraia
Ribeira de Divor
Ribeira de Tera
Ribeira de Alcarrache
Ribeira de Sever
Ribeira de Nisa
Ribeira de Degebe
Ribeira de Vascão
Ribeira de Odeleite
Ribeira de Beliche
Ribeira de Carrapatelo
Ribeira de Alportel
Ribeira de Quarteira
Ribeira de Almargem
Ribeira de Boina
Ribeira de Alcantarilha
Ribeira de Algibre
Ribeira de Arade
Ribeira de Odiáxere
Ribeira de Bensafrim
Ribeira de Alvor
Ribeira de Odelouca
Ribeira de Arade

feet. The Rio Douro at Paradela suddenly curves to the southwest, cutting deeply into the old plateau surface. The plateau is surmounted by Nogueira (4,324 feet [1,318 metres] high), one of the remnants of the ancient mountains.

The Beira provinces. The rocks and landscape of Minho extend southward beyond the Rio Douro into the Beira Litoral, to a point north of Coimbra on the Rio Mondego. Most of central Portugal is included in the Beira provinces. The heart of Beira Alta is an elevated peneplain, partly dissected by the Mondego and Vouga river systems. Like Trás-os-Montes e Alto Douro, this high plain is an extension of the Meseta of Old Castile, Spain. The low, protected valleys, which are subject to climatic conditions very different from those of the high plateau directly above them, are called the *terra quente* (hot country). The high plain of the Beira is bounded to the northwest by the Serra do Caramulo (3,524 feet [1,074 metres]), the Montemuro Mountains (4,534 feet [1,382 metres]), and their foothills. To the south and southeast it is bordered by the great escarpment of the Serra da Estrela, which forms the backbone of the Beira as a gigantic granite and schist block that acts as a barrier between the northern and southern parts of Portugal. To the southwest, the mountain chain extends to the Serra de Açor (4,623 feet [1,409 metres]) and the Serra da Lousã (3,943 feet [1,202 metres]). To the south of Beira Alta and to the east of the Serra da Estrela lies the bleak plateau of Guarda, which has an average height of almost 3,000 feet. Farther south the eroded plateau of Beira Baixa drops to an altitude of from 1,600 feet to 700 feet. It slopes toward the Tagus River in the south and merges almost imperceptibly into the highlands of Alto Alentejo.

The role of Serra da Estrela

Beira Litoral and Estremadura. North of Lisbon, the western provinces of Estremadura and Beira Litoral form a region totally without crystalline rocks. A narrow band of Triassic sandstones run north and south from Coimbra to Tomar. To the west and southwest of the sandstones lies a broad band of Jurassic limestones, interspersed by large areas of Cretaceous sandstones and conglomerates. Erosion has produced a landscape of rolling sandy hills, plateaus of basalt, steep escarpments, and arid, desolate slopes.

Alto Alentejo. The south of Portugal is characterized by the rolling countryside of the Alentejo, formed from ancient schistose, granitic, and marble rocks. With an average height of between 700 and 1,600 feet, Alto Alentejo slopes downward toward the west and south.

Ribatejo. Between Beira Baixa, Estremadura, and Alentejo lies the Ribatejo, a part of a large sedimentary basin forming the lower section of both the Tagus and Sado river systems. The northern edge of the Rio Sado faces the calcareous cliffs of the Serra da Arrábida (1,644 feet [501 metres]) and a long spit of sand dunes extends to the south.

Baixo Alentejo and the Algarve. The southern edge of the plain of Baixo Alentejo rises into the schistose mountains of the Serra do Caldeirão (1,893 feet [577 metres]) in the east. To the west lie the more extensive scarps and hills of the Algarve, composed of Mesozoic limestones and sandstones. The highest peak is the Monchique (2,959 feet [902 metres]), a dissected massif of syenite, an igneous rock.

Rivers and coasts. *River systems.* All the main rivers of Portugal flow from Spain over the edge of the Meseta in a series of defiles and gorges. Consequently none of them is navigable between the two countries (with the possible exception of the lower Minho and Guadiana). The longest river crossing Portugal is the Douro (200 miles), whereas the longest river solely in the country is the Mondego (137 miles), rising in the Serra da Estrela. The principal exclusively Portuguese rivers, all of which are partly navigable, include the Vouga (85 miles), the Mondego, the Sado, and the Zézere (a tributary of the Tagus). North of the Tagus, most of the rivers flow from the northeast to the southwest, whereas most of those in the south flow from the southeast to northwest.

Harbours. The coast has only a few natural harbours, the most important at the mouths of rivers: Lisbon, at the Tagus; and Setúbal, at the Sado. Other harbours depend on the protection of headlands. North of the mouth of the Douro the artificial harbour of Leixões has been built. Sandbars at the entrance of many rivers block passage of ships. Lisbon, with its fine harbour, was the only obvious choice for a large port. The changing fortunes of the ports of the Algarve and the Beira Litoral have reflected the variations of silting along their coasts. Long sandbars extend along most of the Portuguese coast, interrupted by rocky capes and by great lagoons.

The climate. The extreme southwesterly position of Portugal in Europe explains its mild, humid winters and its relatively equable, dry summers. The effect of the ocean and the contrasts of relief offer an extremely varied climate marked by transitional features between west temperate (Marine West Coast) and Mediterranean conditions. In winter the northern half of Portugal is dominated by the North Atlantic cyclones, and in summer the Azores high-pressure system advances northward over the country.

Temperature ranges

In January the average temperatures range from 52° F (11° C) in the southwest to about 45° F (7° C) in the northeast. Severe cold is usually associated with the incursion of high pressure conditions from the Spanish Meseta or from Siberian anticyclones. The highest areas of the Serra da Estrela, or the region of Bragança (to the northeast of Trás-os-Montes e Alto Douro) register temperatures of less than 32° F (0° C), and snow lies on the mountain summits for several months. Evening sea breezes cool the summer heat of the coastal zones. August is usually the hottest month, with an average temperature of 68° F (20° C).

Northern Portugal receives considerable rainfall, and everywhere on the coast humidity is relatively high for the latitude. In the northwest annual rainfall is over 40 inches, and a large part of Minho receives between 40 and 80 inches, with mountain slopes recording up to 100 inches. Near the northwestern frontier, the summer drought is not pronounced, but toward the northeast and especially south of the Tagus the intensity of summer drought is more marked. Porto has a mean annual rainfall of 45 inches, falling within 137 rainy days; Lisbon 24 inches in 99 days; Lagos 15 inches in 90 days. Relief also plays its role, and the Serra da Estrela receives more than 110 inches. The leeward position of Trás-os-Montes e Alto Douro and the Douro Valley partly explains their lower rainfall of 30 to 40 inches in 90 to 130 rainy days.

Vegetation. Portugal exhibits a mixture of Atlantic deciduous flora and Mediterranean and African evergreens. About two-thirds are general European species and about one-third Iberian and African.

Two species of pine (*Pinus pinaster* and *Pinus pinea*) and three species of oak (*Quercus robur*, or english oak; *Quercus toza*, or Pyrenean oak; and *Quercus lusitanica*, or Portuguese oak) predominate in the northern half of Portugal, with notable concentrations of maritime pine on the predominantly siliceous coastal soils. The chestnut, linden, elm, and poplar, like the olive, are widespread. South of the Tagus River there is a quite different vegetation region. Maritime pine is replaced in the western Alentejo by the Italian stone pine (*Pinus pinea*). The dominant trees, however, are the cork and holm oaks. Both cultivated and wild olives grow well.

The far southern Algarve is renowned for its groves of carobs (locusts, or St. John's bread), almonds, and figs, but there are also figs near Tomar and almonds in the upper valley of the Douro in the north. The olive, once concentrated in the south, is now widespread, especially in the Algarve, central Portugal, and the Douro Valley. The Australian eucalyptus flourishes in the poorly drained valleys of the Tagus and the Sado systems. Notable specialized flora include the aquatic plants of the Aveiro lagoons, the beautiful woodlands of the Serras do Arrábida, Sintra, and Bussaco, and the alpine flora of the Serra da Estrela.

THE PEOPLE

Historical origins. *Racial groups.* The Portuguese display diverse racial origins. The Moorish occupation has

left its mark in the dark eyes and sallow skin of many Portuguese, in contrast with the blue- or gray-eyed people in the north. Along the coasts and especially in the south, Negro characteristics are a reflection of the 18th-century slave trade.

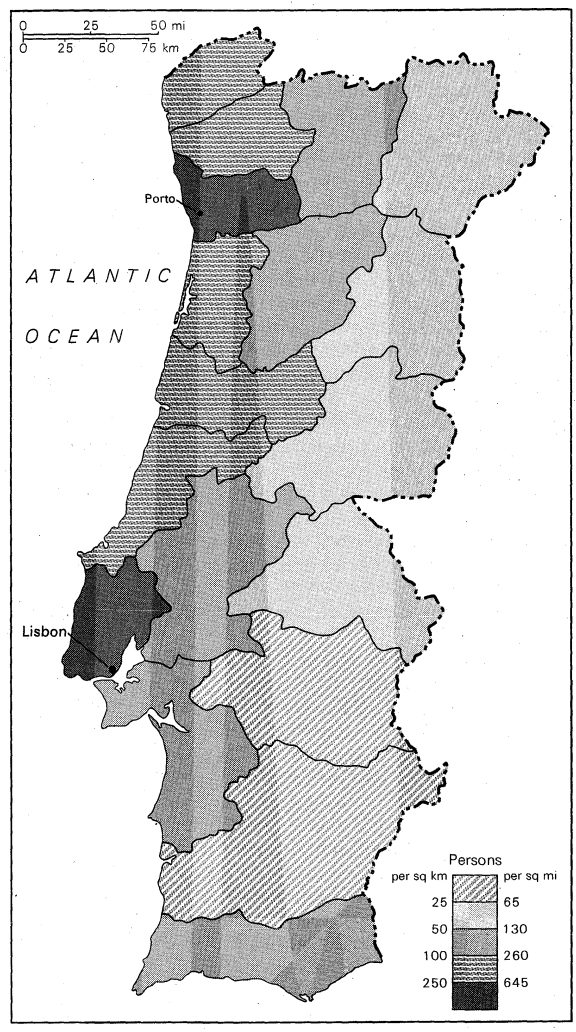
Linguistic features. Although the language appears superficially rooted purely in Latin, with a great similarity to Spanish, it contains a number of words of Germanic and Semitic origins. The language is remarkably uniform throughout Portugal, the only true dialect existing at Miranda do Douro in the northeast.

Religious groups. The vast majority of Portuguese belongs to the Roman Catholic Church, and, although freedom of worship is allowed, there are few adherents to other creeds.

Portugal, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census*
Arquipélago da Madeira				
District				
Funchal	308	797	268,000	253,000
Arquipélago dos Açores				
Districts				
Angra do Heroísmo	274	710	97,000	90,000
Horta	301	780	49,000	41,000
Ponta Delgada	330	854	182,000	159,000
Continente				
Districts				
Aveiro	1,046	2,708	522,000	546,000
Beja	3,954	10,240	269,000	202,000
Braga	1,054	2,730	594,000	617,000
Bragança	2,527	6,545	230,000	177,000
Castelo Branco	2,588	6,704	311,000	252,000
Coimbra	1,527	3,956	434,000	396,000
Évora	2,855	7,393	215,000	176,000
Faro	1,958	5,072	313,000	267,000
Guarda	2,122	5,496	276,000	212,000
Leiria	1,357	3,516	400,000	383,000
Lisboa	1,066	2,762	1,403,000	1,612,000
Portalegre	2,274	5,889	184,000	145,000
Porto	881	2,282	1,192,000	1,315,000
Santarém	2,583	6,689	462,000	435,000
Setúbal	1,989	5,152	376,000	464,000
Viana do Castelo	814	2,108	275,000	250,000
Vila Real	1,637	4,239	323,000	265,000
Viseu	1,938	5,019	477,000	410,000
Total Portugal	35,383†	91,641	8,851,000‡	8,668,000

*Preliminary. †Converted area figures do not add to total given because of rounding. ‡Figures do not add to total given because of rounding.
Source: Official government figures.



Population density of Portugal.

Demography. *Population, births, and mortality.* At the beginning of the 20th century the population of metropolitan Portugal had reached about 5,000,000. By 1930 it had risen to 6,400,000, and at the 1960 census the total was 8,851,000 representing an average density of 242.4 persons per square mile. Portugal's population had risen to about 9,000,000 by the early 1970s; the birth rate stood at about 20 per 1,000 and the mortality rate at 11.1 per 1,000. Although the infant mortality rate had decreased, at 59.2 per 1,000 it was still one of the highest in Europe. About 38 percent of the population is under 20 years of age, 29 percent between 20 and 40 years of age, 21 percent between 40 and 60, and 12 percent over 60 years of age.

Birth and death rates

Migration. Portugal has one of the highest rates of emigration in Europe. Before 1960 most of its migration was to Brazil and a few other Latin American countries. In 1965 about 100,000 persons of eligible working age left the country, while the natural population increase was 115,112. By 1970 emigration to foreign countries stood at 66,360, 33 percent of whom went to France, 30 percent to the Federal Republic of Germany, 0.5 percent to the United States and 17 percent to other North and South American countries (Brazil, Canada, and Venezuela).

Population distribution. Population and settlement reveal striking contrasts between the denser northern and the more sparse southern parts of the country. A number of rural areas have suffered considerable population losses resulting in economic and social depression, particularly in parts of the north, the central east, and the southern coastal areas. The coastal zones between Braga in Minho, and Setúbal just south of Lisbon, with their low-lying plains and urban development, have encouraged a large proportion of the population to settle there. Only the industrial areas of Lisbon, with about 19 percent of the country's population, and Porto, with about 17 percent, are able to absorb their own working populations. Other areas that are currently going through the process of industrialization are Aveiro, Coimbra, Setúbal, and Faro.

Areas such as the humid Minho, the fluvial central plains, and the coastal areas of the Algarve are seriously overpopulated. In the main, however, rural settlement is dispersed, with the inhabitants living in small villages under a system of open-field farming. The Beira Litoral and Estremadura have settlements varying between dispersed and clustered farmsteads. In the region of Aveiro, clusters of farmsteads and other dwellings are strung along roads in strips, which are often of considerable length and density. Fishing, one of the earliest occupations of Portugal, still plays an important role in coastal settlement.

At the time of the 1960 census approximately 23 percent of the population lived in urban communities; by 1970 the percentage had risen sharply to over 37. The growth of Lisbon and Porto has been particularly significant. This urban growth has been achieved at the cost of an equally significant rural depopulation.

Population growth

THE NATIONAL ECONOMY

Agriculture, forestry, and fishing. The picture of Portuguese agriculture is one of a sector in a state of decline

from its past pre-eminence and dwindling as a source of production. Two distinct factors explain this downward trend: first, the incidental and short-term influence of relatively adverse weather conditions and, secondly, the predominance of antiquated agricultural techniques. Small farms strongly predominate particularly in the north where they are too small and made up of too many dispersed holdings to allow integrated farming and rational crop rotation. In the south (except in Algarve) the system of *latifundia* (large estates) prevents more intensive cultivation. Half of the farms in metropolitan Portugal have an area of less than 2.5 acres, and almost 90 percent of them have less than 12 acres. Only 0.3 percent of Portuguese farms, occupying 39 percent of the cultivated area, have more than 494 acres; only three have more than 37,000 acres. As well as the problem of farm size, there is also that of land fragmentation. Although the average size of a complete farm is about 20 acres, the average size of the individual holding may be only 1.4 acres. About 30 percent of the total number of farms are made up of from six to more than 20 non-adjacent holdings. Fewer than two-thirds of the farms are entirely cultivated by their owners; a little over one-sixth are cultivated partly by the owner and partly under rent or sharecropping. Despite the small size and fragmentation, mechanization of agriculture has been slowly progressing.

The migration of workers from the countryside to the towns and abroad affects the various regions of the country in different ways. There is still, however, an excessively high proportion of labourers in agriculture, productivity is quite low, and the late 1960s indicated stagnation rather than a steady improvement of yields.

Efforts have been concentrated in Portugal on costly irrigation projects and the introduction of more efficient techniques of cultivation and farm management to obtain a substantial increase in agricultural productivity. Whereas internally agriculture is feeling the impact of developments in other sectors of the economy, externally it has to face the prospect of being plunged into competition with more advanced methods of agriculture.

Most of the mountainous areas of Portugal are well suited to forestry, and forest products (cork, resins, and pine and eucalyptus lumber) have increased their market value and find a growing demand in industry. For these reasons considerable reforestation has been conducted since the last quarter of the 19th century in areas where yields are low and erosion is rapidly destroying soil fertility. The area of forest land exceeds 7,900,000 acres.

Importance of the fishing industry

Portugal's extensive coastline and the abundance of fish in the surrounding waters have favoured the development of the fishing industry. Sardines, anchovies, and tuna caught near the coasts, together with other species such as cod from the North Atlantic fishing grounds, make a large contribution to food supplies. The fishing industry has prospered, and its products are exported all over the world. There are about 35,309 registered fishermen and about 20,000 additional persons employed in the fishing industry. Matosinhos, Setúbal, Portimão, and Olhão are the main fishery centres.

Industry and power. In the early 1970s, 28 percent of Portugal's active population was still occupied in agriculture, forestry, and fishing, 42 percent was engaged in industry, and 29 percent in the service sector: in 1950, the respective percentages were 49, 25, and 26. Since World War II there has been a decisive change in the economic life of Portugal. New occupations are encouraged as the population increases, and government initiatives aim to raise the standard of living.

Three main groups of industries

In its level of industrial development, Portugal occupies an intermediate, but fairly favourable position among the southern European countries. Since the 1930s attempts have been made to broaden the country's industrial base. Three main groups of industries—foodstuffs and beverages, textiles, and metals—together produce over half of the country's total industrial product and employ about 55 percent of the industrial labour force. Capital intensity varies widely among the industries included in these broad groups, and industrial centres are unevenly distributed throughout the country. The more prosperous and rapidly growing industries are concentrated around the larger towns, especially Lisbon and Porto. The construction of a steel mill at Seixal, the great new shipyards at Cacilhas-Barreiro, the installation of vast chemical plants at Sacavém, and the opening of two oil refineries at Lisbon and Porto have accentuated this industrial concentration.

Portugal has only a small and quite inadequate mineral production. Iron ores and coal are extracted at Torre de Moncorvo (hematites), Marão (magnetites), Montemor-o-Novo (magnetites and iron pyrites), São Pedro da Cova and Pejão (anthracites), and Bussaco (black coal). Other metallic ores, such as wolfram (tungsten ore), cassiterite (tinstone), and copper pyrites are found in small scattered deposits.

Hydroelectric resources are being developed, and a number of projects have been realized in the valleys of the Cávado, Douro, and Zêzere, among others. About 80 percent of the electrical production comes from hydro plants, the remainder from thermal plants. The total consumption of electric power had reached over 7,000,000,000 kilowatt-hours by the early 1970s, compared to only 2,797,650,000 in 1960.

Foreign trade. Portugal has a large foreign trade relative to its total population, and its composition changed considerably during the 1950s and 1960s.

Portugal's foreign trade has shown an annual increase in imports of nearly 10 percent. Food imports have not risen much, but imports of chemicals and fuel have continued to increase rapidly, as well as some groups of manufactured articles, especially motor vehicles and electrical machinery and appliances. The annual increase in exports has been about 12 percent, sustained by a fairly significant expansion in the sales of basic and semifinished materials. Some of the more important exports include cork, cotton yarn and textiles, canned fish, and wines. Portuguese exports to the main countries of the European Economic Community (EEC) have shown the highest growth rates, despite an indifferent performance in exports to Great Britain. Exports to other Western European countries have also increased rapidly. Shipments to Portugal's overseas territories, which account for about one-quarter of total exports, have accelerated somewhat, whereas deliveries to the United States have remained virtually static.

Export commodities

Portugal nevertheless suffers from a significant trade deficit. As a result of a substantial improvement in the balance of invisible trade, including expenditures by the growing influx of foreign tourists, the surplus in current transactions has, however, increased considerably.

National finance. The Portuguese monetary unit is the escudo (100 centavos) with the conto equivalent to 1,000 escudos. The official exchange rate was stabilized at $1 U.S. to 28.75 escudos in 1949, but after the 1971 international monetary reorganization the currency was being allowed to float.

The only bank issuing currency is the Bank of Portugal, founded in 1846. More than 20 commercial banks operate in Portugal and the islands of Madeira and the Azores. In 1959 the Bank of Development was set up to act as an investment bank for both Portugal and the overseas territories.

Monetary policy. The main aim of monetary policy has been price stability. Taxation of industry and incomes is moderate compared to that of other European countries, and there are provisions for a high rate of depreciation on equipment. The channelling of savings to long-term investment is still a weak feature of the system.

The Third Economic Development Plan

Economic prospects. The government's Third Economic Development Plan, covering the period 1968 to 1973, aimed at an average annual growth of real gross domestic product (GDP) of 7 percent. The investments specified for metropolitan Portugal totalled 123,000,000,000 escudos. To achieve the overall growth targets of the plan it was estimated that real gross fixed-asset formation would have to be increased at an annual rate of 8.5 percent, with the total value of fixed assets reaching 22.5 percent of the gross national product (GNP) in 1973 (at 1963 prices), compared with 19 percent in 1967.

Transportation. Although transport and communications have been seriously neglected in the past, there has been a strong effort in recent decades to improve this sector of the economy. The nature of the landscape dictates that metropolitan Portugal relies on many forms of transport.

Roads. Several of the country's main roads date back to ancient times, but now 11,185 miles (18,000 kilometres) of main roads connect the largest towns and extend to the border and ports. Secondary roads link the towns with almost every part of the interior. The total road network of more than 20,000 miles is being extended. By the early 1970s there were over 700,000 motor vehicles, most of which were automobiles.

Railroads. The Portuguese railroad network consists of almost 3,000 miles of track. The main Lisbon–Porto line was electrified in 1965, and by the early 1970s the rail service was accounting for more than 2,200,000,000 passenger-miles and 440,000,000 ton-miles of freight a year.

Shipping. The merchant fleet is one of considerable size, constituting a total of 926,000 gross tons in the early 1970s, compared to a little over 500,000 gross tons in the early 1950s. Well over 8,000 vessels stop at Portuguese ports annually.

Air transport. The Portela de Sacavém airport at Lisbon is an important junction for intercontinental flights and is also the base of the Portuguese airline, Transportes Aéreos Portugueses (TAP), which provides regular service to the overseas territories, as well as the usual commercial routes. There are other important airports at Porto and Faro. By the early 1970s almost 900,000 passengers a year were travelling by Portuguese airlines and more than 22,000,000 ton-miles of freight accounted for.

(I.M.P. do A.)

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of the government. *Constitutional framework.* Portugal has been a republic since the overthrow of the monarchy in 1910, and is administered as a corporative state. According to the constitution of 1933, which was last revised in 1971, the chief of state is the president of the republic. He is indirectly elected by the nation through an electoral college for a seven-year term. The president appoints the chairman of the Council of Ministers (the prime minister), fixes election dates, represents the nation and directs foreign policy, makes and adjusts international treaties, and promulgates the laws and resolutions of the National Assembly.

Council of State and Council of Ministers

The president of the republic is advised when necessary by a Council of State, which is composed of the prime minister and the presidents of the leading political institutions, with ten public men. The prime minister forms his own cabinet (Council of Ministers) with the president's approval. The cabinet's duties consist of endorsing the acts of the president, drawing up laws, approving international conventions and treaties, and generally supervising public administration.

Parliament. The parliamentary body is the National Assembly of 150 members who are elected every four years by direct suffrage. Its duties include discussing and voting upon basic legislation. Bills passed by the National Assembly are sent to the president for promulgation as law. A Corporative Chamber acts as an advisory and technical body, expressing its opinion on bills and drafts of laws, and on the ratification of treaties, conventions and international agreements. It is composed of more than 200 representatives of the existing corporations (comprising all those engaged in any one trade or industry, the employers in guilds [*grêmios*] and the workers in syndicates [sindicatos]), local autonomous bodies, and other national interests.

Other institutions. There are eight economic corporations, covering agriculture, industry, trade, transport and tourism, credit and insurance, entertainments, fisheries and canned goods, and press and printing. Employers and workers collaborate on the drafting of contracts and collective wage agreements, the solving of labour and wage problems; they are members of employers' associations, trade unions, rural Peoples' Centres, and Fishermen's Centres. There is also a corporative organization of social and cultural bodies in the sectors of literature, science, welfare, physical education, and sports.

Local and regional government. Portugal is divided into 18 administrative districts (*distritos*), each with a civil governor who is appointed by and responsible to the central government through the minister of the interior. He administers with the help of a board. The country is also divided into 11 provinces: Minho, Trás-os-Montes e Alto Douro, Beira Litoral, Beira Alta, Beira Baixa, Douro Litoral, Estremadura, Ribatejo, Alto Alentejo, Baixa Alentejo, and Algarve, but this provincial division is no longer used administratively.

Townships and parishes

The 18 districts are subdivided into municipalities, or *concelhos*, and parishes, or *freguesias*. Each municipality is administered by a town council (câmara municipal), the president or mayor of which is a government appointee and whose members are elected by local bodies. Each parish has its council elected by the direct vote of local citizens.

The political process. The president of the republic is elected every seven years by the Electoral College composed of members of the National Assembly, the Corporative Chamber, and representatives of the municipalities and of the governor of the legislative councils of the overseas provinces. Elections for the 150 deputies to the National Assembly take place every four years by direct suffrage. They are preceded by a period for electoral campaigning during which lists of candidates are put up by each constituency. Voting takes place on each list as a whole. Opposition groups (which have no legal political status) are allowed to put up candidates and run an electoral campaign during this period. Lists of government candidates are put up by the Acção Nacional Popular, a nationwide political group that supports government policy, and of which the prime minister is head. The various groups of the opposition (except the outlawed Communists) also put up their lists of candidates in the constituencies they choose. This practice, with the exception of Lisbon and Porto, is confined to areas in which there is a basic liberal or left-wing tendency. The elector casts his vote for the whole list.

Portugal is thus a one-party state with one group giving support to the government. Supporters of the former monarchy are allowed official status as a nonparty group, but democratic opposition has no legal status outside the electoral periods for the National Assembly, which occur every four years. The Portuguese Communist Party is outlawed and only operates clandestinely.

The electorate is restricted to citizens who have reached 21 years of age or have been emancipated by law, who can read and write or who pay a minimum annual tax and have appeared on an official census. Both men and women who fulfill these conditions can vote. Those who have been deprived of their civil rights for any criminal offense may not vote.

The courts. Portuguese justice is carried out independently of the legislature and government through a hierarchy of courts. These range from "courts of the first instance" and courts of appeal to the Supreme Court of Justice. They deal with criminal, civil, and commercial misdemeanours. Special courts handle juvenile-delinquency and domestic-affairs cases, and members of the armed forces are tried by military tribunals. Special courts also review the legality of decisions taken by administrative authorities. Judges are appointed for life by the Higher Council of the Judicature after competitive examination before a panel of judges and professors of law. The attorney general is appointed by the minister of justice, on whose ministry he is dependent, and is represented in all courts and tribunals.

Penal institutions. There is no death sentence in Portugal, and long terms of imprisonment may be commuted after half the term has been served. Imprisonment may be in a political jail, such as the Fort of Peniche; in one of the civil prisons in an urban centre; or in a penal colony. In cases such as juvenile delinquency, prisoners may be remanded to their homes. Political prisoners may be subject to "security measures," as, for example, further terms

Detention of political prisoners

of detention after completing their sentence. A citizen may be detained without trial on serious suspicion of a political crime.

The armed forces. The Ministry of Defense coordinates the country's armed forces, which consist of an army, navy, and air force. Subsidiary organizations exist such as the National Republican Guard (Guarda Nacional Republicana), the Customs (Guarda Fiscal), and the voluntary, paramilitary Portuguese Legion, which is in charge of civil defense. Besides the minister of defense there are ministers of the army and the navy and a secretary and under secretary of state for air.

All male citizens over the age of 18 are subject to compulsory military service. Suspensions may be granted for family, health, or educational and professional reasons. The service consists of between three and four years, two of which consist of active service against the African liberation movements in Angola, Portuguese Guinea, and Mozambique. The reserve list includes men up to the age of 45.

Portugal is divided into seven military regions and eight independent military commands. Angola and Mozambique constitute separate military regions. As a member of NATO, Portugal provides the headquarters for the Iberian Atlantic (Iberlant) Command and various munitions depots and communications centres. An air base at Lajes in the Azores is used by the United States, which under an agreement signed in December 1971 is allowed to remain there until February 1974 in return for more than $500,000,000 in aid to Portugal.

The social sector. *Education.* Primary education in Portugal is compulsory between the ages of six and 14. Illiteracy was officially computed to be 75.9 percent in 1880. By 1950 it was 40.4 percent, by 1960, 32 percent, and by 1970, 21 percent. Private schools supplement the state schools, which provide free education for the majority of people. Secondary training in a high school, or *liceu,* prepares pupils for the universities and for specialized educational establishments such as technical and commercial colleges and teacher-training centres.

The universities

There are five universities: Lisbon Classical University, Lisbon Technical University, Coimbra University (which is one of the oldest in Europe), Porto University, and the Catholic University, founded in 1968, in Lisbon. Besides the universities there are other establishments for higher education such as the Military School, Naval School, and School of Fine Arts. In 1971 the government founded an educational scheme (Acção Social Escolar) to help underprivileged students with such matters as fees, transport, and meals. A youth movement, the Mocidade Portuguesa, was also founded in 1936 by the government chiefly to direct extra-curricular school activities and provide moral and political training for both primary and secondary schoolchildren. Until 1971 membership was compulsory, but in October of that year it was made optional.

Welfare services. The Portuguese welfare system is divided into four types of institutions that insure workers against sickness, disablement, and old age and provide for payment of pensions and family allowances. The first type consists of the Trade Union Provident Funds, which include the People's Centres (Casas do Povo) and Fishermen's Centres (Casas dos Pescadores). The second type is made up of the welfare funds (Fundos de Previdência) for other employee categories. These two are compulsory, unlike the third type—the Mutual Assistance Associations, which have voluntary membership. Finally, there are the provident institutions for the military and civil servants. Both employers and employees contribute to these funds on a basis of about 17 percent and 6.5 percent of wages, respectively. Many big companies conduct their own welfare and sick-benefit schemes and pensions for their employees.

Housing. In general the building of houses in Portugal is financed by private enterprise. Official schemes, however, include low-cost houses for white-collar employees, trade-union members, and public and civil servants. They are built with the cooperation of the government, the municipalities, and the corporative organizations. There are also houses on fixed low rentals that can be purchased by the tenant on a long-term basis and limited-rental houses built in groups by private enterprise with the consent of the municipalities. Institutions such as the Salazar Foundation obtain funds from public benefactors to provide homes for the very poor. A special type of low-rent housing is available for fishermen and is controlled by a central board.

Low-rent housing

Police services. The Portuguese police are under the control of the Ministry of the Interior and are divided into three main categories: the Public Security Police (Polícia de Segurança Publica), the General Security Directorate (Direcção Geral de Seguranca), which replaced the International State Defense Police (Polícia Internacional e de Defesa do Estado) in 1970, and the road police, which in 1971 became a special branch of the National Republican Guard (Guarda Nacional Republicana), which is also responsible for rural armed guards who patrol the countryside and have outposts in the main townships and larger villages. The duties of the Public Security Police are the general maintenance of public order, directing the city traffic, and the prevention of crime. The Security Directorate acts to discover and prevent subversion, to patrol the frontiers, and to seek out illegal political organizations.

Wages and the cost of living. Compared with most other western European countries, the level of wages and salaries is very low in Portugal. This has created an outflow of more than 50,000 emigrants a year (not including several thousand illegal emigrants) to higher income countries—such as France and the Federal Republic of Germany. Wages in Portugal are governed by collective labour contracts in the various industries and have generally risen steadily, although women are still paid less than men. Wage increase has been moderately faster than increase in the price of consumer goods.

Medical services. The hospital network in Portugal includes both state and private institutions. Portugal has ten principal hospitals, nine of which are designated Civil Hospitals; Lisbon has seven and Porto two. The remaining one is designated a University Hospital—in Coimbra. Regional hospitals are located in the main provincial centres and subregional ones in smaller centres. Charity hospitals (the Santas Casas de Misericórdia) give free treatment to the poor. The Civil Hospitals, which are mainly run by the local municipalities with state support and funds from welfare and social organizations, charge according to the means of the patient. There are dispensaries and clinics for mothers and infants all over the country. Special institutions include a cancer hospital and research unit in Lisbon, a school of tropical medicine, a sanatorium in the northern mountains, and a modern rehabilitation service for the disabled near Lisbon. Ancient scourges such as leprosy have been almost stamped out; but a lepers' colony still exists, and a mobile squad of health officers attends cases that do not need isolation. A recent government scheme has attracted considerable sums for building new hospitals and improving old ones.

General social and economic features. The movement from an agrarian to an industrial society in Portugal has meant the erosion of the former feudal class and the consolidation of a middle class comprising not only industrialists and businessmen, professional men and women, and white-collar workers but also technicians and skilled labourers. The new impetus given to industry has not only caused the migration of workers from the land to the urban centres but the bringing of industry to the countryside. Industry, which has been concentrated largely around Lisbon and the country's second city, Porto, has come under a new plan for decentralization. As a result factories are springing up in rural areas, thus increasing the local population and raising the standard of living. A new property-owning class is being created by the pay-as-you-rent house-purchase system in all areas. The emigrant and his family are forming a new social class and consumer group, since remittances sent home by workers abroad contribute substantially to raising the standard of living of a large sector of the once impoverished rural population.

CULTURAL LIFE AND INSTITUTIONS

The cultural background. Portuguese culture is based on a past that dates from prehistoric times into the eras of Roman and Moorish invasions. All have left their traces in a rich patrimony of archeological remains, including prehistoric cave paintings at Escural, the Roman township of Conimbriga, the temple of Diana in Évora, and the typical Moorish architecture of such southern towns as Olhão and Tavira. Throughout the centuries Portugal's arts have been enriched by foreign influences, including Flemish, French, and Italian. The voyages of the Portuguese discoverers opened the country to oriental influences, and the revelation of Brazil's wealth of gold and diamonds fed the Baroque flame in decoration. Buildings are repaired and conserved as part of the national heritage, and museums show the finest examples of the country's arts and crafts. Classic art is fostered in educational establishments, and young artists receive state and other aid as they experiment with modern forms.

Artistic influences of the past

Architecture. Romanic and Gothic influences have given the country some of its greatest cathedrals, and in the late 16th century a national style (Arte Manuelina) was synthesized by adapting several forms into a luxuriant ornamental whole of complicated decoration. The Baroque later dominated decorative themes until French Neoclassicism swept into fashion with its reversion to Gothic imitations. These in turn have yielded to the stark realism of the modern concrete age. Outstanding examples of Portuguese architecture include Jerónimos Monastery in Lisbon, an elaborate example of ornate Manuelin style; Lisbon's Sé (cathedral), which still shows the remains of Romanic construction in part of the façade; the new Palace of Justice in Lisbon, a fine soaring example of austere modern architecture; the Convent of Christ in Tomar, formerly the seat of the Knights Templars; the late-Portuguese-Gothic monastery of Santa Maria at Batalha; Porto's granite Rococo Clérigos tower; and Braga's Romanic cathedral.

Sculpture. After pre-Roman and Roman times, sculpture found rich expression in the magnificent tombs of the 12th and 13th centuries. The Renaissance and Baroque periods also left traces leading to the lyrical period of the 19th and 20th centuries. Outstanding are the examples of late-18th-century Baroque sculptures in wood, of which the Christmas cribs of Machado de Castro are the finest.

Painting. A school of primitive painters headed by Nuno Gonçalves was prominent in the 15th century, and subsequently a wave of Flemish artists influenced the native style, decorating palaces and convents and leaving a rich heritage of religious art. The 19th century saw another rebirth of national art with a late romantic period. A movement of naturalist realism that followed gave way to the experimental 20th century: Vieira da Silva is the country's finest modern abstract painter, and Carlos Boteho is noted for his portraits of Lisbon and street scenes.

Music. Portuguese music reached a peak of national achievement during the Renaissance. Its modern revival was primarily the work of Luis de Freitas Branco, whose Neoclassic tradition has been perpetuated by Joly Braga Santos. A new group of young composers is acquiring prestige both at home and in international competitions. There is a very active National Broadcasting Symphony Orchestra, and Porto has had its own symphony orchestra since 1962, the year in which the Chamber Orchestra was set up by the Gulbenkian Foundation, of the late Portuguese oil magnate Calouste Gulbenkian. In 1971 the Lisbon Municipal Orchestra was created, replacing the defunct Lisbon Philharmonic Orchestra. The official workers' recreation group, the Federação Nacional de Alegria no Trabalho (FNAT), created the Portuguese Operatic Company, which encourages national singers, producers, and conductors. Folk music and the traditional *fado*, a song form that expresses the sad romantic mood of the nation, remain as the country's fundamental forms of musical expression.

Literature. The Portuguese language became synthesized in the 12th century, and a lyrical quality was outstanding both in poetry and prose. A soldier-poet, Luiz Vaz de Camões, gave expression to the nation's epic genius in the 15th century, as the poet Fernando Pessoa expressed its decadence in the 19th century. Lyrical poetry still flourishes. The tendency of fiction, however, has been to deviate from the romantic form of the 19th and early 20th centuries and take a more realistic form. José Maria Eça de Queiróz was the outstanding realist novelist at the end of the 19th century; his works include *Os Maias* (1888; "The Mayans") and *A Cidade e As Serras* (1901; "The City and the Mountains"). In the 20th century the late Aquelino de Ribeiro was the outstanding regionalist novelist whose works include *Jardin das Tormentas* (1913; "Garden of Torments") and *O Homem que Matou o Diabo* (1930; "The Man Who Killed the Devil"), while Ferreira de Castro is the outstanding modern realist novelist with *A Selva* (1930; *Jungle*, 1934) and *Os Emigrantes* (1928; "The Emigrants").

Theatre and cinema. After a long period of stagnation the theatre is resuming a more active role. The film industry is also recovering after years of depression with aid from the Gulbenkian Foundation. While the state has for years promoted a ballet company whose works are based on Portuguese regional folk dancing and music, the Gulbenkian Foundation has now formed a ballet company with a repertoire in the classical and modern tradition. International opera, theatre, and ballet are prominent in regular seasons in the Lisbon São Carlos Opera House and other theatres.

Folk traditions and popular culture. The folk traditions still survive, particularly in the northern countryside where the old songs and dances grace local religious festivals. The government is also doing much to promote and preserve them both as a national tradition and a tourist attraction. Some of the best examples of regional dances are the *vira*, *chula*, *corridinho*, *tirana*, and fandango. National dress has been generally abandoned for everyday use, but it is still worn in the northern Minho province for marriage and other festivals. Traditional garments such as the green-and-red stocking cap of the Alentejo cattleman with his *samarra* (a short jacket with a fox fur collar) and *cifões* (chaps) survive. In the northern Trásos-Montes e Alto Douro province, the shepherd wears a straw cloak.

Regional fairs abound throughout the year, many of them combining with religious festivals. Christmas and other religious customs include the burning of the yule log in the porch (atrium) of the country church, the fight between Saint George and the Dragon in the streets of Monção at Corpus Christi. A folk museum in Lisbon features the furniture, tools, and customs of the various provinces of Portugal, including the pottery, regional embroideries, and other crafts that are a living part of rural life.

Cultural institutions. *Academies and societies.* Among Lisbon's chief institutions are the Geographical Society with its library, museum of overseas exhibits, and lecture halls; the Lisbon Academy of Sciences, the Portuguese Academy of History, the National Academy of Fine Arts, the Music Academy of Santa Cecília, and the Musical Circle. There were also organizations concentrating on literary activities, including the Society of Writers and Composers, which had been disbanded by government order on a charge of subversion but which has now been reformed. Nonpolitical popular organizations, such as the Voice of the Worker (Voz do Operário), give classes and lectures to workers, the youth movement (Mocidade Portuguesa) organizes cultural classes and entertainment for the young, and official students' associations run extra-curricular activities for university students.

Museums, libraries, and art galleries. Portugal is rich in museums, libraries, and art galleries. In Lisbon, besides the Gulbenkian Foundation which houses Calouste Gulbenkian's collections, the most outstanding are the Museum of Ancient Art, the Coach Museum, which contains Europe's finest collection of ancient vehicles, the Museum of Contemporary Art, and the Ricardo Espirito Santo School of Decorative Arts, which trains craftsmen in



Terraced vineyards along a tributary of the Rio Douro in the province of Trás-os-Montes e Alto Douro, Portugal.
Yan—Rapho Guillumette

ancient handicrafts. The Biblioteca Nacional (National Library) and Palácio Nacional de Ajuda (National Palace of Ajuda) have fine libraries, while outside Lisbon the libraries of the Convento de Mafra and the University of Coimbra have historical importance.

Press and broadcasting. A new law was passed in 1971 that gave greater freedom of action to the press and restricted official censorship to news concerning the war against the African liberation movements and other matters pertaining to national defense when deemed necessary.

Lisbon has five morning and four evening newspapers. Other main newspapers are published in most of the large Portuguese towns and in Madeira and the Azores. Periodicals include a glossy review of international news, women's magazines, technical, ecclesiastical, and juvenile publications.

Broadcasting facilities consist of both a state network and private stations. The state Emissora Nacional is maintained by license fees paid by listeners. It possesses a network of 47 medium and shortwave transmitting stations. The most important privately owned stations are Radio Clube Português, Radio Renascença, and Emissores Associados de Lisboa, Lda.

Radiotelevisão Portuguesa holds the concession for operating television in Portugal. Regular programs were introduced in 1957, and in the early 1970s the station was on the air for some 11 hours a day on two channels. Colour television is to be introduced on the system's second channel.

THE OUTLOOK

Portugal is in a phase of relatively significant political, economic, social and cultural change. When Marcello Caetano succeeded Antonio Salazar as prime minister in 1968, he directed government activities toward freeing the country from its tangled bureaucracy, easing economic restrictions, reforming industry, and passing laws for relaxing censorship and for religious freedom. A complete scheme of educational reform was introduced, and intellectual life in Portugal began to experience liberalization and change. Coordinated efforts by the authorities to preserve the country's historic and artistic heritage and the ancient skills of the craftsman continued, but at the same time a definite stirring was felt in modern artistic endeavour. This awakening partly stemmed from the easing of restraints such as censorship and partly from the increasing contact with European cultural movements. Modern technology continued to change the style of life more and more rapidly in both town and countryside. While emigration and the war against the African liberation movements in the overseas territories drained manpower from labour, women were increasingly absorbed into the professions, business, and industry. This in itself was a radical social change in a country where middle and upper class women were still traditionally kept at home. Increased purchasing power at home and revenue from emigrants abroad have fostered a higher standard of living, particularly among people living in the rural areas.

Experimentation in all art forms, a modern adaptation of architectural concepts to meet the exigencies of modern living, and the spread of popular education through the press, radio, and television are all playing their part in producing a nation aiming to achieve the social and economic progress that will enable it to participate equally in the European community. (J.Sh.)

BIBLIOGRAPHY. HERMANN LAUTENSACH, *Bibliografia Geográfica de Portugal* (1948), a useful bibliography; ARISTIDES AMORIM GIRAO, *Atlas de Portugal*, 2nd ed. (1958), the standard atlas; *Guia de Portugal*, 6 vol., a detailed regional guide by the Gulbenkian Foundation, in course of publication.

General works: PIERRE BIROT, *Le Portugal: étude de géographie régionale* (1950), an excellent concise source; ARISTIDES AMORIM GIRAO, *Geografía de Portugal*, 15 pt. (1941–43), an interesting, if somewhat dated, handbook; HERMANN LAUTENSACH, *Portugal auf Grund eigener Reisen und der Literatur*, 2 vol. (1932–37), the first major survey of Portugal; ORLANDO RIBEIRO, *Portugal*, vol. 5 of *Geografía de España y Portugal* (1955), the best general survey; *Portugal, o Mediterrâneo e o Atlântico*, 3rd ed. (1967), a synthesizing treatment emphasizing the originality of Portugal; DAN STANISLAWSKI, *The Individuality of Portugal: A Study in Historical-Political Geography* (1959).

Regional studies: Five books published for the XVIth International Congress of Geography, give good regional coverage: JORGE DIAS, *Minho, Trás-os-Montes, Haut-Douro* (1949); MARIANO FEIO, *Le Bas Alentejo et l'Argarve* (1949); A.F. MARTINS, *Le Centre littoral et le Massif Calcaire de l'Estremadure* (1949); VIRGINIA RAU and GEORGES ZBYSZEWSKI, *Estremadura et Ribatejo* (1949); and ORLANDO RIBEIRO, *Le Portugal Central* (1949). Other interesting regional studies are: G. FERRO, *L'Algarve, Monografia regionale* (1956); DAN STANISLAWSKI, *Portugal's Other Kingdom: The Algarve* (1963); *Landscapes of Bacchus: The Vine in Portugal* (1970). The periodical *Finisterra* (issued twice yearly), is also a very good source of regional information.

Annual surveys and serials: Several Public Services of the country maintain annual surveys and serials, such as Direcção-Geral dos Serviços Hidráulicos (hydraulics), Instituto Nacional de Estatística (statistics and demography), Serviços Geológicos de Portugal (geological surveys and mines), Serviço Meteorológico Nacional (climate), and Serviço de Reconhecimento e de Ordenamento Agrário (soils and vegetation). These contain a wealth of useful information.

Special topics: EUGENIO DE CASTRO CALDAS, *Modernização da Agricultura Portuguesa* (1960); A. HUBERT, *Le Portugal et son économie* (1961); J.L. PINTO MACHADO, *Alguns Problemas do Mundo Rural Português* (1965); VALENTIM X. PINTADO, *The Structure and Growth of the Portuguese Economy* (1964); ORLANDO RIBEIRO, *A Evolução Agrária no Portugal Mediterrâneo* (1970); *Alguns Elementos Sobre a Economia Portuguesa* (1963); *III° Plano de Fomento para 1968–1973*, 4 vol. (1968); *Portugal*, OECD Annual Economic Surveys; official English translation of the 1971 revised text of the Portuguese constitution, published by the Secretariat of State for Information and Tourism, Lisbon; "Any More Questions?" (1970), an informative booklet on Portuguese institutions and culture, also published by the Secretariat of State for Information and Tourism; RODNEY GALLOP, *A Book of Folkways* (1961), intensive research on old customs and usage; ANN BRIDGE and SUSAN LOWNDES, *The Selective Traveller in Portugal*, rev. ed. (1967), an intelligent guide to the architecture, art, and culture of Portugal.

(I.M.P. doA./J.Sh.)

Portugal, History of

Portugal became an independent monarchy in the 12th century AD. Its name derives from Portus Cale, a pre-Roman or Roman settlement near the mouth of the Douro River. The southern part of the Roman province of Gallaecia was occupied by a Germanic people, the Suebi, in

AD 411, and Portucale (Porto) was held by them. They were subdued by the Visigoths, whose state was overthrown by the Muslim invasions in the 8th century. A Christian territory of Portugal was constituted in 868; it later became a county and was extended to Coimbra. Afonso Henriques assumed the title of king (1139) and annexed Lisbon. His successors took Alentejo and Algarve from the Muslims (by 1252). Portugal now includes the "adjacent islands" of Madeira and the Azores, first settled in the 15th century.

PRE-ROMAN, ROMAN, GERMANIC, AND MUSLIM PERIODS

The earliest human remains found in Portugal are Neanderthal-type bones from Furninhas. Most peninsular Paleolithic industries are represented, but a distinct culture first emerges in the Mesolithic middens of the lower Tagus Valley, dated *c.* 5500 BC. Neolithic cultures entered from Andalusia. There are varied types of beehive huts and passage graves: agriculture, pottery, and the working of soft metals followed by the same route. In the first millennium BC, Celtic peoples entered the peninsula by the Pyrenees, and many groups were projected westward by natural pressure. Hallstatt cultures brought ironworking to the Tagus Valley. Phoenician and later Carthaginian influence reached southern Portugal in the same period. By 500 BC, Iron Age cultures predominated in the north. Celtic hilltop settlements (*castros, citânias*) retained their vitality after the Roman conquest.

Celtic tribes

After the Second Punic War (218–201 BC), Rome dominated the eastern and southern seaboards of the peninsula, and Celtic peoples who had partially absorbed the indigenous population occupied the west. A Celtic federation, the Lusitani, resisted Roman penetration under the brilliant leadership of Viriathus; but after his assassination (*c.* 140 BC), Decius Junius Brutus was able to march northward through central Portugal, cross the Douro River, and subdue the Gallaeci. Julius Caesar governed the territory for a time. In 25 BC Augustus founded Augusta Emerita (Mérida) as capital of Lusitania, which incorporated the present central Portugal. Gallaecia, to the north of the Douro, became a separate province under the Antonines. In Roman times the region of Beja and Évora was a wheat belt. The valley of the Tagus was famous for its horses and farms, and there were important mines in Alentejo. Notable Roman remains include the "Temple of Diana" at Évora and the site of Conimbriga (Condeixa). Christianity reached Lusitania in the 3rd century and in the 4th Gallaecia, where the heretical and ascetic teaching of Priscillian long held sway.

Suebi, Visigoths, and Muslims

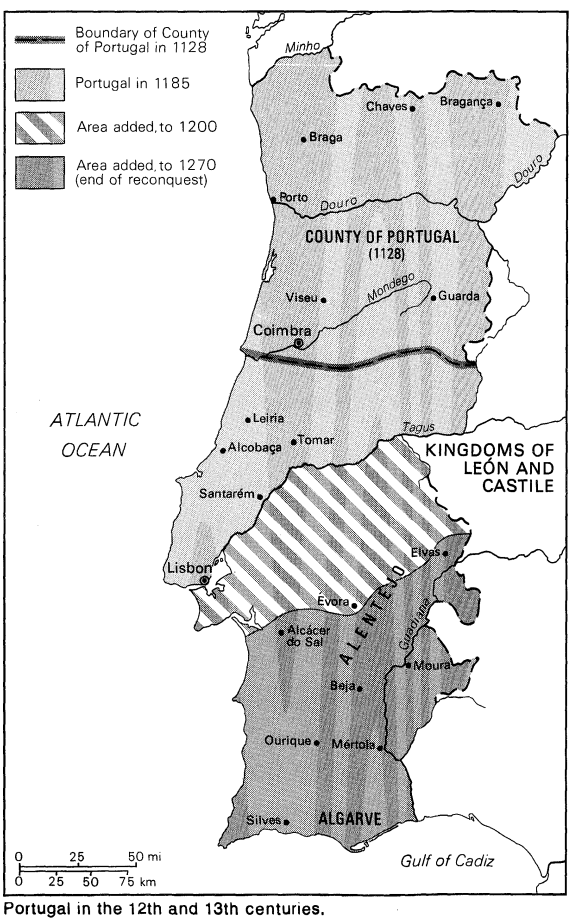
With the collapse of the Rhine frontier (AD 406), barbarian peoples forced their way into Gaul and crossed the Pyrenees. A Germanic tribe, the Suebi, was settled in southern Gallaecia, their rulers residing at or near Bracara Augusta (Braga) and Portucale. They annexed Lusitania and for a time overran the rest of the peninsula, but the Visigoths were sent to subdue them and extinguished their monarchy (AD 469). The records are silent until *c.* 550, when the Suebic monarchy had been restored and was reconverted to Catholicism by St. Martin of Braga. The Visigoth Leovigild again overthrew the Suebic monarchy and annexed the territory, but it remained distinct from Gothic Spain. St. Martin's church grouped together the bishoprics of the Suebic territory until *c.* 660, when the ecclesiastical divisions were adjusted to the old Roman provincial system and the region to the south of the Douro was restored to Lusitania. With the Muslim invasion of 711, the only serious Gothic resistance was made at Mérida; on its fall the northwest submitted. Berber troops were placed in central Portugal and Galicia, but with the revolt of the Berbers and famine (740–750) the latter was evacuated. Braga was abandoned, but the rural population remained or was restored. When ʿAbd ar-Raḥmān I set up the Umayyad monarchy at Córdoba (756), there was some resistance in the west, and he perhaps stationed Berber allies at Mérida and Coimbra. Lisbon was independent for a few years (*c.* 805). The restoration of the Christian sees of Galicia, the discovery of the supposed tomb of St. James, and the erection of his shrine at Santiago de Compostela were followed by the organization of the frontier territory of Portucale (868) by Vimara Peres; Coimbra was annexed by the Christians but later lost again.

THE COUNTY AND KINGDOM OF PORTUGAL (TO 1383)

By the 10th century the county of Portugal (north of the Douro) was held by Mumadona Dias and her husband Hermenegildo Gonçalves and their descendants, one of whom was tutor and father-in-law to the Leonese ruler, Alfonso V. But when this dynasty was overthrown by the Navarrese-Castilian house of Sancho III the Great, the western county lost its autonomy. Sancho's son, Ferdinand I of Castile, reconquered Coimbra in 1064 but entrusted it to a Mozarabic governor. When the African Almoravids annexed Muslim Spain, Alfonso VI of León (1065–1109) and Castile (1072–1109) provided for the defense of the west by calling on Henry, brother of Duke Eudes (Odo) of Burgundy, whom he married to his illegitimate daughter Teresa and made count of Portugal. From 1095, therefore, Henry and Teresa (who used the title of queen) ruled Portugal and Coimbra. On the death of Alfonso VI, his realms passed to his legitimate daughter Urraca and her little son Alfonso (VII). Henry of Portugal sought power but had achieved little when he died in 1112, leaving Teresa with an infant son, Afonso Henriques. Teresa's intrigues with her Galician favourite, Fernando Peres of Trava, lost her the support of the Portuguese barons, and in 1128 followers of her son defeated and drove her into exile. Thus Afonso Henriques became count of Portugal. Though at first obliged to submit to his cousin Alfonso VII, he began to use the title of king, traditionally because of a victory over the Muslims at Ourique (July 25, 1139; but perhaps a symbolic legend). In 1143 his cousin accepted his autonomy, but the title of king was formally conceded only in 1179, when Afonso Henriques placed Portugal under the direct protection of the Holy See, promising an annual tribute. He had captured Santarém (March 1147) and Lisbon (October 1147), the latter with the aid of English, French,

Afonso Henriques

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London



Portugal in the 12th and 13th centuries.

German, and Flemish crusaders bound for Palestine. An English priest, Gilbert of Hastings, became first bishop of the restored see of Lisbon. Although the new Moroccan dynasty of the Almohads struck back (1179–84), when Afonso I died (Dec. 6, 1185) the Portuguese frontier was firmly established on the Tagus. The new military orders, the Templars, those of Calatrava (from *c.* 1156), and of Santiago (from *c.* 1170), etc., governed castles and territory on the frontier, and the Cistercians were responsible for the introduction of agriculture and architecture in central Portugal (Alcobaça).

The kingdom and the reconquest. Although Afonso Henriques began to grant charters to new settlements, it was his son Sancho I (1185–1211) who enfranchised many municipalities (*concelhos*), especially in eastern and central Portugal. The privileges of these communities were embodied in charters (*forais*), which attracted settlers from the more feudal north. Even Muslims were enfranchised, though on the other hand many of them were enslaved. Assisted by passing crusaders, Sancho captured Silves in Algarve (1189); but in 1190 an al-Muwāhhidūn army from Africa advanced to the Tagus and, although Lisbon, Santarém, and Tomar stood firm, the Muslims recovered Silves in 1191, together with most of the land south of the Tagus. In his later years Sancho was involved in a quarrel with Pope Innocent III over the payment of tribute due to the Holy See and with the bishop of Porto, who received Innocent's support. But peace was made before his death, and it was left to his son Afonso II, the Fat (1211–23), to endeavour to strengthen the power of the throne at the expense of the church.

Though Afonso II was an unwarlike king, his followers were beside the Castilians at the great Christian victory of Las Navas de Tolosa in 1212 and, again assisted by crusaders, recovered Alcácer do Sal in 1217. Meanwhile Afonso repudiated the bequests of large estates made by his father to his brothers and accepted those to his sisters only after a war with León and in a form, settled by the pope, that recognized Afonso's sovereignty. In the first year of his reign, Afonso called a meeting of the Cortes at Coimbra, to which the nobility and prelates were summoned (representatives of the commoners were not to appear until 1254). Both estates obtained important concessions; in fact, the position of the church and the orders was now so strong that Afonso II and his successors were involved in recurrent conflicts with Rome. Afonso himself instituted (from 1220) *inquirições*, or royal commissions, to investigate the nature of holdings and recover whatever had been illegally taken from the crown. In his last years, Afonso quarrelled with the archbishop of Braga, who was supported by Pope Honorius III, defied the papacy, and was excommunicated.

The Cortes

Little is known of the reign of his son Sancho II (1223–*c.* 1246), but the reconquest of Alentejo was now completed and much of Algarve was reduced. On his accession, Sancho found the church in full ascendancy as a result of the agreement made before his father's death. Conflicting reports exist of Sancho's own government, but in his later years the kingdom seems to have slipped into anarchy. At all events, his younger brother Afonso, who had become count of Boulogne by his marriage (1238) with Matilde, daughter of Count Raynald I of Dammartin, was granted a papal commission (1245) to take over the government, and Sancho was ordered to be deposed by papal bull. When Afonso reached Lisbon (late 1245 or early 1246), he received the support of the church and of the inhabitants of Lisbon and other towns. After a civil war lasting two years, Sancho II retired to Toledo, dying there in January 1248.

On his arrival, the count of Boulogne had already declared himself king as Afonso III, and the death of Sancho without issue gave his usurpation the mantle of legality. He brought together the divided kingdom, completed the reconquest of Algarve, transferred the capital from Coimbra to Lisbon, and, fortified by the support of the towns, summoned the Cortes at Leiria at which for the first time commoners representing the municipalities made their appearance (1254). His conquest of Algarve aroused the jealousy of Castile. Two campaigns were fought (1250 and 1252), Afonso probably being worsted, and peace was made by means of a marriage pact. Although still the husband of Matilde of Boulogne, Afonso married Beatriz, illegitimate daughter of Alfonso X of Castile, holding the disputed territory of Algarve as a fief of Castile until such time as the eldest son of the marriage should reach the age of seven, when Algarve was to return to Portugal. This marriage led to a dispute with the Holy See, in which Afonso was placed under an interdict. Despite his early connection with Rome, Afonso refused to give way; and in 1263 the bigamous marriage was legalized and his eldest son, Dinis, legitimized. Shortly afterward, Afonso launched *inquirições*, as a result of which the church was deprived of much property. The prelates protested these actions of the royal commissions, and most of them subsequently left the country. Although Afonso was excommunicated and threatened with deposition, he continued to defy the church until shortly before his death early in 1279.

First representation of commons

The achievements of Afonso's reign—the completion of the reconquest, the assertion of the royal power before the church, and the incorporation of the commoners in the Cortes—indicate important institutional advances. Under his son Dinis (1279–1325), Portugal was to come into closer touch with western Europe and to acquire a university, the elements of a national literature, and a navy. The chartering of fairs and the increased use of minted money bear witness to the growth of commerce, and the planting of pine forests to hold back the sand dunes near Leiria illustrates Dinis' concern for shipbuilding and agriculture. In 1317 Dinis engaged a Genoese admiral, Emmanuele Pessagno, to build up his navy, having already adopted various measures to stimulate foreign trade. He founded the University of Coimbra (at first in Lisbon) in 1290 and was both a poet and patron of literature. Yet he was especially famed as the "farmer king" (*rei lavradór*) for his interest in the land.

Despite Dinis' attachment to the arts of peace, Portugal was several times involved in strife during his reign. In 1297 the Treaty of Alcañices with Castile confirmed Portugal's possession of Algarve and provided for an alliance between Portugal and Castile. In the later years of Dinis' reign, his son, the future Afonso IV, rebelled more than once, being persuaded to submit by the influence of his mother Isabel, daughter of Peter III of Aragon. This remarkable woman, later canonized as St. Elizabeth of Portugal and popularly known as the Rainha-Santa (the Holy Queen), successfully exercised her influence in favour of peace on several occasions.

St. Elizabeth of Portugal

Disputes with Castile. Afonso IV (1325–57) was also involved in various disputes with Castile. Isabel, who had retired to the Convent of Santa Clara at Coimbra, continued to intervene in favour of peace; but on her death in 1336 war broke out, and terms were not made till 1340, when Afonso himself with a Portuguese army joined Alfonso XI of Castile in the great victory over the Muslims on the Salado River in Andalusia. Afonso's son Pedro was married (1336) to Constança (died 1345), daughter of the Castilian infante Juan Manuel; but soon after the marriage he fell in love with one of her ladies, Inês de Castro, by whom he had several children. Afonso was persuaded to countenance the assassination of Inês in 1355, and one of Pedro I's earliest acts on his accession was to take vengeance on her murderers. During his short reign (1357–67), Pedro devoted himself to the dispensation of justice; his judgments, which he executed himself, were severe and often violent, and his iron rule was tempered only by fits of revelling.

Pedro's son by Constança, Ferdinand (1367–83), inherited a wealthy throne almost free of external entanglements; but the dispute between Pedro the Cruel and Henry of Trastamara (later Henry II) for the Castilian throne was raging, and on the murder (1369) of the former, several Castilian towns offered Ferdinand their allegiance, which he was unwise enough to accept. Henry II duly invaded Portugal in 1369, and by the Peace of Alcoutim (1371) Ferdinand was constrained to renounce his claim and to promise to marry Henry's daughter; however, he instead took a Portuguese, Leonor Teles, although she was already

married and in spite of the protests of the commoners of Lisbon. He also made alliance (1372) with the English in the person of John of Gaunt, duke of Lancaster, who had married the elder daughter of Pedro the Cruel and claimed the Castilian throne. In 1372 Ferdinand provoked Henry II, who invaded Portugal and besieged Lisbon. Unable to resist, Ferdinand was forced to repudiate his alliance with John of Gaunt and to act as an ally of Castile, surrendering various castles and persons as hostages. It was only on the death (1379) of Henry that Ferdinand dared openly to challenge Castile again. In 1380 the English connection was resumed, and in the following year John of Gaunt's brother, Edmund of Langley, earl of Cambridge (afterward duke of York), took a force to Portugal for the invasion of Castile and betrothed his son Edward to Ferdinand's only legitimate child, Beatriz. In mid-campaign, Ferdinand came to terms with the enemy (August 1382), agreeing to marry Beatriz to a Castilian prince. She did, in effect, become the wife of John I of Castile; and when Ferdinand died, prematurely decrepit, Leonor Teles became regent and Castile claimed the Portuguese crown.

Alliance with the English

Leonor had long been the paramour of the Galician João Fernandes Andeiro, count of Ourém, who had intrigued with both England and Castile and whose influence was much resented by Portuguese patriots. Opponents of Castile chose as their leader an illegitimate son of Pedro I, John, master of Aviz, who killed Ourém (December 1383) and, being assured of the support of the populace of Lisbon, assumed the title of defender of the realm. The regent fled to Alenquer and thence to Santarém, and the king of Castile came to her aid, soon however, relegating her to a Spanish convent. Lisbon was besieged for five months (1384), but an outbreak of plague obliged the Castilians to retire.

THE HOUSE OF AVIZ, 1383–1580

The legitimate male line of Henry of Burgundy ended at Ferdinand's death, and when the Cortes met at Coimbra (March–April 1385), John was declared king and became the founder of a new dynasty. This result was not unopposed, for many of the nobility and clergy still considered the queen of Castile the rightful heiress; but popular feeling was strong, and John I had valuable and trusted allies in Nuno Álvares Pereira, "the Holy Constable," his military champion, and in João das Regras, his chancellor and jurist.

Independence assured. A number of towns and castles still held out for Castile when in August 1385 John I of Castile and a considerable army made their appearance in central Portugal. Although much outnumbered, the Portuguese won the great Battle of Aljubarrota (Aug. 14, 1385), in which the Castilian chivalry was dispersed and John of Castile himself barely escaped. The victory, followed by secondary successes won by Nuno Álvares, assured John I of his kingdom and made him a desirable ally. A small force of English archers had been present at Aljubarrota. Now the Treaty of Windsor (May 9, 1386) raised the Anglo-Portuguese connection to the status of a firm, binding, and permanent alliance between the two crowns. John of Gaunt duly went to the Peninsula (July 1386) and attempted an invasion of Castile in conjunction with John I. This was not successful, but the Portuguese king married Gaunt's daughter Philippa of Lancaster (1387), who introduced various English usages into Portugal. The truce arranged with Castile in 1387 was prolonged at intervals until peace was finally concluded in 1411.

The victory of John of Aviz may be regarded as a victory of the national spirit against the feudal attachment to established order. As much of the older nobility had adhered to Castile, John rewarded his followers at their and at the crown's expense. Meanwhile, commerce prospered, and the marriage of John's daughter Isabella to Philip the Good of Burgundy was to be followed by the growth of close trading relations between Portugal and Philip's County of Flanders. With the conclusion of peace with Castile, John found an outlet for the activities of his frontiersmen and of his own sons in the conquest of Ceuta (1415), from which may be dated the great age of Portuguese expansion.

Trade with Flanders

During the short reign of John's eldest son, Edward (Duarte; 1433–38), an unsuccessful attempt was made to conquer Tangier in 1437 by John's third son, Prince Henry the Navigator, and his younger brother Ferdinand, who was captured by the Moors and died, still unransomed, in 1443. On Edward's death, his son Afonso V was still a child, and his brother Pedro, duke of Coimbra, had himself made regent (1440) instead of the widow, Leonor of Aragon. But Pedro's own regency was later challenged by the powerful Bragança family, descended from Afonso, illegitimate son of John of Aviz and Beatriz, daughter of Nuno Álvares Pereira. This family continued to set the young king against his uncle, who was forced to resign the regency, driven to take up arms, and killed at Alfarrobeira (May 1449). Afonso V (1438–81) proved unable to resist the demands of the Braganças, who now became the wealthiest family in Portugal. Having married Joan, daughter of Henry IV of Castile, Afonso laid claim to the Castilian throne and became involved in a lengthy struggle with Ferdinand and Isabella in the region of Zamora and Toro, where he was defeated in 1476. He then sailed to France to entreat the aid of Louis XI, in which he failed, and on his return concluded with Castile the Treaty of Alcáçovas (1479), abandoning the claims of his wife Joan. Afonso never recovered from his reverse, and during his last years his son John administered the kingdom.

Consolidation of the monarchy. John II (1481–95) was as cautious, firm, and jealous of the royal power as his father had been openhanded and negligent. At the first Cortes of his reign, he exacted a detailed oath of homage that displeased his greatest vassals. A suspicion of conspiracy enabled him to arrest Fernando II, duke of Bragança, and many of his followers; the duke was sentenced to death and executed (1484) at Évora, and John himself stabbed James (Diogo), duke of Viseu (1484). As well as attacking the power of the nobility, John lessened the effects of the unfavourable treaty with Castile. Calculating and resolute, he later received the epithet "the Perfect Prince."

John II was predeceased by his legitimate son and therefore succeeded by his cousin, the duke of Beja, as Manuel I (1495–1521), known as "the Fortunate." Manuel, who assumed the title of "lord of the conquest, navigation, and commerce of India, Ethiopia, Arabia, and Persia," inherited, because of the work of John II, a firmly established autocratic monarchy and a rapidly expanding overseas empire. Drawn toward Spain by the common need to defend their overseas interests as defined by the Treaty of Tordesillas (1494), Manuel nourished the hope of joining the whole peninsula under the house of Aviz; he married Isabella, eldest daughter of Ferdinand and Isabella, who, however, died (1498) in giving birth to a son, Miguel da Paz. This child was recognized as heir to Portugal, Castile, and Aragon but died in infancy. Manuel then married Isabella's sister Maria (died 1517) and thirdly Eleanor, sister of the emperor Charles V.

As a condition of the marriage with Isabella, Manuel was required to "purify" Portugal of Jews. John II had admitted many Jewish refugees expelled from Spain (1492) and had taxed them heavily but was also to supply ships for them to leave Portugal. This was not done, and Manuel now ordered all Jews to leave within ten months, by October 1497. On their assembly in Lisbon, every effort was made to secure their conversion by promises or by force. Some who resisted were allowed to go, but the rest were "converted" under promise that no inquiry should be made into their beliefs for 20 years. As "Christians" they could not be forced to emigrate, and they were indeed prohibited from leaving Portugal. In April 1506 a large number of these "new Christians," or *marranos*, were massacred in Lisbon during a riot, but Manuel afterward protected them and allowed many to emigrate to Holland, where their experience of Portuguese trade was put at the service of the Dutch.

Forced conversion of Jews

If Manuel failed to realize his dream of ruling Spain, his son John III (1521–57) lacked the power to resist Castili-

an influence. A pious, retiring man, he was ruled by his wife, Catherine, sister of the emperor Charles V, and encouraged the installation of the Inquisition (1536); the first auto-da-fé was held in 1540. The Society of Jesus, established in 1540, soon controlled education in Portugal. In 1529 the settlement by the Treaty of Saragossa (Zaragoza) of a dispute over the possession of the Moluccas removed an obstacle to Portuguese–Spanish understanding, and the line of Tordesillas was matched by a similar line in the Pacific; all the territories of the New World were in theory divided between Spain and Portugal, while the Reformation (as well as the discoveries) had come between the latter and its English ally.

John III was succeeded by his grandson Sebastian (1557–78), then a child of three. As a boy, Sebastian became obsessed with the idea of a crusade against Morocco. Fanatically religious, he had no doubts of his own powers and listened only to flatterers. He visited Ceuta and Tangier in 1574 and began in 1576 to prepare a large expedition against Larache: this departed in June 1578 and on August 4 was utterly destroyed by the Moors in the Battle of the Three Kings near Alcazarquivir (Ksar el-Kebir). Sebastian himself was killed, 8,000 of his men are said to have died, and 15,000 were captured. Only a handful escaped.

The disaster of Larache

Sebastian was succeeded by his great-uncle, Cardinal Henry (1578–80), a brother of John III. His age and celibacy made it certain that the Portuguese throne would soon pass from the direct line of Aviz. Philip II of Spain, nephew of John III and husband (by his first marriage) of John's daughter Maria, had already made his preparations and, on the death of the cardinal-king (Jan. 31, 1580), summoned the authorities to obey him. An army under the great duke of Alba entered Portugal in 1580; the resistance of António the prior of Crato (illegitimate son of John III's brother Luís), acclaimed António I at Santarém, collapsed; and Philip II of Spain became Philip I of Portugal (1580–98). The claims of the duchess of Bragança (Catherine, daughter of John's brother Edward) were disregarded.

MEDIEVAL SOCIAL AND ECONOMIC DEVELOPMENT

Medieval Portugal comprised regions of considerable diversity. In the north, the old aristocracy of Leonese descent owned large estates worked mainly by serfs. In the southern territory that had been reconquered from Muslim rule, there were many towns, often separated by districts almost barren and depopulated. The initiative in settling these areas was taken by Cistercian monks, who had reached Portugal by 1143; later kings such as Sancho I and Afonso III established *concelhos* (municipalities), granting to them by the issue of *forais* (charters) many privileges designed to attract population. Tax concessions were often given, and freedom was promised to serfs or to Christian captives after a year's residence. But in the south, the *concelhos* were burdened with defense duties, the *cavaleiros-vilãos* (villein knights) being obliged to horse and arm themselves, and the *peões*, or less substantial men, being obliged to serve as foot soldiers in defense of the country and perhaps also on a *fossado* (raid) into Muslim territory.

Court officials

At court, the king was advised by his *curia regis* (court or council), comprising the *majordomus curiae*, the head of the administration, the military chief or *signifer*, the *dapifer curiae* (steward of the household), the chancellor (an official whose origins in Portugal were Burgundian rather than Visigothic), and any members of the greater aristocracy, the *ricos-homens*, who might be at court. This class also comprised the bishops and abbots and masters of the orders of knighthood; many held private civil or military authority. The lesser nobility were without such rights. Below them came various classes of free commoners, such as the *cavaleiros-vilãos* and the *malados*, men who had commended themselves to protectors. There were numerous serfs and slaves.

By the end of the medieval period several shifts in the social structure had occurred. Many of the old aristocracy lost their position at the advent of the house of Aviz, and the new nobility, exemplified in the house of Bragança, was often of bureaucratic or ministerial origin. The privileges granted to *concelhos* forced landlords to compete for labour, and by the time of Afonso III money payments had largely replaced labour services on the estates. Representatives of the commoners, first attending the Cortes in 1254 on behalf of the *concelhos*, took an increasing part in politics. The Cortes were very frequently called in the reigns of John I, Duarte, and Afonso V; but the avenues of power had become wider by the 16th century, and John III's proposal (1525) to call them only every ten years aroused no opposition. Although the trade guilds were slow in developing, they took some part in determining local taxation in the 13th century. Trade increased, Portuguese merchants having connections with the Low Countries from the time of Afonso Henriques and with England from the early 13th century. The political crisis of 1385 was followed by intense inflation and many debasements: there was thereafter no national gold currency until 1435, when West African sources began to be tapped.

THE DISCOVERIES AND THE EMPIRE

The idea of expansion into Africa was a logical result of the completion of the reconquest in the peninsula, and the conquest of Ceuta in North Africa (1415) probably provided the impulse toward further expansion. The simple idea of fighting the Muslims on their own soil was linked with more complicated motives: the desire to explore in a scientific sense, the hope of finding a way to the rich spice trade of the Indies, and the impulse to spread the Christian faith. These purposes were gradually molded together into a national enterprise, though at first they represented the hopes and aspirations of one man, Prince Henry. The third son of John I and Philippa of Lancaster, known rather inaccurately as "the Navigator" (he himself never went farther afield than Tangier), Henry became (1420) master of the Order of Christ, which King Dinis had founded (1319). The resources of the Order were used to draw together skilled geographers and navigators and to equip a series of expeditions that only gradually began to bear fruit.

Henry the Navigator

The date of Prince Henry's earliest expedition is not exactly known but appears to have been about 1418, when the island of Porto Santo was visited; the first call at Madeira probably dates from 1419. An attempt was made to settle in the Canaries, and between 1427 and 1431 the Azores were visited by Portuguese seamen. Both the Azores and Madeira were then uninhabited, and their colonization proceeded fairly rapidly from *c.* 1445. Sugar was exported to Europe and gave the islands great economic importance. Meanwhile, Prince Henry's ships were probing the African coast, passing Cape Bojador in 1434 and Rio de Oro in 1436. The unsuccessful expedition against Tangier (1437) was followed by a break in the discoveries, but in 1439 Prince Henry was authorized to colonize the Azores; and from 1440 further expeditions equipped with a new and lighter ship, the caravel, reached the Bay of Arguin (1443), Cape Verde (1444), and by Henry's death (1460) had explored the coast as far south as Sierra Leone.

Under Afonso V, three military expeditions were sent against Morocco (1458, 1463, and 1471); by the last of them, Tangier and Arzila were captured. The African explorations were not entirely neglected, but it remained for John II, with his sharp sense of the national interest, to found a fortress and trading post in the Gulf of Guinea at Elmina (São Jorge da Mina, 1481–82). Diogo Cam (Cão) discovered the mouth of the Congo in 1482 and then advanced to Cape Cross, 200 leagues southward (1486). A native monarchy, the Manicongo, sought conversion and alliance with the Portuguese: its first Christian king, Afonso I (1505–40), made Mbanza a centre of Portuguese influence, but the "kingdom of Congo" fell into internal strife and Portuguese interests were transferred to the neighbouring kingdom of Angola. Paulo Dias de Novais founded Luanda, the first city in West Africa south of the Equator, in 1576. In 1488, Bartolomeu Dias at length rounded the Cape of Good Hope and reached the East African coast, and the seaway to India lay open.

Bartolomeu Dias rounds Cape of Good Hope

His return was followed in 1493 with the news that Christopher Columbus had, as he thought, discovered the "Indies" by sailing across the Atlantic. Much as this news must have perturbed the Portuguese, Columbus brought no news of the spiceries or the cities of the East; and John II ordered the preparation of an expedition to India by way of the Cape of Good Hope, though this sailed only after his death. John also contested the Spanish claim to all lands discovered west of the Atlantic, and by the Treaty of Tordesillas (June 7, 1494), Spain's rights were limited to what lay more than 370 leagues west of the Cape Verde Islands. Thus the territory that was to become Brazil was reserved for Portugal.

The Treaty of Tordesillas had confirmed Portugal's rights to the exploration of Africa and the seaway to India. In July 1497 Vasco da Gama set sail with four ships on the first expedition to India. They reached Calicut (Kozhikode) the following spring, and the survivors put into Lisbon in the autumn of 1499 with specimens of Oriental merchandise (see also INDIAN SUBCONTINENT, HISTORY OF THE). A second fleet was prepared under Pedro Álvares Cabral, who touched the Brazilian coast (April 22, 1500) and claimed it for Portugal. One of Cabral's ships, under Diogo Dias, discovered Madagascar in 1500; João da Nova discovered Ascension in the following year and St. Helena in 1502. Tristão da Cunha sighted the island named after him in 1506 and went on to explore Madagascar. Meanwhile, trading posts in India had been established by Cabral at Cochin and Calicut (1501) and by João da Nova at Cannanore. In 1502 Vasco da Gama made tributary to Portugal the ruler of Kilwa (Quiloa) in East Africa.

Settlements in the Far East

In 1505 Francisco de Almeida arrived as viceroy of India, strengthening the African station at Kilwa and supporting the ruler of Cochin against the *zamorin* of Calicut. The control of sea trade now instituted became the chief source of Portuguese wealth in the East. It was assured by the defeat of Muslim naval forces off Diu in 1509. Almeida's successor, Afonso de Albuquerque, conquered Goa (1510), which he made the seat of Portuguese power, and Malacca (1511); sent two expeditions to the Moluccas (1512 and 1514); and captured Hormuz in the Persian Gulf (1515). Soon after, Fernão Peres de Andrade reached Canton in China; in 1542 Portuguese merchants were permitted to settle at Liampo (Ning-po), and in 1557 they founded the colony of Macao (Macau).

Albuquerque was responsible for this conception of a system of strong points that secured the trade of the Orient to Portugal for nearly a century. Goa soon became the chief port of western India; Hormuz controlled the Persian Gulf, and Malacca the gateway from the Indian Ocean to the South China Sea, while a string of fortified trading posts secured the coast of East Africa, the gulf and the shores of India and Ceylon (see CEYLON, HISTORY OF: *Ceylon under foreign rulers, 1505–1947*). Farther east, less fortified settlements were set up with the consent of the native rulers from Bengal to China, and the trade of the principal spice islands was in Portuguese hands. The preservation of the whole system was entrusted to a governor, who sometimes held the rank of viceroy, at Goa; and although Portuguese arms had both triumphs and reverses, their control of the Oriental trade remained substantial, if never complete, until the 17th century, when the Dutch, at war with the joint crown of Portugal and Spain and deprived of their traditional trade with Lisbon, began to seek spices from their source and effectively demolished the Portuguese monopoly.

UNION OF SPAIN AND PORTUGAL, 1580–1640

After Philip II of Spain had occupied Portugal in 1580, the island of Terceira in the Azores held out for António of Crato, who himself sought alliances in England and France. In 1582 a French expedition to establish him in the Azores was defeated, and in 1589 an English attempt upon Lisbon, led by Sir Francis Drake and Sir John Norris, failed dismally. But although António died in Paris in 1595, the true symbol of Portuguese independence was not the prior of Crato but King Sebastian himself. The Portuguese people refused to believe that he was dead and nourished a messianic faith in his reappearance, of which four pretenders sought to avail themselves, the last as late as 1600 and as far afield as Venice.

Meanwhile, Philip arrived in Portugal and was accepted as King Philip I (1580–98) by the Cortes held at Tomar (1581). He undertook to preserve Portuguese autonomy, to consider the union as a personal one like that of Aragon and Castile under Ferdinand and Isabella, to appoint only Portuguese to the administration, to summon Cortes frequently, and to be accompanied by a Portuguese council in Madrid. These undertakings were, however, neglected by Philip II (III of Spain, 1598–1621) and completely violated by Philip III (IV of Spain, 1621–40).

Exploitation by the Spanish

Portuguese resentment against Spanish rule was increased by the failure of these kings to visit Portugal, the appointment of Spaniards to Portuguese offices, the loss of trade consequent on Spain's foreign wars, and the levying of taxation to sustain these wars. In 1624 the Dutch seized Bahia in Brazil, only to be expelled by a joint Spanish and Portuguese expedition (1625). But in 1630 they occupied Pernambuco and the adjoining sugar estates, which they held for a generation. The final straw was the conde-duque de Olivares' plan (1640) of using Portuguese troops against the equally discontented Catalans. Two Portuguese insurrections, in 1634 and 1637, had failed to attain dangerous proportions, but in 1640 Spain's powers were extended to the utmost by war with France and revolt in Catalonia. The French minister, Cardinal de Richelieu, already had agents in Lisbon, and a leader was found in John, duke of Bragança, a grandson of the duchess Catherine (niece of John III) whose claims had been overridden in 1580 by Philip II of Spain. Taking advantage of the unpopularity of the governor, Margaret of Savoy, duchess of Mantua, and her secretary of state Miguel de Vasconcelos, the leaders of the party of independence carried through a nationalist revolution on Dec. 1, 1640. Vasconcelos was almost the only victim; the Spanish garrisons were driven out; and on December 15 the duke of Bragança was crowned as John IV (1640–56).

THE HOUSE OF BRAGANCA, 1640–1910

Although the Cortes confirmed the accession of the dynasty of Bragança and John's coronation on Jan. 28, 1641, the success of the new regime was not finally assured till 1668, when Spain at last recognized Portuguese independence. Faced with the threat of a Spanish invasion, John had sent missions to the courts of Europe in quest of alliances. France now refused a formal treaty. The Dutch, having seized northern Brazil, accepted a truce in Europe and proceeded to capture Angola from Portugal. John made a treaty (1642) with Charles I of England, which was made void by Charles's execution (1649). Meanwhile, the Portuguese defeated the Spaniards at Montijo (May 26, 1644) and warded off several invasions. In 1654 they negotiated a treaty with the English Commonwealth, obtaining aid in return for commercial concessions. The Dutch were now finally expelled from Pernambuco in northern Brazil. By a secret article of the Treaty of the Pyrenees (1659), France promised Spain to give no further aid to Portugal, but in 1661 Portugal signed a treaty of alliance with the restored English monarchy. In 1662 Charles II of England married John's daughter Catherine of Bragança and, in return for a large dowry, including the cession of Bombay and Tangier, provided arms and men for the war with Spain. Portuguese defense was organized by the German soldier Friedrich Hermann von Schönberg (later duke of Schomberg); in June 1663 Sancho Manuel, count of Vila Flor, defeated Don John of Austria at Ameixial, and in June 1665 von Schönberg won the important victory of Montes Claros. Peace was finally made by the Treaty of Lisbon early in 1668.

Treaties with England

On the death of John IV, his second son Afonso VI (1656–83) was 13 years of age. Afonso's mother, Luísa de Gusmão, acted as regent until, in June 1662, he began to rule. Afonso himself was feebleminded, but the country was capably governed by Luiz de Vasconcelos e Sousa, conde de Castelo Melhor, until 1667. The French prin-

cess, Maria Francesca of Savoy, who married Afonso in the previous year, now entered into an intrigue with his more personable brother Pedro (afterward Pedro II). They contrived to dismiss Castelo Melhor and to have Maria Francesca's marriage annulled. She at once married Pedro (1668), who was declared regent. Afonso was imprisoned until his death. During the reign of Pedro II (1683–1706), Portugal recovered from the strain of the Spanish wars and began to benefit from the discovery of gold and precious stones in Brazil. The first strike of gold in Minas Gerais took place in 1693, and in the last years of the 17th century, considerable wealth was being extracted; however, it was not until 1728 that diamonds were discovered: the mineral wealth of Brazil then formed a very substantial part of the revenue of the Portuguese crown.

Mineral wealth from Brazil

The 18th century. The War of the Spanish Succession (1701–13) saw Portugal's recent friends, England and France, on opposing sides; and although Pedro sought at first to remain neutral, Portugal joined the Anglo-Austrian Grand Alliance in 1703, by which it afforded a base for the archduke Charles (later the emperor Charles VI) to conduct his war for the Spanish throne. In the same year (December 27) the English envoy, John Methuen, also concluded the treaty that bears his name, by which the exchange of port wine for English woollens became the basis for Anglo-Portuguese trade. Although the treaty of 1654 had secured great privileges for English merchants in Lisbon, neither it nor the treaties of 1642 and 1661, by which the traditional alliance was restored, had created trade. This was now done, and by reason of the wealth that soon poured into Lisbon from Brazil, the English merchants gained a commanding position in the trade of Portugal. The political treaties of 1703 proved less fruitful. The Portuguese general António Luís de Sousa, marquess of Minas, entered Madrid in 1706; but French and Spanish forces were victorious at Almansa (1707), and in 1711 the French admiral René Duguay-Trouin sacked Rio de Janeiro. At the conclusion of the war Portugal negotiated a peace treaty with France (April 1713) but did not conclude peace with Spain until 1715.

Under Pedro's son John V (1706–50), Portugal attained a degree of prosperity unknown since the restoration. The tax of a royal fifth levied on the precious metals and stones of Brazil gave the monarchy an independent source of wealth. The Cortes, which had met irregularly since 1640, was no longer summoned, and government was carried out by ministers appointed by the king. John V desired the absolute authority enjoyed by Louis XIV. He did not fail to convert his wealth into papal and other dignities: the archbishop of Lisbon became a patriarch (1716), and Pope Benedict XIV in 1749 gave John the title "his most faithful majesty"; and royal academies, palaces, and libraries were inaugurated. But in his later years his ministers proved inadequate and the kingdom sank into stagnation. On John's death, his son Joseph (1750–77) appointed as minister Sebastião José de Carvalho e Melo, later conde de Oeiras and marquês de Pombal, who soon gained a complete ascendancy over the king and endeavoured to replace the stagnant absolutism with a more active type of despotism that, with some qualifications, deserves the epithet enlightened. His full powers date from his efficient handling of the crisis caused by the disastrous Lisbon earthquake of November 1755; but even before this he had reformed the sugar and diamond trades, set up a national silk industry (1750), and formed one chartered company to control the sardine and tunny-fishing industry of Algarve and another to trade with northern Brazil. In 1756 he founded the Junta do Comércio, a board of trade with powers to limit the privileges enjoyed by the English merchants under the treaties of 1654 and 1661, and set up the General Company for Wines of Alto Douro (Companhia dos Vinhos do Alto Douro) to control the port wine trade. Industries for the manufacture of hats (1759), cutlery (1764), and other articles were set up with varying success.

Enlightened despotism of Pombal

Pombal's methods were arbitrary and his enemies numerous. His reform of the wine industry provoked a riot in Porto (1757) that was savagely repressed; but his principal victims were the Jesuits, expelled in 1759 from all the Portuguese dominions, and the nobility, in particular José Mascarenhas, duque de Aveiro, and the Távora family, who were accused of an attack on the King (Sept. 3, 1758), condemned, and executed (Jan. 12, 1759). Having eliminated the Jesuits from the educational system, Pombal applied regalist principles in the reform of the University of Coimbra (1772) and the royal board of censorship (1768), which supervised the system of lower education from 1771.

Expulsion of Jesuits

While Pombal succeeded in modifying the ascendancy of the British merchants in Portugal, he invoked the English alliance in 1762 when Spain, prompted by the renewal of the Bourbon Family Compact alliance with France, invaded Portugal. The Portuguese Army was reformed by Wilhelm von Schaumburg-Lippe, and an English force was led by James O'Hara, 2nd Baron Tyrawley, and John Campbell, 4th earl of Loudoun. Peace was signed in February 1763 at Fontainebleau. On the death of Joseph (Feb. 24, 1777), his daughter Maria I (1777–1816), who had married his brother, her uncle (Pedro III), acceded; Pombal was dismissed (1777) and eventually found guilty on several charges. His successors made peace with Spain by the Treaty of San Ildefonso (1777).

The French Revolutionary and Napoleonic Wars. Maria I suffered from melancholia after the loss of her consort (1786) and eldest son John (1788). In 1792 her mental balance was further disturbed, probably by the news of the French Revolution, and she ceased to reign. Her son, who on her death became John VI (1816–26), then ruled in her name and in 1799 became prince regent. In 1793 Portugal joined England and Spain against France, sending a naval division to assist the English Mediterranean fleet and an army to the Catalan front. The Peace of Basel (July 1795), by which Spain abandoned its allies, left Portugal still at war. Although subjected to pressure from the French Directory and from the Spanish minister, Manuel de Godoy, Portugal remained unmolested until 1801, when Godoy sent an ultimatum and invaded Alentejo. By the Peace of Badajoz (June 1801), Portugal lost the town of Olivenza and paid an indemnity.

From the Peace of Amiens (1802) until 1807, Portugal was once more immune from attack, though subjected to continuous pressure to break off the English connection. By the Berlin decree of Nov. 21, 1806, Napoleon sought to close all continental ports to British ships. Portugal endeavoured to maintain neutrality, but the secret Franco-Spanish Treaty of Fontainebleau (October 1807) provided for its eventual dismemberment by Napoleon I and Godoy. Already Gen. Andoche Junot was hastening across Spain with a French army, and on November 27, the prince regent and the royal family and court embarked on a fleet lying in the Tagus River and were escorted by British vessels to Brazil; the court remained 14 years at Rio de Janeiro. Junot declared the Braganças deposed, but his occupation of Portugal was challenged in August 1808 by the arrival of Sir Arthur Wellesley (later duke of Wellington) and 13,500 British troops in Mondego Bay. Winning the victories of Roliça (August 17) and Vimeiro (August 21), Wellesley enabled his superiors to negotiate the Convention of Sintra (August 31), by which Junot was allowed to evacuate Portugal with his army. A second French invasion (1808–09) led to Sir John Moore's death at La Coruña (January 1809) and the reembarkation of the British forces. In February, William Carr (later Viscount) Beresford was placed in command of the Portuguese Army, and in March Marshal N.J. de Dieu Soult advanced from Galicia and occupied Porto. Wellesley returned to Portugal in April, drove Soult from the north, and after his victory of Talavera de la Reina in Spain (July), withdrew to Portugal. The third French invasion followed in August 1810 when Marshal André Masséna with Marshal Michel Ney and Junot entered Beira province. Defeated by Wellington at Bussaco (September 27) near Coimbra, the French found themselves facing the prepared lines of Torres Vedras, north of Lisbon, where they wintered amid great privations. By the spring of 1811 they could only retreat and, on March 5,

Flight of royal family to Brazil

Lines of Torres Vedras

began the evacuation of Portugal, harassed all the way by English and Portuguese attacks and crossing the frontier after a defeat at Sabugal (April 3).

Portugal made peace with France on May 30, 1814. It was represented at the Congress of Vienna but played little part in the settlement. The series of Anglo-Portuguese treaties concluded between the years 1809 and 1817, however, was important insofar as it extended many of the conditions of the Anglo-Portuguese alliance to Brazil and had an influence on the future of Africa. England's efforts to get Portuguese collaboration in suppressing the slave trade resulted in the treaty of Jan. 22, 1815, and in the additional convention of 1817, by reason of which Portugal's claims to a considerable part of Africa were formally recognized.

Constitutionalism. The Napoleonic campaigns had caused great devastation in Portugal, and the absence of the royal family and the presence of a foreign commander (Beresford) combined with revolutionary agitation and the influence of Spanish liberalism to produce an atmosphere of discontent and restlessness. On Dec. 16, 1815, Brazil was raised to the rank of a kingdom united with Portugal, and John VI, who succeeded in March 1816, showed no desire to return to Portugal. In 1817 Beresford suppressed a conspiracy in Lisbon, and the Masonic leader Gen. Gomes Freire de Andrade was executed. Unrest increased, and when Beresford himself went to Brazil (March 1820) to press John to return, a constitutionalist revolution began in Porto (Aug. 24, 1820), spread over the country, and led to the formation of a junta in Lisbon (October 4). On Beresford's return (October 10) he was not allowed to land, and British officers were expelled from the army. A constituent assembly was summoned that drew up a very liberal constitution, thus confronting John VI with an accomplished fact.

John's reluctance to return was at last overcome, and he left his elder son Pedro to govern Brazil, landing at Lisbon on July 3, 1821. He swore to uphold the constitution, but his wife, Carlota Joaquina, and their second son, Dom Miguel, refused to take the oath and were sentenced to banishment, though this was not carried out. The Portuguese constitutionalists, not appreciating the determination of Brazil not to yield up its status as a kingdom, sought to compel Pedro to return; but he, rather than sacrifice the rule of the Braganças in Brazil, declared for Brazilian independence (Sept. 7, 1822) and became emperor of Brazil as Pedro I (see also BRAZIL, HISTORY OF). This enabled his brother Miguel to appeal to absolutist forces in Portugal to overthrow the constitutionalists, and an insurrection led by Miguel almost succeeded (April 30, 1824); but through the action of the foreign ministers, John VI was restored and Miguel went into exile in Vienna (June 1824).

Independence of Brazil

War of the Two Brothers. John VI acknowledged the independence of Brazil in 1825, assuming *pro forma* the imperial title and then yielding it to Pedro; but when John died (March 10, 1826), no provision had been made for the succession except that his daughter Maria Isabel was now named regent. Pedro, as Pedro IV of Portugal, issued from Brazil a charter providing for a parliamentary regime by the authorization of the monarchy and not based on the sovereignty of the people. He then made a conditional abdication (May 1826) of the Portuguese throne in favour of his daughter Maria da Glória, age seven, provided that she should marry her uncle Miguel and that he should swear to accept the charter. This compromise could not be effective. The absolutists had hoped that Pedro would resign all rights to the Portuguese crown, and the council of regency hesitated to publish the charter until Gen. João Carlos de Saldanha (later duque de Saldanha) forced their hand. In 1827 Miguel took the oath and was appointed regent, landing in Lisbon in February 1828. His supporters at once began to persecute the liberals. A form of Cortes met in Lisbon, and in July 1828 repudiated Pedro's claims and declared Miguel rightful king.

Only the island of Terceira in the Azores sustained the liberal cause. In June 1829, however, a regency on behalf of Maria da Glória was set up in Terceira; and in 1831 Pedro, having abdicated the Brazilian throne, went to Europe and began to raise money and men for the conquest of Portugal. In February 1832 the expedition sailed to Terceira, and in July the liberals, led by Pedro, disembarked at Mindelo near Porto, which city they soon occupied. The rest of the country, however, stood by Miguel, who besieged the liberals in Porto for a year (July 1832–July 1833). By now, Miguelite enthusiasm had waned; and António José de Sousa Manuel, duque de Terceira, and Capt. (later Sir) Charles Napier, who had taken command of the liberal navy, made a successful landing in Algarve (June 1833). Terceira advanced on Lisbon, which fell in July 1833, and Miguel capitulated at Evora-Monte in May 1834.

Further political strife. The War of the Two Brothers ended with the exile of Miguel (June) and the death of Pedro (Sept. 24, 1834). Maria da Glória became queen as Maria II (1834–53) at the age of 15. While Maria necessarily came under the influence of the successful generals of the civil war, her principal aim was to defend her father's charter (which had been granted by the crown) from those who demanded a "democratic" constitution like that of 1822, asserting the ultimate sovereignty of the nation.

In September 1836 the latter, thenceforth called Septembrists, seized power. The chartist leaders rebelled and were exiled, but by 1842 the Septembrist front was no longer united, and António Bernardo da Costa Cabral restored the charter. In 1846 the movement of Maria da Fonte, a popular rising against higher taxation to improve roads and reforms in public health in which almost all parties joined, put an end to Costa Cabral's government but left Portugal divided between the Septembrists, who held Porto, and Saldanha, now in Queen Maria's confidence, in Lisbon. Saldanha negotiated for the intervention of other members of the Quadruple Alliance (formed in April 1834 by England, France, Spain, and Portugal), and a combined British and Spanish force received the surrender of the Porto junta in June 1847 and ended the war with the Convention of Gramido (June 29, 1847). Saldanha governed until 1849, when Costa Cabral resumed office only to be overthrown in April 1851. Saldanha then held office for five years (1851–56), and the period of peace at length permitted the country to settle down. The "Regeneration" ended civil strife and established party government.

Septembrists and chartists

Maria II was succeeded by her eldest son by her second husband, Ferdinand of Saxe-Coburg, Pedro V (1853–61), who married Stephanie of Hohenzollern-Sigmaringen in 1858. He gave promise of being a capable and conscientious monarch but died of typhoid fever on Nov. 11, 1861. His brother Luís I (1861–89) seemed to have inherited a country that had recovered from the Napoleonic invasions and from civil wars, political strife, and *pronunciamentos*. But although the main parties were now defined as Historicals (*i.e.*, radicals) and Regenerators (or moderates), the alternation of governments gradually ceased to reflect any semblance of popular feeling, and in the last years of Luís' reign republicanism was already beginning to gain ground.

Colonial policies. With the accession of Carlos I (1889–1908), a serious dispute with Great Britain occurred. Portugal's possessions in Africa had been recognized by Great Britain in the treaty of 1815, but more recently Germany and Belgium had entered the colonial field; and at the Conference of Berlin (1885), the definition of "effective occupation" was adopted as the basis for possession of colonial territories. A colonial movement had gained momentum in Lisbon, and a Portuguese scheme, known as the "Rose-Coloured Map," which laid claim to a colony stretching across Africa from Angola to Mozambique, was recognized by France and Germany (1886). Although the Marquess of Salisbury registered a protest (1888), the Portuguese foreign minister Henrique de Barros Gomes sent Maj. Alexandre Alberto da Rocha de Serpa Pinto to the Shire Highlands (modern Malawi), with a view to their annexation. He became involved in a fight with the Makololos, who were under British protection, and a series of communications between London and

The "Rose-Coloured Map"

Lisbon ended in the dispatch of the British ultimatum of Jan. 11, 1890, demanding the withdrawal of all Portuguese from the Shire. Amid great popular excitement, Barros Gomes had no alternative but to comply, and the government resigned. The incident caused the deepest resentment in Portugal, not only against the ancient ally but also against the monarchy, which was menaced by an attempted republican revolution in Porto (Jan. 31, 1891).

During the following years, the Portuguese African colonies were defined as a result of the Anglo-Portuguese treaty of July 1891, but the financial position of the country was so bad that it seemed unlikely that the efforts to consolidate the African colonies would succeed. In 1897 it became clear that Portugal would require a considerable loan, and Germany demanded to partake in any assistance that was offered. On Aug. 30, 1898, A.J. (later the earl of) Balfour, temporarily in charge of the British Foreign Office, concluded a secret Anglo-German convention assigning spheres of influence in the Portuguese colonies to Great Britain and Germany in the event of such a loan. This was, however, denounced by the prime minister, Lord Salisbury, and in 1899 when the Germans endeavoured to persuade the Portuguese to accept a loan, Salisbury's action and the imminent danger of a conflict in the Transvaal caused an Anglo-Portuguese approximation. On Oct. 14, 1899, the ancient treaties of alliance were reaffirmed in a secret declaration, later made public (the so-called Windsor Treaty).

Meanwhile, the financial situation showed little improvement, and the republicans continued to progress. In 1906 João Franco, formerly a Regenerator, came to power as champion of the failing monarchist cause. Unable to obtain the support of the other monarchists, he began to govern by decree. Although Franco bravely undertook to reform the finances and administration, he was accused of illegally advancing money to the King. These scandals were followed by rumours of plots, and on Feb. 1, 1908, Carlos and his heir, Luís Filipe, were assassinated as they rode in an open carriage in Lisbon. Whether or not the regicides were isolated fanatics or agents of a hidden organization such as the Carbonária, a republican secret society, the deed was applauded by the republicans, who immediately began their preparations for a final attack on the monarchy.

Assassination of Carlos I

King Manuel II (1908–10) found no unity among the monarchist politicians. The general election of August 1910 showed republican majorities in Lisbon and Porto, and on October 3 the murder by a madman of one of the republican leaders, the distinguished physician Miguel Bombarda, offered the pretext for a rising that was already organized. Armed civilians, soldiers, and the men aboard some ships in the Tagus began the republican revolution on October 4 and, after faltering, their movement, in which António Machado Santos played the predominant part, succeeded on October 5. Manuel escaped to Ericeira and thence by sea to Gibraltar and to England. On his death in 1932 his body was returned to Portugal.

Social and economic conditions. From the time of the Portuguese overseas conquests, the flow of wealth from the new territories and trading posts, if it sustained the court and capital, did little to improve the domestic economy, still largely rural. The favourable financial position of the late 15th century, derived from trade in slaves, gold, and spices, did not long survive into the 16th century, when the expenses of maintaining far-flung and unproductive foreign stations and the depredations of pirates quickly absorbed any surpluses. There were few native industries. Not only were manufactured goods such as cloth, tapestry, and metalware imported but also basic foodstuffs, salt meat, cured fish, and dairy produce. Agriculture was little regarded, and insufficient land was available for small holdings. During the years of Spanish domination, the ports were closed to English merchants; and by the time they were reopened after 1640 the flow of trade had found new channels, and the Dutch and English had outstripped Portugal as colonial powers. The discovery of gold in Brazil at the end of the 17th century brought about a revival in the country's economy, but gold production was in decline by 1750, while the diamond market was saturated. The later 18th century saw a series of protectionist measures, many introduced by Pombal. The Methuen Treaty with England (1703) had strengthened the port wine trade at the expense of Portuguese cloth; later further attempts were made to improve the export value of port wine. Support was also given for the production of woollen goods, linen, paper, porcelain, and cutlery and to the tunny and sardine fisheries. Pombal attempted to create an educated bourgeoisie, but the Portuguese textile industry could not withstand mechanized competition. After 1850 public works, railways, and ports were given priority, but Portugal had to wait until the 20th century for any sustained attack upon its economic difficulties.

Economic difficulties

THE REPUBLIC, FROM 1910

The new regime formed a provisional government under the presidency of Teófilo Braga, a well-known writer. This in turn issued a new electoral law giving the vote to adult Portuguese males and presided over the election of a constituent assembly, which opened on June 19, 1911. The constitution was passed by the assembly on August 20, and the provisional government surrendered its authority a few days later (August 24) to the new president, Manuel José de Arriaga.

Evolutionists, Unionists, and Democrats. Although a monarchist invasion was unsuccessfully attempted by Henrique de Paiva Couceiro in October 1911, the main danger to the new regime came from its internal divisions. For the moment, it was fairly united in denouncing monarchism and persecuting the church. The religious orders were expelled (Oct. 8, 1910) and their property confiscated. The teaching of religion in primary schools was abolished and the Roman Catholic Church disestablished. The conditions under which Catholics and monarchists were imprisoned attracted attention abroad, and it was only gradually that this legislation was modified. New universities were founded at Lisbon and Porto; but the task of destruction proved easier than that of construction, and before long the republicans were divided into Evolutionists (moderates), led by António José de Almeida; Unionists (centre party), led by Manuel de Brito Camacho; and Democrats (the left wing), led by Afonso Augusto da Costa. A number of prominent republicans had no specific party. The whirligig of republican political life offered little improvement on the monarchist regime, and in 1915 the army showed signs of restlessness. Gen. Pimenta de Castro formed a military government and permitted the monarchists to reorganize, but a Democratic revolution (May 14) led to his arrest and consignment to the Azores. President Arriaga resigned and was succeeded by Braga and then by Bernardino Machado (Aug. 5, 1915–Dec. 8, 1917). The Democratic regime, in which Costa was paramount, was ended by the revolution (December 1917) of Maj. Sidónio Pais, who established a "New Republic" of a right-wing tendency, supported at first by the Unionists. On their withdrawal (March 1918), a new National Republican Party gained control. Sidónio Pais' "presidentialist" regime was abruptly ended with his assassination on Dec. 14, 1918, when, after the provisional presidency of Adm. João do Canto e Castro, power passed gradually back to the Democrats.

Roman Catholic Church disestablished

Meanwhile, on the outbreak of World War I, Portugal had proclaimed its adhesion to the English alliance (Aug. 7, 1914) and on November 23 committed itself to military operations against Germany. On September 11 a first expedition left to reinforce the African colonies, and there was fighting in northern Mozambique, on the Tanganyika frontier, and in southern Angola, on the frontier of German South West Africa. In February 1916 Portugal seized German ships lying in Portuguese ports, and Germany declared war (March 9). A Portuguese expeditionary force went to the western front in 1917, under Gen. Fernando Tamagnini de Abreu; on April 9, 1918, they were under heavy German attack in the Battle of Lys. By the Treaty of Versailles (1919) Portugal received 0.75 percent of the indemnity payable by Germany and the

Fighting during World War I

Quionga (Kionga) area of Mozambique captured by Portuguese forces in East Africa.

Almeida completed his term of office as president (Oct. 5, 1919–Oct. 5, 1923), but ministries succeeded one another in rapid succession. In 1921 the founder of the republic, António Machado Santos, was among those murdered by enemies of Sidónio Pais; and although António Maria da Silva contrived to govern in the Democratic interest for a year and nine months, his fall in November 1923 was followed by a number of short-lived ministries. Revolutionary movements grew more frequent as the Democratic Party lost its cohesion, and there were signs in the army of impatience with the political turmoil. Although the Democrats obtained a clear majority in 1925 and Manuel Teixeira Gomes (1923–25) yielded the presidency to Bernardino Machado without incident, a military revolt occurred in Lisbon on Feb. 2, 1926. It was quelled, but on May 28 Commander José Mendes Cabeçadas Júnior and Gen. Manuel de Oliveira Gomes da Costa led a successful rebellion at Braga. Machado was deposed from the presidency, and a provisional government was formed.

Military revolt

The "New State." At first, Cabeçadas was head of the provisional government and Gomes da Costa minister of war; but the former was regarded as too close to the politicians, and Gomes da Costa unseated him. Within a few weeks he too was deposed and his place taken by his foreign minister, Gen. (later Marshal) António Óscar de Fragoso Carmona (July 9, 1926). Elected president of the republic in March 1928, he remained in office until his death in April 1951. After an attempt at revolution in February 1927, which resulted in considerable bloodshed, Carmona's government was not seriously interrupted. The program of the military regime was merely to restore order. To remedy the financial plight of the country, it had been proposed to borrow money from the League of Nations, but the conditions offered included supervision of the finances, which was regarded as offensive to national sovereignty. The loan was, therefore, rejected, and Carmona in 1928 called on António de Oliveira Salazar to take the Ministry of Finance.

Salazar, professor of economics at Coimbra University, fully controlled all expenditure and revenue and embarked on a complete overhaul of the administration. As minister of finance from 1928 to 1940, he produced an unbroken series of budgetary surpluses that restored Portugal's financial credit. As prime minister from 1932, he ushered in the new Constitution of 1933; as a minister of colonies in 1930, he prepared the Colonial Act governing the administration of Portugal's overseas provinces; and as minister for foreign affairs (1936–47), he guided Portugal through the difficulties caused by the Spanish Civil War and during World War II practiced a form of neutrality compatible with the Anglo-Portuguese Alliance. In May 1940 a concordat was signed with the Vatican, which clarified the position of the Catholic Church in Portugal. The church was reconfirmed in its possession of most of the property it had held before 1910, religious instruction was restored in state schools, private church schools were allowed, and church marriages were to be recognized. On the death of Carmona in 1951, Salazar temporarily assumed under the constitution the attributes of the presidency until Gen. Francisco Craveiro Lopes was sworn in as president in August 1951.

Constitution of 1933

By 1953 Salazar had been a member of the government for 25 years. His "New State," as defined by the Constitution of 1933, provided for a president elected for a period of seven years who appointed a prime minister and a variable number of ministers; a national assembly of deputies elected quadrennially as a block, meeting for at least three months annually; and a corporative chamber consisting of representatives of trades and professions acting as a consultative assembly. At general elections from 1934 onward, all seats in the National Assembly went to government (National Union Party) supporters, although in 1949, 1953, 1957, and 1969 there were a few opposition candidates. General Craveiro Lopes was succeeded (1958) as president by Rear Adm. Américo de Deus Rodrigues Tomás, elected by a large majority over Brig. Humberto Delgado, who shortly afterward went into exile. Meanwhile, the determination of the Indian government to annex Portuguese India led to India's severing diplomatic relations with Portugal (August 1955) and to mass invasions of the Portuguese possessions by Indian "passive resisters," who were repelled. In 1955 Portugal entered the United Nations, and in December 1955 referred the question of free access to the enclaves of Dādra and Nagar Haveli (Havili) to the International Court of Justice, which in April 1960 recognized Portugal's right of passage to the enclaves and India's control of the surrounding territory. On Dec. 18, 1961, India invaded Goa, Diu, and Damão (Daman); the subsequent debate in the UN was stifled by the U.S.S.R.'s veto.

Salazar had made it plain that "decolonization" was not a possible policy for him. Shortly afterward, northern Angola was invaded from the Congo, and a number of settlers and others were murdered. Attempts were later made to invade Guinea and Mozambique, involving a substantial Portuguese military commitment. In Portugal a small rising at Beja failed in January 1962.

The New State pressed on with electrification schemes (the most important being that of the Douro), industrialization (steelworks, assembly of motor vehicles, shipbuilding, etc.), and public works (the Salazar Bridge at Lisbon, the largest in Europe). Many new schools were built. Internal migration swelled the population of the cities, especially Lisbon, and the tide of emigration turned from the New World to Europe, particularly France.

In September 1968 Salazar (died July 1970) suffered a serious stroke, and the president, after consulting the Council of State, asked Marcelo José das Neves Alves Caetano, former minister and rector of Lisbon University, to form a government. Caetano, despite opposition from hard-line Salazarists in cabinet and National Assembly, cautiously introduced some liberalizing measures, including a relaxation of press censorship.

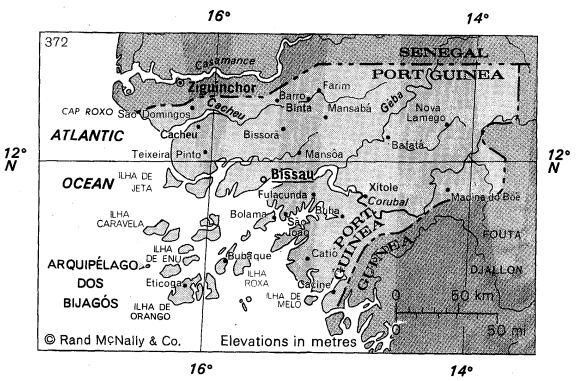
BIBLIOGRAPHY. Recent histories of Portugal in English are H.V. LIVERMORE, *A History of Portugal* (1947) and *A New History of Portugal* (1966 and 1969); and CHARLES E. NOWELL, *A History of Portugal* (1952). The standard histories in Portuguese are FORTUNATO DE ALMEIDA, *História de Portugal*, 6 vol. (1922–29); and DAMIAO PERES (ed.), *História de Portugal*, 7 vol. (1928–35). There is a convenient *Dicionário de História de Portugal*, ed. by J. SERRAO (1963). For prehistory, see H.N. SAVORY, *Spain and Portugal* (1968); and on Portuguese origins, H.V. LIVERMORE, *Origins of Spain and Portugal* (1971). OLIVEIRA MARQUES has published a guide to medieval Portugal entitled *A Sociedade medieval portuguesa* (1964); the classic account is still ALEXANDRE HERCULANO DE CARVALHO E. ARAUJO, *História de Portugal*, 4 vol. (1846–53; 7th ed., 8 vol., 1914–16). The best general account of the Portuguese expansion is E. PRESTAGE, *The Portuguese Pioneers* (1933), but much work has since been done on economic aspects: see especially V. DE MAGALHAES GODINHO, *L'Économie de l'empire portugais* (1969). DAMIAO PERES, *História dos Descobrimentos Portugueses* (1959; Eng. trans., *A History of the Portuguese Discoveries*, 1960), is a more recent treatment of Prestage's theme; and C.R. BOXER, *The Portuguese Seaborne Empire, 1415–1825* (1969), carries it forward until the separation of Brazil. F. MAURO, *Le Portugal et l'atlantique au XVII^e siècle, 1570–1670* (1960), deals with the overseas economy in the 17th century. On the Inquisition, see HERCULANO, *Da origem e estábelecimento da Inquisição em Portugal*, 3 vol. (1854–59; Eng. trans., *History of the Origin and Establishment of the Inquisition in Portugal*, 1 vol., 1926); and A.J. SARAIVA, *Inquisição e Cristãos-Novos*, 2nd ed. (1969). A.D. FRANCIS, *The Methuens and Portugal, 1691–1708* (1966), discusses the famous treaties. T.D. KENDRICK has examined the repercussions of *The Lisbon Earthquake* (1956). A. SILBERT, *Le Portugal méditerranéen à la fin de l'ancien régime, XVIII^e–début du XIX^e siècle*, 2 vol. (1966), is a substantial study of underlying agrarian conditions. An outline of the 19th century is given in R.J. HAMMOND, *Portugal and Africa* (1966). The Anglo-Portuguese crisis of 1890 is dealt with by P.B. WARHURST in *Anglo-Portuguese Relations in South-Central Africa, 1890–1900* (1962); and by E. AXELSON in *Portugal and the Scramble for Africa, 1875–1891* (1967). The story of the republic is told by the Spanish historian JESUS PABON, who does not like it, in *La Revolución portuguesa*, 2 vol. (1941–45). Works on Salazar include F. CLEMENT C. EGERTON, *Salazar, Rebuilder of Portugal* (1943); and MICHAEL DERRICK, *The Portugal of Salazar*

(1938), which deal with his earlier years; see also R. VON GERSDORFF, *Portugals Finanzen* (1961); and A. PASQUIER, *L'Économie du Portugal* (1961).

(H.V.L.)

Portuguese Guinea

The West African territory of Portuguese Guinea is roughly triangular in shape and is officially known as the Province of Portuguese Guinea. The United Nations, however, considers it to be a non-self-governing territory, and it is referred to as Guinea-Bissau by African nationalists. It is bounded by Senegal to the north, by the Republic of Guinea to the east and south, and by the Atlantic Ocean to the west. The Bijagós (Bissagos) Archipelago and other islands off the coast also form part of the territory, which is administered as a political unit from the capital, Bissau. The total area is 13,948 square miles (36,125 square kilometres), which includes low-lying ground submerged at high tide.



PORTUGUESE GUINEA

While at the time of the 1960 census the population numbered about 520,000, the present population total remains uncertain because of the guerrilla war, which has been waged against Portuguese rule by African nationalists since January 1963. According to the 1970 census the population numbers 487,000. The conflict in the territory has taken the form of a war of attrition and has played havoc with an already fragile economy. In the early 1970s, Portuguese Guinea was an embattled country with extremely little prospect of an early return to normal conditions. (For coverage of historical aspects, see WEST AFRICA, HISTORY OF.)

The landscape. *Relief.* Almost all of the country is low-lying, except in the northeast where some of the outlying ridges of the Fouta Djallon plateau rise to heights of between 300 and 600 feet. Along the coast there is an extremely dense network of rias (drowned valleys), which deeply indent a partially flooded coastal plain, bordered by swamps to seaward. In the interior the ground slowly rises to form an interior plain to the northwest, between the Rio Geba and the Senegal frontier, and low plateaus to the northeast and east. Crisscrossed by meandering rivers, these plains, which form part of the Sénégal River Basin, are flatlands with some minor relief features and outcrops of lateritic (reddish iron-bearing) rock. Farther to the east, a transitional plateau between the plains and the Fouta Djallon highlands of neighbouring Guinea is formed by the Planalto (plateau) de Bafatá in the middle of the country and the Planalto de Gabu in the northeast. Off the coast a string of low, muddy islands is hardly distinguishable from swampland at high tide. Among these are Caió, Pecixe, Bissau, Areias, Bolama, Como, and Melo. In previous centuries these islands were used by the Portuguese as stepping-stones to the interior; from 1890 to 1941 the capital was at Bolama on the island of the same name. Farther out to sea, the Bijagós Islands were for many years a lair of piratical Africans who could easily hide in the heavily forested archipelago in order to prey both upon the mainland and upon unwary navigators. The main islands of the Bijagós group are Bubaque, Rubane, Orangozinho, Roxa, Galinhas, Meio, Poilão, Uno, Orango, Uracane, Eguba, Carache, Caravela, Formosa, and Ponta. The principal rivers are navigable by flat-bottomed craft far into the interior but split the country into sections between which communication is difficult. From north to south the main rivers are the Cacheu, followed by the Mansôa, the Geba and the Corubal, the Rio Grande de Buba, and the Cacine.

The coastal islands

Climate, vegetation, and animal life. The climate is hot and wet; rainfall occurs between May and October, after which the dry season lasts from November to April. The coast has a monsoonal climate (characterized by seasonal moisture-bearing winds from the sea) with an abundant rainfall, which amounts to 40 to 80 inches a year in the north. The interior is influenced by the savanna (grassland) type of climate, which has greater variation in rainfall and temperature. Everywhere the temperature is high, May and April being the hottest months with temperatures of about 84° F (29° C), and January and December the coldest with temperatures of about 73° F (23° C). There are three vegetation zones. The first is formed by the coastal swamps and plains that are covered with mangrove and palm trees and in some places with rice fields. The second is formed by the interior plain, which is heavily forested. The third zone is a continuation of the savanna that occurs in neighbouring Senegal and Guinea. Portuguese Guinea has a great variety of aquatic birds, including the pelican and the flamingo. Crocodiles, snakes, buffalo, gazelles, apes, eagles, parrots, hyenas, and leopards abound, but lions are now rarely seen. Sharks are found in the estuaries.

Human settlement. There is a striking difference between the swampy, coastal part of the country to the west and the drier hinterland to the east. Patterns of settlement differ accordingly. While the islands and the coast are inhabited by palm oil and rice growers, the interior is used by pastoralists for grazing and is cultivated by peanut (groundnut) producers. The Balante, or Balanta, people, who occupy the area between the Rio Geba and the Casamance River in Senegal, are renowned as rice growers. Villages are almost nonexistent, each extended family being lodged in a hut complex where strongly built storerooms, cattle pens, and intricate palisaded corridors protect the clan from cattle thieves or marauders from other Balante families. Another coastal people, the Bram (Brame), are famous for their skill in farming without irrigation. The Bijagós people inhabiting the Bijagós Archipelago are more at home at sea than on land. Inland are the cattle-raising Fulani and the agricultural Malinke (Mandingo) peoples; both peoples are Islāmized and most of their villages have a mosque.

The Portuguese versus the nationalists

Upon these traditional patterns the war has superimposed a new form of settlement—the strategic hamlet, established by the Portuguese to regroup and protect Africans who have not been won over by the nationalists. Especially among the coastal Mandyako and Balante, the authorities and the nationalists are engaged in a contest for the allegiance of each tribe. In the nationalist-held territories, new regroupings have also taken place in the forests to escape air attacks from the Portuguese. Few sectors of the country remain unaffected either by insurgency or by governmental countermeasures. Many shifts of population also have taken place. As a result of the war, Bissau, the capital, has expanded to accommodate refugees from the interior and has also become an important garrison town. Its population in 1970 numbered 62,000; it is an administrative, commercial, and communications centre with many gardens and fine public buildings. The other city of note, Bolama, the former capital, is slowly decaying. Inland, Bafatá, a market town and military and administrative centre, is important. Other significant posts on the coastal plain are Catió and Teixeira Pinto. Farther east, Bissorã, Mansôa, and Nova Lamego (Gabú) are peanut-collecting points or garrison headquarters. Historically significant Farim and Cacheu on the Rio Cacheu have declined as a result of the war in progress.

People and population. In the past the main population movement was that of the Islāmized pastoralists of the east toward the animist peoples of the coast—a move-

ment partially checked by the Portuguese conquest. The main ethnic divisions among the Africans were last reported at the 1950 census, when they were as follows: Balante 160,000; Fulani (Foulah) 108,000; Mandyako (Mandjak) 72,000; Malinke (Mandingo) 64,000; Pepel (Papel) 36,000; Bram 16,000; Biafada 12,000; Bijagós 10,000; Felup 8,000; Bayot 4,000; Nalu 3,000; Soninke (Sarakole) 2,000; Susu 2,000. While no ethnic statistics were available after the 1960 census, one may consider that the Balante now constitute about 30 percent of the total, the Fulani 20 percent, the Mandyako 14 percent, the Malinke 12.5 percent, the Pepel 7 percent, and others 16.5 percent. In theory, each group originally had its tribal territory (known as *chão* in local parlance); but a Balante expansion, coupled with further movements by the inland people and war, have tended to make the tribal pattern complex if not confused.

The most expansion-prone people are the Balante who, belonging to a stateless society centred around the family, have shown themselves receptive to nationalist slogans of emancipation from the Portuguese and from whatever chiefs the administration has tried to give them. An individualist people, largely animist in belief, they constitute the main body of guerrilla forces. The Fulani, who are divided into subgroups according to the degree of their miscegenation with the people they previously conquered, are essentially Muslim pastoralists. Their society is hierarchical, the relationship between chiefs and their vassals being semifeudal in character. Polygamous if they can afford it, and accustomed to fighting against the animists they try to convert, the Fulanis are impervious to Portuguese culture; paradoxically, however, they have become tactical allies of the Portuguese army against the guerrillas, whom they see as a threat to their religion, feudal society, and traditions. Some smaller groups that are Fulani in culture have, however, joined the nationalists in order to emancipate themselves from the authority of elders and tribal lords. The Malinke, who are agriculturalists, are more sedentary than the Fulani, but their true vocation, now that they are no longer warriors, is trade. Skilled craftsmen, they also seek to convert their neighbours to Islām. The Mandyako are experienced in growing palm trees; the Portuguese and the insurgents are now contesting their territory. The Bram, Biafada, Felup, Bayot, Nalu, Susu, and Bijogo all live in close proximity to the sea; they are farmers and are little influenced by either Portuguese or Islāmic culture.

Assimilados and *mestiços*

Portuguese Guinea also possesses a group of detribalized Africans formerly categorized as *assimilados*; numbering only a few thousand, they are acquainted with Portuguese and European culture and are chiefly civil servants or whitecollar workers. They are not to be confused with the significant colony of Cape Verdean *mestiços* (mulattos) living in the country, who in general tend to side with the Portuguese, though the nationalist leadership, seeking to control the Cape Verde Islands as well as Portuguese Guinea itself, is also largely drawn from this group. Other nationalist leaders are Portuguese Guinean *mestiços*. The non-Africans, mainly Portuguese civil servants, number only a few thousand; the Portuguese have never used the territory for settlement. There are also, however, about 35,000 Portuguese troops permanently garrisoned in the country and who, it might be said, are hastening the promotion of Portuguese culture. Together with the Cape Verdeans, the Portuguese hold almost all posts of authority in the territory, except in insurgent areas.

Each tribe speaks its own vernacular, which in turn is divided into dialects. The Balante vernacular, for example, is divided into Berassé, Benaga, Betxa, and other dialects. The great variety of languages spoken fall into two categories. The first of these is the Mande group, which may be subdivided into a Mande-tan grouping—which in Portuguese Guinea includes Soninke (spoken by the Sarakole and the diminutive tribe of Jacanca) and the Malinke, Bambara, and Dyula languages—and the Mande-fu grouping, which includes the Susu (Soso) language. The second category, the Atlantic (West Atlantic) language group, includes all other African languages spoken in Guinea, among which Nalu, the Balante dialects, Mandyako, Pepel, Bram, Biafada, Bijagós, Felup, Bayot, and the Fulani dialects are prominent. Apart from this mixture of about 20 languages and dialects, the lingua franca is *crioulo*, a sort of Portuguese patois with an Africanized syntax and vocabulary. Resembling the *crioulo* spoken in the Cape Verde Islands, it exerts a unifying influence in the rural areas and is used by the nationalists as a teaching medium.

Portuguese is the official language and is spoken not only by the Portuguese themselves but also by other educated persons. Some Arabic is known by Muslim scholars.

Christianity has made few inroads in Portuguese Guinea. In 1964 there were about 26,500 Catholics, and there were virtually no Protestants. Islām is the dominant religion, having more than 200,000 adherents. Islām in West Africa consists of a series of religious orders; in Portuguese Guinea most adherents belong to the Qādirīyah or Tijānīyah orders. Portugal supports the expansion of Islām to help counteract Marxist influence. Even in Bissau, which is predominantly Catholic, the government builds mosques and also subsidizes pilgrimages to Mecca for Fulani and Malinke leaders. Traditional animist beliefs are nevertheless still strong, even among those inhabitants who have formally adopted Christianity or Islām.

The average population density for the country is estimated to be about 40.5 persons per square mile.

Portuguese Guinea, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
Circumscriptions (*circunscrições*)				
Bijagós	570	1,478	9,000	13,000
Fulacunda	1,573	4,075	35,000	13,000
São Domingos	642	1,663	24,000	21,000
Municipalities (*concelhos*)				
Bafatá	2,326	6,025	71,000	88,000
Bissau	304	788	56,000	119,000
Bissorã	367	950	45,000	21,000
Bolama	58	150	5,000	7,000
Cacheu*	1,076	2,788	70,000	71,000
Catió	902	2,338	38,000	10,000
Farim	1,197	3,100	57,000	15,000
Gabu	3,479	9,013	70,000	75,000
Mansôa	560	1,450	41,000	35,000
Total Portuguese Guinea	13,056†‡	33,815†‡	521,000	487,000‡

*Area and population totals for Cacheu *concelho* include Bula *concelho* which was established March 12, 1970. No breakdown of area or adjusted census figures is available. †Areas for subdivisions are land areas including some water area. The total given is the land area including some water area; the total area of Portuguese Guinea is 13,948 sq mi (36,125 sq km). ‡Figures do not add to total given because of rounding.
Source: Official government figures; UN.

The economy. The economy is based on traditional farming and on cattle raising. It is now in a chaotic state as a result of the war; even before it started, however, the economy was far from strong. African farmers sell their products to European companies, who are represented by local white or mestiço middlemen. Rice is the staple food. The production of swamp rice from the coastal plain amounts to about 170,000 tons; formerly, in record years, some was exported. The Portuguese as well as the insurgents endeavour to increase its production. Oil palm products are also associated with rice cultivation in the islands and on the coast. Coconuts are exported. Peanuts (groundnuts) form the main cash crop, being cultivated —traditionally by Muslims—in the vicinity of Bafatá, Farim, and Nova Lamego and accounting for about 60 percent of exports. Oil seeds of one kind or another—including not only peanuts but also palm products and copra—altogether account for about 90 percent of total exports, which are mostly sent to Portugal through the monopolistic Companhia União Fabril (CUF). The company has, however, been hurt economically by the various nationalist activities, which have included a boycott.

Farming and ranching

Cattle raising is particularly important among the Fulani and the Balante. In 1970, 260,000 head of cattle, 174,000 goats, and 113,000 pigs were registered; some hides and skins are exported. Fishing is insignificant. Industry is negligible, except for some oil-processing plants in Bissau.

While oil prospecting has been conducted and bauxite deposits are known to exist, the rate of investment has declined, apparently because of war operations. The balance of trade has shown a regular deficit since 1957. In 1968, for example, Portuguese Guinea imported 530,100,000 escudos (28.75 escudos = $1 U.S.; 69 escudos = £1 sterling on December 1, 1970) and exported 111,000. The value of imports, which amounted to about 297,000,000 escudos in 1961, has since grown, while exports, which amounted to 211,200,000 escudos in the same year, have subsequently declined. Insecurity has brought the economic life of the territory, formerly in the hands of CUF and its subsidiaries, into a decline. Public financing now depends upon significant subsidies from Portugal. In 1968 governmental revenue amounted to 333,000,000 escudos, which included a loan of 59,200,000 escudos from Lisbon in addition to substantial sums for health and education. The Third Six-Year Development Plan for the territory, which covers the period from 1968 to 1973, provided for investments of 210,000,000 escudos for 1969, most of which were allocated for mineral prospecting and transportation.

Balance of trade

Transport. There are no railways. Rivers provide the cheapest means of transporting peanuts and palm products, although nationalist guerrillas sometimes ambush rivercraft. In 1966 there were about 1,900 miles of roads, although the distinction between roads and improved tracks in some instances is not clear. For military purposes, a major effort is being made to build more paved roads. The main road artery runs from Bissau to Bafatá, with subsidiary roads leading to Nova Lamego in the northeast, to the riverine ports, and to neighbouring countries. Road traffic to Senegal and to the Republic of Guinea has, however, almost come to a standstill. Bissau is the only modern port. Centrally located, it handles a modest annual traffic of about 200,000 net tons of cargo; about 100 ships call each year. The only other port of note is Binda on the Rio Cacheu. Although all sea and air links with independent African states have been severed, the port of Bissau remains linked not only with Portugal but also with the Portuguese territories of Angola, Cape Verde, and Mozambique by a concessionaire steamship service. Bissau airport handles international air traffic, but this is now restricted to aircraft flying to and from Lisbon and the Cape Verde Islands. Military airfields have been built at Bafatá, Nova Lamego, and other major towns. Insecurity on the roads caused by guerrilla attacks has led to increased reliance on passenger air traffic.

Administration and social conditions. The administration of Portuguese Guinea is by a governor who is a Portuguese general. He is assisted by a secretary general. A legislative council of 14 members (of which three members are elected by direct but nonuniversal suffrage) has limited legislative powers and consultative authority. A government council composed of higher civil servants (including the commander in chief of land, sea, and air forces) and three members elected by the legislative council help the governor in his legislative and administrative functions. Illiterates being barred from the polls, the vote is largely restricted to Portuguese and to educated Africans. As an overseas province of Portugal, Portuguese Guinea is represented in the Portuguese National Assembly, in the Portuguese Corporative Chamber (an advisory body in the Portuguese central government), and in the Portuguese Overseas Council. The territory is divided into ten municipalities (*concelhos*) and three circumscriptions (*circunscrições*) headed by Portuguese administrators. Among the Muslim tribes, powers of the chiefs upholding Portuguese authority have been augmented by the Portuguese administration.

War has disrupted traditional life. Except in the outlying Bijagós and among the most central Fulanis, where the writ of the Muslim chiefs is still obeyed, most of the country is passing through a period of unprecedented change. This is particularly noticeable in education. In 1956, the African population was 99.7 percent illiterate. Eleven years later, in the period 1967 to 1968, there were 76 Catholic primary and rural schools, 48 army-run schools, and 47 official administration schools. There were also two secondary schools and a training college for primary school teachers. In the late 1960s the African nationalists claimed to have more than 150 schools in rural areas controlled by them. There are no institutions of higher learning in the country. Some scholarships are granted by the Portuguese for studies in Portugal. The nationalists have sent some scholars to study in the U.S.S.R. and Eastern Europe.

Cultural life. Bissau has a museum and a library. National radio programs are broadcast by the Portuguese from Bissau. The local press is represented by the Government Gazette, by a cultural bulletin, by a monthly trade publication, and by a daily newspaper whose publication has frequently been interrupted in recent years.

Prospects for the future. Portuguese Guinea is at war, and while it is difficult to reconcile conflicting claims by the contenders, defense is clearly the main activity of the population on either side. The Portuguese have 35,000 European and African soldiers under arms against 5,000 to 10,000 nationalist guerrillas whose strongholds are along the borders of the Republic of Guinea except in the northwest corner, in the region where the Balante live, and in some areas adjacent to Senegal. Sizable areas are said to be controlled by the Partido Africano da Independência da Guiné e Cabo Verde (PAIGC) led by Amílcar Cabral, a *mestiço* agronomist from the Cape Verde Islands who was assassinated in Conakry, Guinea, in early 1973. His partisans claim to have forced the Portuguese and their African allies out of some of their southeastern outposts and are reported to be reorganizing the African population along socialist (communal) lines by establishing hospitals, schools, and cooperatives of their own. Portuguese sources, on the other hand, report that the population resists "liberation" and point out the fact that African contingents are included in the Portuguese forces.

Strengths of the opposed military forces

It appears, however, that in the areas they dominate, the nationalist guerrillas are able to mine and ambush army convoys and to shell Portuguese fortified camps. To counteract their influence, the Portuguese evacuate outlying villages, concentrate their populations around garrisons, and protect the villagers as they till their fields. In 1969 and 1970 the Portuguese, under the slogan "For a Better Guinea," introduced a new policy stressing rural development, including the construction of schools and health clinics, in the hope of severing ties between the insurgents and communities whose allegiance was wavering. The new policy resulted in progress for some areas previously neglected. Most of the country is in the throes of a modernizing process disruptive of old traditions, the outcome of which is as yet hard to gauge.

BIBLIOGRAPHY

Geography: D.M. ABSHIRE and M.A. SAMUELS (eds.), *Portuguese Africa: A Handbook* (1969); A. TEIXEIRA DA MOTA, *Guiné Portuguesa*, 2 vol. (1954), the standard work with English and French summaries at the end of both volumes; AGENCIA GERAL DO ULTRAMAR, *Guiné: Pequena Monografía* (1961), a summary outline; R.J. HARRISON CHURCH, *West Africa*, 6th ed. (1968); CENTRO DE ESTUDOS DA GUINE PORTUGUESA, *Boletim Cultural da Guiné Portuguesa*, a scholarly and general quarterly with statistics.

Economy: BANCO NACIONAL ULTRAMARINO, *Boletim Trimestral*, an economic quarterly with sections on Portuguese Guinea; INSTITUTO SUPERIOR DE CIENCIAS SOCIAIS E POLITICA ULTRAMARINA, *Cabo Verde, Guiné, São Tomé e Príncipe* (1966), an unsurpassed wealth of details on various aspects including economy.

History and politics: JOAO BARRETO, *História da Guiné, 1418–1918* (1938), the only historical treatise; BASIL DAVIDSON, *The Liberation of Guiné* (1969), pronationalist report; C.O.N.C.P., *Guinée et Cap-Vert* (1970), an analysis of current developments seen from the PAIGC angle.

(R.P.)

Positivism and Logical Empiricism

As a philosophical ideology and movement, Positivism first assumed its distinctive features in the work of the French philosopher Auguste Comte, who also named and systematized the science of sociology. It then developed through several stages known by various names, such as Empiriocriticism, Logical Positivism, and Logical Empiricism, and finally, in the mid-20th century, flowed into the movement known as Analytic and Linguistic philosophy (*q.v.*).

Affirmations and negations

The basic affirmations of Positivism are (1) that all knowledge regarding matters of fact is based upon the "positive" data of experience, and (2) that beyond the realm of fact is that of pure logic and pure mathematics, which were already recognized by the Scottish Empiricist and Skeptic David Hume as concerned with the "relations of ideas" and, in a later phase of Positivism, were classified as purely formal sciences. On the negative and critical side, the Positivists became noted for their repudiation of metaphysics; *i.e.*, of speculation regarding the nature of reality that radically goes beyond any possible evidence that could either support or refute such "transcendent" knowledge claims. In its basic ideological posture, Positivism is thus worldly, secular, antitheological, and antimetaphysical. Strict adherence to the testimony of observation and experience is the all-important imperative of the Positivists. This imperative is reflected also in their contributions to ethics and moral philosophy, and most Positivists have been Utilitarians to the extent that something like "the greatest happiness for the greatest number of people" was their ethical maxim. It is notable, in this connection, that Auguste Comte was the founder of a short-lived religion, in which the object of worship was not the deity of the monotheistic faiths but humanity.

There are distinct anticipations of Positivism in ancient philosophy. Though the relationship of Protagoras, a 5th-century-BC Sophist, for example, to later Positivistic thought was only a distant one, there was a much more pronounced similarity in the classical Skeptic Sextus Empiricus, who lived at the turn of the 3rd century AD, and in Pierre Bayle, his 17th-century reviver. Moreover, the medieval Nominalist William of Ockham had clear affinities with modern Positivism. An 18th-century forerunner who had much in common with the Positivistic antimetaphysics of the following century was the German thinker Georg Lichtenberg.

Positivism clearly has its proximate roots, however, in the French Enlightenment, which stressed the clear light of reason, and in the 18th-century British Empiricism, particularly that of Hume and of Bishop George Berkeley, which stressed the role of sense experience. Comte was influenced specifically by the Enlightenment Encyclopaedists (such as Denis Diderot, Jean d'Alembert, and others) and, especially in his social thinking, was decisively influenced by the founder of French Socialism, Claude-Henri, comte de Saint-Simon, whose disciple he had been in his early years and from whom the very designation Positivism stems.

THE SOCIAL POSITIVISM OF COMTE AND MILL

Comte's law of three phases

Comte's Positivism was posited on the assertion of a so-called law of the three phases (or stages) of intellectual development. There is a parallel, as Comte saw it, between the evolution of thought patterns in the entire history of man, on the one hand, and in the history of an individual's development from infancy to adulthood, on the other. In the first, or so-called theological, stage, natural phenomena are explained as the results of supernatural or divine powers. It matters not whether the religion is polytheistic or monotheistic; in either case, miraculous powers or wills are believed to produce the observed events. This stage was criticized by Comte as anthropomorphic; *i.e.*, as resting on all-too-human analogies. Generally, animistic explanations—made in terms of the volitions of soullike beings operating behind the appearances—are rejected as primitive projections of unverifiable entities.

The second phase, called metaphysical, is in some cases merely a depersonalized theology: the observable processes of nature are assumed to arise from impersonal powers, occult qualities, vital forces, or entelechies (internal perfecting principles). In other instances, the realm of observable facts is considered as an imperfect copy or imitation of eternal ideas, as in Plato's metaphysics of pure Forms. Again, Comte charged that no genuine explanations result; questions concerning ultimate reality, first causes, or absolute beginnings are thus declared to be absolutely unanswerable. The metaphysical quest can lead only to the conclusion expressed by the German biologist and physiologist, Emil du Bois-Reymond: *"Ignoramus et ignorabimus"* ("We are and shall be ignorant"); it is a deception through verbal devices and the fruitless rendering of concepts as real things.

The sort of fruitfulness that it lacks can be achieved only in the third phase, the scientific, or "positive," phase —hence the title of Comte's magnum opus: *Cours de philosophie positive* (1830–42; Eng. trans., *The Positive Philosophy of Auguste Comte*, 1853)—because it claims to be concerned only with positive facts. The task of the sciences, and of knowledge in general, is to study the facts and regularities of nature and society and to formulate the regularities as (descriptive) laws; explanations of phenomena can consist in no more than the subsuming of special cases under general laws. Mankind reached full maturity of thought only after abandoning the pseudoexplanations of the theological and metaphysical phases and substituting an unrestricted adherence to scientific method.

Classification of the sciences; sociology

In his three stages Comte combined what he considered to be an account of the historical order of development with a logical analysis of the levelled structure of the sciences. By arranging the six basic and pure sciences one upon the other in a pyramid, Comte prepared the way for Logical Positivism to "reduce" each level to the one below it. He placed at the fundamental level the science that does not presuppose any other sciences—viz., mathematics—and then ordered the levels above it in such a way that each science depends upon, and makes use of, the sciences below it on the scale: thus arithmetic and the theory of numbers are declared to be presuppositions for geometry and mechanics, astronomy, physics, chemistry, biology (including physiology), and sociology. Each higher level science, in turn, adds to the knowledge content of the science or sciences on the levels below, thus enriching this content by successive specialization. Psychology is conspicuously missing in Comte's system of the sciences. Anticipating some ideas of 20th-century Behaviourism and physicalism, Comte assumed that psychology should become a branch of biology (especially of brain neurophysiology), on the one hand, and of sociology, on the other. As the "father" of sociology, Comte maintained that the social sciences should proceed from observations to general laws, very much as (in his view) physics and chemistry do. He was skeptical of introspection in psychology, being convinced that, in attending to one's own mental states, these states would be irretrievably altered and distorted. In thus insisting on the necessity of objective observation, he was close to the basic principle of the methodology of 20th-century Behaviourism.

Among Comte's disciples or sympathizers were Cesare Lombroso, an Italian psychiatrist and criminologist, and M.-P.-E. Littré, J.-E. Renan, and Louis Weber, all French philosophers.

Despite some basic disagreements with Comte, the 19th-century English philosopher John Stuart Mill, also a logician and economist, must be regarded as one of the outstanding Positivists of his century. In his *System of Logic* (1843), he developed a thoroughly Empiricist theory of knowledge and of scientific reasoning, going even so far as to regard logic and mathematics as empirical (though very general) sciences. The broadly synthetic philosopher Herbert Spencer, author of a doctrine of the "unknowable" and of a general evolutionary philosophy, was, next to Mill, an outstanding exponent of a Positivistic orientation.

THE CRITICAL POSITIVISM OF MACH AND AVENARIUS

The influences of Hume and of Comte were also manifest in important developments in German Positivism, just prior to World War I. The outstanding representatives of this school were a philosophical critic of the physics of Newton, an Austrian, Ernst Mach, who was also an original thinker as a physicist and excelled as a historian of mechanics, thermodynamics, and optics, and Richard Avenarius, founder of a philosophy known as Empiriocriticism.

Mach's phenomenalism

Mach, in the introductory chapter of his book *Beiträge zur Analyse der Empfindungen* (1886; Eng. trans., *Contributions to the Analysis of the Sensations*, 1897), reviving the Humean antimetaphysics, contended that all factual knowledge consists of a conceptual organization and elaboration of what is given in the elements; *i.e.*, in the data of immediate experience. Very much in keeping with the spirit of Comte, he repudiated the transcendental Idealism of Kant. For Mach, the most objectionable feature in Kant's philosophy was the doctrine of the *Dinge an sich* —*i.e.*, of the "things-in-themselves"—the ultimate entities underlying phenomena, which Kant had declared to be absolutely unknowable though they must nevertheless be conceived as partial causes of man's perceptions. Hermann von Helmholtz, a wide-ranging scientist and philosopher and one of the great minds of the 19th century, by contrast, held that the theoretical entities of physics are, precisely, the things-in-themselves—a view which, though generally Empiricist, was thus clearly opposed to the Positivist doctrine. Theories and theoretical concepts, according to the Positivist understanding, were merely instruments of prediction. From one set of observable data, theories formed a bridge over which the investigator could pass to another set of observable data. Positivists generally maintained that theories might come and go, whereas the facts of observation and their empirical regularities constituted a firm ground from which scientific reasoning could start and to which it must always return in order to test its validity. In consequence, most Positivists were reluctant to call theories true or false but preferred to consider them merely as more or less useful.

The task of the sciences, as it earlier had been expressed by the German physicist Gustav Kirchhoff, was the pursuit of a compendious and parsimonious description of observable phenomena. Concern with first causes or final reasons was to be excluded from the scientific endeavour as fruitless, or hopeless (if not meaningless). Even the notion of explanation became suspect and was at best taken (as already in Comte) to be no more than an ordering and connecting of observable facts and events by empirically ascertainable laws.

Anti-atomism of Mach and Ostwald

Mach and, along with him, Wilhelm Ostwald, the originator of physical chemistry, were the most prominent opponents of the atomic theory in physics and chemistry. Ostwald even attempted to derive the basic chemical laws of constant and multiple proportions without the help of the atomic hypothesis. To the Positivist the atom, since it could not be seen, was to be considered at best a "convenient fiction" and at worst an illegitimate ad hoc hypothesis. Hans Vaihinger, a subjectivist who called himself an "idealistic Positivist," pursued the idea of useful fictions to the limit, and was convinced that the concept of the atom, along with the mathematical concepts of the infinite and the infinitesimal, and those of causation, free will, the economic man, and the like, were altogether fictitious, some of them even containing internal contradictions.

The anti-atomistic strand in the thought of the Positivists was an extreme manifestation of their phobia regarding anything unobservable. With the undeniably great success of the advancing microtheories in physics and chemistry, however, the Positivist ideology was severely criticized, not only by some contemporary philosophers but also by outstanding scientists. The Austrian Ludwig Boltzmann and the German Max Planck, for example, both top-ranking theoretical physicists, were in the forefront of the attack against Mach and Ostwald. Boltzmann and Planck, outspoken Realists, were deeply convinced of the reality of unobservable microparticles, or microevents, and were clearly impressed with the ever-growing and converging evidence for the existence of atoms, molecules, quanta, and subatomic particles. Nevertheless, the basic Positivist attitude was tenaciously held by many scientists, and striking parallels to it appeared in American Pragmatism and instrumentalism; in parts of the work of the Pragmatists Charles Sanders Peirce, William James, and John Dewey, there is a philosophy of pure experience essentially similar to that of Mach.

The Positivism of Avenarius

Though Richard Avenarius has not become widely known, he too anticipated a good deal of what the American Pragmatists propounded. His Positivism, like that of Mach, comprised a biologically oriented theory of knowledge. From the needs of organisms in their adaptation to the exigencies of their environment develop the conceptual tools needed for prediction of future conditions. In Avenarius' view, the raw material of the construction of the concepts of common sense and of the sciences, however, was "the given"; *i.e.*, the data of immediate sensory experience. Just as Mill in the 19th century considered ordinary physical objects as "permanent possibilities of sensation," so Mach and Avenarius construed the concepts pertaining to what men commonsensically regard as the objects of the real world as "complexes of sensations." Thus, it was maintained that a stone, for example, is no more than a collection of such sensory qualities as hardness, colour, and mass. The traditional assumption that there must be an underlying substance that has these properties was repudiated. To the question "What would be left over if all of the perceptible qualities were stripped (in thought) away from an observable object?" these Positivists answered: "Precisely nothing." Thus the concept of substance was declared not only superfluous but meaningless as well.

In similar fashion, the concept of causation was explicated not as a real operating principle but as regularity of succession or as functional dependency among observable or measurable variables. Because these dependencies are not logically necessary, they are contingent and ascertained by observation, and especially by experimentation and inductive generalization.

The Newtonian doctrine according to which space and time are absolute or substantive realities had been incisively criticized by the 17th-century Rationalist Gottfried Leibniz, and was subjected by Mach to even more searching scrutiny. While Leibniz had already paved the way for the conception of space and time as exclusively a matter of relations between events, Mach went still further in attacking the arguments of Newton, then still highly persuasive, in favour of a dynamic and absolute space and time. In particular, the inertial and centrifugal forces that arise in connection with accelerated or curvilinear motions had been interpreted by Newton as effects of such motions with respect to a privileged reference medium imagined as an absolute Cartesian mesh system graphed upon a real space. In a typically Positivistic manner, however, Mach found the idea quite incredible. How, he asked, could an absolutely empty space have such powerful effects? Mach conjectured that any privileged reference system must be generated not by an imperceptible grid but by material reality—specifically, by the total mass of the universe (galaxies and fixed stars), an idea that later served as an important starting point for Einstein's general theory of relativity and gravitation.

Influence of Mach and Avenarius

The Positivist theory of knowledge, as proposed by Mach and Avenarius, impressed many scholars, most notable among whom was probably the leading British logician and philosopher Bertrand Russell in one of the earlier phases of his thought. In a work entitled *Our Knowledge of the External World* (1914), Russell analyzed the concept of physical objects as comprising classes of (perceptual) aspects or perspectives, an idea that later stimulated the work of Rudolf Carnap, an outstanding philosophical semanticist and Analyst, entitled *Der logische Aufbau der Welt* (1928; Eng. trans., *The Logical Structure of the World*, 1967). Mach remained the most influential thinker among Positivists for a long time, though some of his disciples, like Josef Petzoldt, are now largely forgotten. But *The Grammar of Science* (1892),

written by Karl Pearson, a scientist, statistician, and philosopher of science, still receives some attention; and in France it was Abel Rey, also a philosopher of science, who, along the lines of Mach, severely criticized the traditional mechanistic view of nature. In the United States, John Bernard Stallo, a German-born American philosopher of science (also an educator, jurist, and statesman), developed a Positivistic outlook, especially in the philosophy of physics, in his book *The Concepts and Theories of Modern Physics* (1882), in which he anticipated to a degree some of the general ideas later formulated in the theory of relativity and quantum mechanics.

LOGICAL POSITIVISM AND LOGICAL EMPIRICISM

A first generation of 20th-century Viennese Positivists began its activities, strongly influenced by Mach, around 1907. Notable among them were a physicist, Philipp Frank, mathematicians Hans Hahn and Richard von Mises, and an economist and sociologist, Otto Neurath. This small group was also active during the 1920s in the Vienna Circle of Logical Positivists, a seminal discussion group of gifted scientists and philosophers that met regularly in Vienna, and in the related Berlin Society for Empirical Philosophy.

The Vienna Circle and the Berlin Society

These two schools of thought, destined to develop into an almost worldwide and controversial movement, were built on the Empiricism of Hume, on the Positivism of Comte, and on the philosophy of science of Mach. Equally important influences came from several eminent figures who were at the same time scientists, mathematicians, and philosophers—G.F. Bernhard Riemann, the author of a non-Euclidean geometry; Hermann von Helmholtz, a pioneer in a broad range of scientific studies; Heinrich Hertz, the first to produce electromagnetic waves in his laboratory; Ludwig Boltzmann, a researcher in statistical mechanics; Henri Poincaré, equally eminent in mathematics and philosophy of science; and David Hilbert, distinguished for his formalizing of mathematics. Most significant, however, was the impact of Albert Einstein, as well as that of the three great mathematical logicians of the past 100 years—the ground-breaking German Gottlob Frege and the authors of the monumental *Principia Mathematica* (1910–13), Bertrand Russell and Alfred North Whitehead.

The earlier Positivism of Viennese heritage. The confluence of ideas from these sources and the impressions that they made upon the Vienna and Berlin groups in the 1920s gave rise to the philosophical outlook of Logical Positivism—a label supplied in 1931 by A.E. Blumberg and the Minnesota philosopher of science Herbert Feigl. The leader of the Vienna Circle between 1924 and 1936 was Moritz Schlick, who in 1922 succeeded to the chair (previously held by Mach and Boltzmann) for the philosophy of the inductive sciences at the University of Vienna. By 1924 an evening discussion group had been formed with Schlick, Hans Hahn, Otto Neurath, Victor Kraft, Kurt Reidemeister, and Felix Kaufmann as the prominent active participants. The most important addition to the circle was Rudolf Carnap who joined it in 1926. One of the early activities was the study and critical discussion of the *Tractatus Logico-Philosophicus* (1922) of Ludwig Wittgenstein, a seminal thinker in Analytical and Linguistic philosophy. It seemed at the time that the views of Carnap and Wittgenstein, though formulated and elaborated quite differently, shared a large measure of basic agreement. Parallel, but not completely independent, developments occurred in the Berlin group, in which Hans Reichenbach, Richard von Mises, Kurt Grelling, and Walter Dubislav were the leading spirits.

Scientific, antimetaphysical orientation

Both the Vienna and Berlin groups consisted mainly of philosophically interested scientists or scientifically trained and oriented philosophers. Schlick had already anticipated some of the basic epistemological tenets of the groups in his *Allgemeine Erkenntnislehre* (1918; "General Theory of Knowledge"). But the philosophical outlook was sharpened and deepened when, in the late 1920s, the Viennese Positivists published a pamphlet *Wissenschaftliche Weltauffassung: Der Wiener Kreis* (1929; "Scientific Conception of the World: The Vienna Circle"), which was to be their declaration of independence from traditional philosophy—and, in the minds of its authors (Carnap, Hahn, and Neurath, aided by Waismann and Feigl), a "philosophy to end all philosophies."

Language and the clarification of meaning. The basic ideas of Logical Positivism were roughly as follows: the genuine task of philosophy is to clarify the meanings of basic concepts and assertions (especially those of science) —and not to attempt to answer unanswerable questions such as those regarding the nature of ultimate reality or of the Absolute. Inasmuch as an extremely ambitious Hegelian type of metaphysics, Idealistic and absolutist in orientation, was still prevalent in the German-speaking countries, there were many who believed that the antidote was urgently needed. Moreover, the Logical Positivists also had only contempt and ridicule for the ideas of the German Existentialist, Martin Heidegger, whose interminable torment regarding such questions as "Why is there anything at all?" and "Why is what there is, the way it is?" and especially his pronouncements about Nothingness seemed to them to be not only sterile but so confused as to be nonsensical. The Logical Positivists viewed metaphysics as a hopelessly futile way of trying to do what great art, and especially poetry and music, already do so effectively and successfully. These activities, they held, are expressions of visions, feelings, and emotions and, as such, are perfectly legitimate as long as they make no claims to genuine cognition or representation of reality. What Logical Positivism recommended positively, on the other hand, was a logic and methodology of the basic assumptions and of the validation procedures of knowledge and of evaluation.

An adequate understanding of the functions of language and of the various types of meaning is another of the fundamentally important contributions of the Logical Positivists. Communication and language serve many diverse purposes: one is the representation of facts, or of the regularities in nature and society; another is the conveying of imagery, the expression and arousal of emotions; a third is the triggering, guidance, or modification of actions. Thus, they distinguished cognitive–factual meaning from expressive and evocative (or emotive) significance in words and sentences. It was granted that, in most utterances of everyday life (and even of science), these two types of meaning are combined or fused. What the Logical Positivists insisted upon, however, was that the emotive type of expression and appeal should not be mistaken for one having genuinely cognitive meanings. In such expressions as moral imperatives, admonitions, and exhortations there is, of course, a factually significant core; viz., regarding the (likely) consequences of various actions. But the normative element—expressed by such words as "ought," "should," "right," and their negations (as in "Thou shalt not . . . .")—is by itself not cognitively meaningful but has primarily emotional and motivative significance.

Cognitive versus emotive meaning

Early statements about moral value judgments, such as those by Carnap or by A.J. Ayer, a more radical British Positivist, seemed shocking to many philosophers, to whom it seemed that, in their careless formulation, moral norms were to be treated like expressions of taste. Equally shocking was their condemnation as nonsense (really non-sense; *i.e.*, complete absence of factual meaning) of all moral, aesthetic, and metaphysical assertions. More adequate and delicate analyses, such as that of the American Positivist Charles Stevenson, were soon to correct and modify those extremes. By proper allocation of the cognitive and the normative (motivative) components of value statements, many thinkers rendered the originally harsh and implausible Positivist view of value judgments more acceptable. Nevertheless, there is—in every Positivistic view—an ineluctable element of basic, noncognitive commitment in the acceptance of moral, or even of aesthetic, norms.

The verifiability criterion of meaning and its offshoots. The most noteworthy, and also most controversial, contribution of the Logical Positivists was the so-called verifiability criterion of factual meaningfulness. In its original form, this criterion had much in common with

the earlier Pragmatist analysis of meaning (as in Charles Sanders Peirce and William James). Schlick's rather careless formulation: "The meaning of a [declarative sentence] is the method of its verification," which was really intended only to exclude from the realm of the cognitively meaningful those sentences for which it is logically inconceivable that either supporting or refuting evidence can be found, was close to the Pragmatist and, later, to the operationalist slogan that may be paraphrased by: "A difference must make a difference in order to be a difference"—or (more fully explicated): only if there is a difference in principle, open to test by observation, between the affirmation and the denial of a given assertion does this assertion have factual meaning. To take the classical example from Hume's analysis of the concept of causation, there is no difference between saying "*A* is always followed by *B*" and saying "*A* is necessarily always followed by *B*." That all *effects* have causes is true by virtue of the (customary) definitions of "cause" and "effect"; it is a purely formal or logical truth. But to say (instead of speaking of effects) that all *events* have causes is to say something factual—and conceivably false. (It should be noted that these rather crude uses of "causality" and "necessity" were later replaced by much more subtle analyses.)

Untestability of causation, ether, simultaneity

One of the most important examples that stimulated the formulation of the meaning criterion was Einstein's abandonment, in 1905, of the ether hypothesis and of the notion of absolute simultaneity. The hypothesis that there exists a universal ether, as a medium for the propagation of light (and of electromagnetic waves generally), had been quite plausible and was widely accepted by physicists during the second half of the 19th century. To be sure, there were a number of serious difficulties with the idea: the properties that had to be ascribed to the ether were difficult to conceive in a logically compatible manner; and the ether hypothesis in the last stage of its development (by the Dutch physicist Hendrik Lorentz and the Irish physicist George Fitzgerald) had become objectionable in that it sought to provide excuses for the absolute unobservability of that mysterious, allegedly all-pervasive, universal substance. Similarly, it had become impossible, except at the price of intolerably ad hoc hypotheses, to maintain the notions of absolute time and of absolute simultaneity. Thus Einstein, by eliminating these empirically untestable assumptions, was led to his special theory of relativity.

Modifications of the meaning criterion

Several important changes in the formulation of the meaning criterion took place in the ensuing decades from 1930 to 1960. The original version formulated in terms of verifiability was replaced by a more tolerant version expressed in terms of testability or confirmability. Obviously, universal propositions, such as "All cats have claws," being only partially supportable by positive instances (one cannot examine every cat that exists), are not conclusively verifiable. Nevertheless, scientists do accept lawlike statements on the basis of only incomplete, as well as indirect, verification—which is what "confirmation" amounts to. It was in coming to this juncture in his critique of Positivism that Karl Popper, an Austro-English philosopher of science, in his *Logik der Forschung* (1935; Eng. trans., *The Logic of Scientific Discovery*, 1959), insisted that the meaning criterion should be abandoned and replaced by a criterion of demarcation between empirical (scientific) and transempirical (nonscientific, metaphysical) questions and answers—a criterion that, according to Popper, is to be testability, or, in his own version, falsifiability; *i.e.*, refutability. Popper was impressed by how easy it is to supposedly verify all sorts of assertions—those of psychoanalytic theories seemed to him to be abhorrent examples. But the decisive feature, as Popper saw it, should be whether it is in principle conceivable that evidence could be cited that would refute (or disconfirm) a given law, hypothesis, or theory. Theories are (often) bold conjectures. It is true that scientists should be encouraged in their construction of theories—no matter how far they deviate from the tradition; it is also true, however, that all such conjectures should be subjected to the most severe and searching criticism and experimental scrutiny of their truth claims. The growth of knowledge thus proceeds through the elimination of error; *i.e.*, through the refutation of hypotheses that are either logically inconsistent or entail empirically refuted consequences.

Despite valuable suggestions in Popper's philosophy of science, the Logical Positivists and Empiricists continued to reformulate their criteria of factual meaningfulness. The Positivist Hans Reichenbach, who emigrated from Germany to California, proposed, in his *Experience and Prediction* (1938), a probabilistic conception. If hypotheses, generalizations, and theories can be made more or less probable by whatever evidence is available, he argued, then they are factually meaningful. In another version of meaningfulness, first adumbrated by Schlick (under the influence of Wittgenstein's thought), the philosopher's attention is focussed on concepts rather than on propositions. If the concepts in terms of which theories are formulated can be related, through chains of definitions, to concepts that are definable ostensibly—*i.e.*, by pointing to or exhibiting items or aspects of direct experience—then those theories are factually meaningful. This is the version also advocated by Richard von Mises, an Austro-American mathematician and methodologist, in his *Positivism* (1951) and, later, more technically elaborated by Carnap, in *Minnesota Studies in the Philosophy of Science*, vol. 1 (1956).

Other issues. These views of meaningfulness are essentially refinements of the doctrine of so-called protocol sentences, developed in the late 1920s and early 1930s and elaborated especially by Carnap, by Otto Neurath, a polymath sociologist and philosopher, and also (with some differences) by Schlick. Protocol sentences, originally conceived along the lines of an interpretation, developed in the Vienna Circle, of Wittgenstein's elementary propositions, were identified as those sentences that make statements about the data of direct experience. But Neurath—and independently also Popper—warned of the danger that this doctrine might lead to subjective Idealism and recommended that it be given a rational reconstruction on an intersubjective basis. Thus, Neurath and Carnap preferred that a physicalistic thing-language be employed as the starting point and testing ground of all knowledge claims. Propositions in this language would describe objectively existing, directly observable states of affairs or events. Because all objective and intersubjective knowledge was seen, in such a physicalism, to rest on statements representing things and their properties, relations and ongoing processes as they are found in unbiassed, and presumedly theory-free, observation, the physicalists were thus proclaiming a first thesis of the so-called Unity of Science principle. Though Mach had proceeded from the basis of (neutral) immediate experience, his insistence on the unity of all knowledge and all science was retained—at least in general spirit—by the later Positivists. In this view, all classifications of the sciences, or divisions of their subject matter, were seen as artificial, valuable at best only administratively, but without philosophical justification.

Physicalism and the Unity of Science

Sharply to be distinguished from this first thesis of the Unity of Science is a second that formulates a reductionism of a very different type: whereas the first thesis concerns the unity of the observational basis of all the sciences, the second proposes (tentatively) a unity of the explanatory principles of science. Reductions within physics itself, such as that of thermodynamics to the kinetic theory of heat (statistical mechanics) and of optics to electromagnetics; and, beyond that, the explanation of chemical phenomena, with the help of the quantum theory, in terms of atomic and molecular processes; and, furthermore, the progress that has been made in the physical explanation of biological phenomena (especially in the recent development of molecular biology)—all of these encourage the idea of a unitary set of physical premises from which the regularities of all of reality could be derived. But it must be admitted that, in contrast to the first thesis (which, by comparison is almost trivial), the second, being a bold conjecture about future reductions in the sciences, might well prove to be limited in the scope

of its validity. The most controversial part of this reductionist ideology, however, concerns the realms of organic life and especially that of mind; it concerns, in other words, the reducibility of biology to physics and chemistry and of psychology to neurophysiology—and (though this is clearly utopian at present) of both ultimately to basic physics.

The most serious alternative to this reducibility thesis of the Unity of Science movement is the theory of emergent evolution, according to which life or mind (or both) are genuinely novel forms of reality that could not possibly have been derived from, or predicted by, any laws or theories of the lower or earlier levels of existence.

Historically, it may be plausible that the notorious perplexities of the traditional problem of how mind relates to body motivated both the phenomenalistic Positivists as well as the Behaviourists and physicalists. In either view, the mind–body problem conveniently disappears; it is branded as a metaphysical pseudoproblem. The phenomenalism of Mach and the early Russell was expressed in a position called neutral monism, according to which both psychological and physical concepts are viewed as logical constructions on the basis of a neutral set of data of immediate experience. There are thus not two realities—the mental and the physical; there are merely different ways of organizing the experiential data. In the Behaviourist–physicalist alternative, on the other hand, the philosopher, considering the concepts that are ordinarily taken to characterize private mental acts and processes, defines them on the basis of publicly (intersubjectively) observable features of the behaviour, including the linguistic behaviour, of man.

Neutral monism and mental states

The absolute privacy of mental events was first criticized, however, by Carnap and later by an Oxford Analytical philosopher, Gilbert Ryle; and Wittgenstein, in an argument against the very possibility of a private language, maintained that, unless men have objective criteria for the occurrence of mental states, they cannot even begin to communicate meaningfully with each other about their direct experiences. Wittgenstein thus repudiated the traditional view according to which a man's knowledge of other persons' minds must be based on analogical inference from his own case. In a similar vein, an American psychologist, B.F. Skinner, tried to account for man's acquisition of subjective terms in his language by a theory of verbal behaviour. A man learns to describe his mental states, explained Skinner, from the utterances of others who ascribe these states to him by virtue of their observation of his behaviour (*e.g.*, in the social context; or when a certain stimulus situation prevails in his environment).

Dispositions as conditionals

Both Carnap and Ryle have emphasized that many mental features or properties have a dispositional character. Dispositional terms, whether used in psychology or more broadly, have to be understood as shorthand expressions for test conditions—or test-result conditionals. Thus, even in ordinary life, a man appraises, for example, the intelligence of a person in the light of what he does, how he does it, how fast he does it, when confronted with various tasks or problems. Just as such physical properties as malleability, brittleness, or thermal or electrical conductivity must be defined in terms of what happens when certain conditions are imposed, so also mental dispositions are to be construed as similarly hypothetical; *i.e.*, as (in the simplest case) stimulus–response relationships.

The later Positivism of Logical Empiricism. Logical Positivism, essentially the doctrine of the Vienna Circle (*c.* 1924–38), underwent a number of important changes and innovations in the middle third of the century, which suggested the need for a new name. The designation Positivism had been strongly connected with the Comte–Mach tradition of instrumentalism and phenomenalism. The emphasis that this tradition had placed, however, on the positive facts of observation and their negative attitude toward the atomic theory and the existence of theoretical entities in general were no longer in keeping with the spirit of 20th-century science. Nevertheless, the requirement that hypotheses and theories be empirically testable, though it became more flexible and tolerant, could not be relinquished. It was natural, then, that the word "empiricism" should occur in any new name. Accordingly, retaining the term "logical" in roughly its same earlier meaning, the new name "Logical Empiricism" was coined.

The status of the formal and a priori. The intention of the word "logical" was to insist on the distinctive nature of logical and mathematical truth. In opposition to Mill's view, according to which even logic and pure mathematics are empirical (*i.e.*, are justifiable or refutable by observation), the Logical Positivists—essentially following Frege and Russell—had already declared mathematics to be true only by virtue of postulates and definitions. Expressed in the traditional terms used by Kant, logic and mathematics were recognized as a priori disciplines (valid independently of experience) precisely because their denial would amount to a self-contradiction, and statements within these disciplines were expressed in what Kant called analytic propositions; *i.e.*, propositions that are true or false only by virtue of the meanings of the terms they contain. In his own way, the broad-ranging scholar Gottfried Leibniz had seen this in the 17th century long before Kant. The truth of such a simple arithmetical proposition as, for example, "$2 + 3 = 5$" is necessary, universal, a priori, and analytic because of the very meaning of "2," "+," "3," "5," and "=." Experience could not possibly refute such truths because their validity is established (as Hume said) merely by the "relation of ideas." Even if—"miraculously"—putting two and three objects together should on some occasion yield six objects, this would be a fascinating feature of those objects (rabbits, perhaps); but it would not in the least tend to refute the purely definitional truths of arithmetic.

The analytic and synthetic in mathematics

The case of geometry is altogether different. Geometry can be either an empirical science of natural space or an abstract system with uninterpreted basic concepts and uninterpreted postulates. The latter is the conception introduced in rigorous fashion by a late-19th-century mathematician, David Hilbert, and, still later, by a United States geometer, Oswald Veblen. In the axiomatizations that they developed, the basic concepts, called primitives, are implicitly defined by the postulates: thus, such concepts as point, straight line, intersection, betweenness, and plane are related to each other in a merely formal manner. The proof of theorems from postulates, and with explicit definitions of derived concepts (such as of triangle, polygon, circle, or conic section), is achieved by strict deductive inference. Very different, however, is geometry as understood in practical life, and in the natural sciences and technologies, in which it constitutes the science of space. Ever since the development of the non-Euclidean geometries in the first half of the 19th century, it has no longer been taken for granted that Euclidean geometry is the only geometry uniquely applicable to the spatial order of physical objects or events. In Einstein's general theory of relativity and gravitation, in fact, a four-dimensional Riemannian geometry with variable curvature was successfully employed, an event that amounted to a final refutation of the Kantian contention that the truths of geometry are "synthetic a priori." With respect to the relation of postulates to theorems, geometry is thus analytic, like any other rigorously deductive discipline; but the postulates themselves, when interpreted—*i.e.*, when construed as statements about the structure of physical space—are indeed synthetic, but also a posteriori; *i.e.*, their adequacy depends upon the results of observation and measurement.

Developments in Linguistic Analysis and their offshoots. Important contributions, beginning in the early 1930s, were made by Carnap, by Kurt Gödel, a Moravian-American mathematical logician, and others to the logical analysis of language. Charles Morris, a Pragmatist concerned with Linguistic Analysis, had outlined the three dimensions of semiotics (the general study of signs and symbolisms): syntax, semantics, and pragmatics (the relation of signs to their users and to the conditions of their use). Syntactical studies, concerned with the formation and transformation rules of language

(*i.e.*, its purely structural features), soon required supplementation by semantical studies, concerned with rules of designation and of truth. Semantics, in the strictly formalized sense, owed its origin to Alfred Tarski, a leading member of the Polish school of logicians, and was then developed by Carnap and applied to problems of meaning and necessity. As Wittgenstein had already shown, the necessary truth of tautologies simply amounts to their being true under all conceivable circumstances. Thus the so-called eternal verity of the principles of identity (p is equivalent with itself), of noncontradiction (one cannot both assert and deny the same proposition), and the principle of excluded middle (any given proposition is either true or false; there is no further possibility) is an obvious consequence of the rules according to which the philosopher uses (or decides to use) the words "proposition," "negation," "equivalence," "conjunction," "disjunction," and others. Quite generally, questions regarding the meanings of words or symbols are answered most illuminatingly by stating the syntactical and the semantical rules according to which they are used.

Ordinary versus artificial language

Two different schools of thought originated from this basic insight: (1) the philosophy of "ordinary language" Analysis—initiated by Wittgenstein, especially in his later work, and (following him) developed in differing directions in the works of Gilbert Ryle and John Langshaw Austin, both Oxford philosophers, of the Cambridge Analyst John Wisdom, and others; and (2) the ideology, essentially that of Carnap, usually designated as logical reconstruction, which builds up an artificial language. In the procedures of "ordinary language" Analysis, an attempt is made to trace the ways in which a person commonly expresses himself. In this manner, many of the traditional vexatious philosophical puzzles and perplexities are shown to arise out of deviant uses of language. (Lewis Carroll had already anticipated some of these oddities in his whimsical manner in *Alice in Wonderland.*) The much more rigorous procedures of the second school—of Tarski, Carnap, and many other logicians—rest upon the obvious distinction between the language (and all of its various symbols) that is the object of analysis, called the object language, and that in which the analysis is formulated, called the metalanguage. If needed and fruitful, this process can be repeated, in that the erstwhile metalanguage can become the object of a metametalanguage, and so on—without the danger of a vicious infinite regress.

With the help of semantical concepts, an old perplexity in the theory of knowledge can then be resolved. Positivists have often tended to conflate the truth conditions of a statement with its confirming evidence, a procedure which has led to certain absurdities committed by phenomenalists and operationalists, such as the pronouncement that the meanings of statements about past events consist in their (forthcoming future) evidence. Clearly, the objects—the targets or referents—of such statements are the past events. Thus the meaning of a historical statement is its truth conditions; *i.e.*, the situation that would have to obtain if the historical statement is to be true. The confirmatory evidence, however, may be discovered either in the present or in the future. Similarly, the evidence for an existential hypothesis in the sciences may consist, for example, in cloud-chamber tracks, spectral lines, or the like, whereas the truth conditions may relate to subatomic processes or to astrophysical facts. Or, to take an example from depth psychology, the occurrences of unconscious wishes or conflicts are the truth conditions for which the observable symptoms (Freudian lapses, manifest dream contents, and the like) serve merely as indicators or clues; *i.e.*, as items of confirming evidence.

The third dimension of language (in Morris' view of semiotic)—*i.e.*, the pragmatic aspect—has not been as fully and formally analyzed as the other two dimensions. There is even some disagreement as to whether some of the cognitive activities (such as verifying, refuting, or interpreting), in addition to the noncognitive functions (such as the motivative and persuasive appeals), are to be included in studies of pragmatics.

Undecidability; probability

One of the most surprising and revolutionary offshoots of the metalinguistic (formal) analyses was Kurt Gödel's discovery, in 1931, of an exact proof of the undecidability of certain types of mathematical problems, a discovery that dealt a severe blow to the expectations of the Formalistic school of mathematics championed by Hilbert and his collaborator, Paul Bernays. Before Gödel's discovery, it had seemed plausible that a mathematical system could be complete in the sense that any well-formed formula of the system could be either proved or disproved on the basis of the given set of postulates. But Gödel showed rigorously (what had been only a conjecture on the part of the Dutch Intuitionist L.E.J. Brouwer and his followers) that, for a large class of important mathematical systems, such completeness cannot be achieved.

Both Carnap and Reichenbach, in their very different ways, made extensive contributions to the theory of probability and induction. Impressed with the need for an interpretation of the concept of probability that was thoroughly empirical, Reichenbach elaborated a view that conceived probability as a limit of relative frequency and buttressed it with a pragmatic justification of inductive inference. Carnap granted the importance of this concept (especially in modern physical theories) but attempted, in increasingly refined and often revised forms, to define a concept of degree-of-confirmation that was purely logical. Statements ascribing an inductive probability to a hypothesis are, in Carnap's view, analytic, because they merely formulate the strength of the support bestowed upon a hypothesis by a given body of evidence.

CURRENT CRITICISMS AND CONTROVERSIES

Logical Positivism and Logical Empiricism have from their very beginnings been subjected to searching criticisms. At first it was the verifiability criterion of meaningfulness that produced a storm of oppositions. One group of critics asked whether the criterion was meaningful in the light of its own standard. Carnap replied that the criterion itself was not intended as a factual assertion but rather as a proposal for a better and clearer use of language. Nevertheless, the Logical Empiricists felt that the (tolerant–liberal) formulation of the meaning criterion —far from being an arbitrary injunction—came rather close to what enlightened common sense, and especially the scientific attitude, intended by the difference between sense and non-sense.

Scientific explanation

Other recent and current criticisms concern the Logical Empiricist views regarding the nature of scientific explanation, in regard to which matters are not as simple as they were originally conceived to be. Closer attention to the history of scientific theories has revealed important discontinuities, or revolutions, in the conceptual schemes of the sciences. The significant role of statistical (or probabilistic) explanations in most modern sciences, for example, is receiving increasingly sophisticated analyses.

Nevertheless, the prevalence of scientific revolutions and anarchism or subjectivism in scientific method has been exaggerated in differing ways by several scholars. As has been conceded by all competent philosophers of science and even by the greatest scientist-philosophers of the century—Albert Einstein, Niels Bohr, Erwin Schrödinger, and others—there is no straight logical path, no standard recipe, by which to move from the data of observation and arrive at scientific theories. It may also be admitted that, though scientific creativity has psychologically much in common with artistic creation, the criteria of appraisal are certainly quite different. And, although, from the present critical point of view, all and any scientific assertions are in principle subject to revision, it is nonetheless felt to be grotesque to deny the relative stability of the empirical laws that serve as the testing ground of alternative theories.

BIBLIOGRAPHY

Literature on Classical Positivism: J. WATSON, *Comte, Mill and Spencer* (1895); W.M. SIMON, *European Positivism in the Nineteenth Century* (1963); JOHN STUART MILL, *Auguste Comte and Positivism* (1865); AUGUSTE COMTE, *Cours de philosophie positive*, 6 vol. (1830–42; Eng. trans. and cond. by H. MARTINEAU, *The Positive Philosophy of Auguste Comte*,

2 vol., 1853). For further references to the ethical views of the classical Positivists, see the bibliography of Utilitarianism in the *Encyclopedia of Philosophy*.

Literature of Logical Positivism and Logical Empiricism: Many of the original classics, both books and articles, are listed in the ample bibliography of A.J. AYER (ed.), *Logical Positivism* (1959), an anthology that contains, among other important essays, RUDOLF CARNAP's "Psychology in Physical Language." The early history of Viennese Positivism is well told in VICTOR KRAFT, *Der Wiener Kreis: Der Ursprung des Neopositivismus* (1950, 2nd ed. 1968; Eng. trans., *The Vienna Circle*, 1953, reprinted 1969). Another important source is J. JOERGENSEN, *The Development of Logical Empiricism*, vol. 2, no. 9 of the *International Encyclopedia of Unified Science* (1951). For a brief account of the European movement of Logical Positivism and its migration and impact in the United States, see H. FEIGL, "The Wiener Kreis in America," in D. FLEMING and B. BAYLIN (eds.), *The Intellectual Migration: Europe and America 1930–1960* (1969). Books, mainly in the foundations of the sciences, but also in philosophy of language and epistemology, many by the leading Logical Empiricists, are listed in the ample *Bibliography and Index*, in HERBERT FEIGL and CHARLES MORRIS (eds.), *International Encyclopedia of Unified Science*, vol. 2, no. 10 (1969). Of direct relevance are the major works of R. Carnap, O. Neurath, M. Schlick, P. Frank, H. Reichenbach, E. Nagel, C.G. Hempel, R. von Mises, and Charles Morris. For criticisms, those of KARL R. POPPER may be used; and the intellectual autobiography of Carnap, the 26 descriptive and critical essays, and his replies, in P.A. SCHILPP (ed.), *The Philosophy of Rudolf Carnap* (1963). For recent reactions, see P. ACHINSTEIN and S.F. BARKER (eds.), *The Legacy of Logical Positivism: Studies in the Philosophy of Science* (1969); and *Minnesota Studies in the Philosophy of Science*, vol. 1–5 (1956–70).

(H.Fe.)

Postal Systems

In the modern world, the term postal system may be defined as the institution—almost invariably under the control of a government agency—that makes it possible for any person to send a letter, packet, or parcel to any addressee, in his own country or abroad, in the expectation that it will be conveyed according to certain established standards of regularity, speed, and security. The service is paid for in advance by the sender according to a relatively simple scale of fees based on weight. Prepayment is ordinarily made by means of postage stamps or a franking machine impression, and the addressee is not usually called upon to make any payment for delivery.

It may seem tedious to catalog at such length characteristics that are familiar to everyone. Their very familiarity, however, and the consequent unthinking acceptance of them make it important to emphasize that until recently postal systems lacked many of these features. Though the basic need for a system to exchange written communications has been felt by all human societies and has been met in many ways, the evolution of varied postal systems adopted by different societies through the centuries into the basically similar pattern of today's state monopoly service has been a long and difficult process.

Today the governments of many countries use their postal systems to provide a range of services that often have no direct connection with the traditional function of exchanging letters. To provide for the collection, transport, and delivery of letters throughout a country, it has been necessary to establish a network of post offices extending into the remotest areas. Such a network of offices, staffed by agents of the state, provides an efficient banking and money-transfer service in areas in which it would be uneconomic for a commercial or state bank to establish a branch office. Many governments also pay various social security benefits—such as pensions and family allowances—through vouchers that can be cashed at post offices; they also collect certain taxes, normally through the sale of licenses, and revenue or tax stamps. In fact, a wide range of ancillary services is provided through the postal system. In some African and Asian countries, for instance, the postal system helps distribute antimalarial drugs. While the variety of such services significantly indicates the ever-increasing use made of postal systems throughout the modern world, they lie outside the scope of this article, which concentrates on the traditional postal or mail services. The functioning of these services is described, followed by the chief characteristics of the postal service in certain countries, the international system (including the functioning of the Universal Postal Union), and, finally, the impact of modern technology on postal work.

POSTAL OPERATIONS AND MANAGEMENT

The raw material of the postal services, always a single object that demands individual treatment, is something sent by one person or entity to another person or entity who may be anywhere else in the world. Letters and parcels alike go in all shapes and sizes, subject only to the limits of weight and dimensions prescribed by postal legislation. Yet if postal services are to be efficient and economical, the processing of these items, as far as is possible, must be on a mass production basis. Thus, the basic function of postal organization is to convert the individual item as rapidly as possible into something that can be handled on a bulk basis, ensuring, however, that it finally regains its individual status.

The sorting process

The collection and sorting of individual items by the most economic method, concentrating together all items that are going to the same place or in the same direction, involves the use of local transport, usually operated by the postal services themselves, and sorting offices. The size of the sorting office depends on local requirements, but some are, in fact, large industrial concerns that handle several million items a day and employ thousands of personnel. Certain ancillary operations are involved, such as, for instance, the cancelling (by machine in all except the smallest offices) of stamps that have served to prepay the postal charge.

The next stage is to transport the grouped items to different destinations that may be intermediate sorting offices, perhaps a railway sorting carriage. Items are combined by further sorting with mail from other sources to permit grouping the mails for final destinations.

The third stage is the arrival of the mail at the sorting office of the final destination, where it is sorted systematically. The items finally recover their identity and are grouped together for delivery to the individual address. In most countries, delivery is on a house-to-house basis, although boxes at a local post office are sometimes used.

All stages have to be planned and dovetailed together to meet an overall standard of performance. The transport of mail between sorting offices is normally by services not necessarily directly operated by the post office—*i.e.*, by truck, bus, train, marine shipping, or air services. In some countries, however, the administration operates its own air service, usually by night, to supplement the public services. Post office counter services (*i.e.*, the public offices that existed originally as depots where the public could buy stamps, make inquiries, and post and collect correspondence) have in many countries gone beyond the role of accessory to the postal service proper and pose problems of organization that in many ways are completely separate from those of the mail-handling services.

Postal services, because of the nature of the items or business handled, require extensive manpower. Many countries are striving to mechanize or even automate sorting, transport, and counter processes. Postal organization depends largely on geography; large agglomerations of population present the greatest problems. Furthermore, geography usually determines the routing of mails and the intermediate steps between the posting and destination offices. Traffic problems have forced postal managers, over the years, to develop their own approach to the operational aspects of postal organization. They developed their own work study methods and operational research techniques long before these terms were invented.

Problems of postal management

Apart from postal operations as such, postal management is concerned with the efficient administration and deployment of large bodies of manpower, the organization of large transport fleets, many aspects of property management, and financial and economic problems, particularly in an increasing number of countries where the postal services are expected to depend on their

postal revenue to meet all of their costs (including those related to capital expenditure). In this connection, the labour-intensive nature of postal work presents particular difficulties and is obliging some postal administrations to review traditional standards of performance (and charges) against a contemporary background of rapid and easy communications other than by means of letters.

HISTORY

Message-relay systems of the ancient world. Since good communications were clearly essential for governing the extensive empires of the ancient world, it is not surprising that the earliest historical references to postal systems should concern Egypt about 2000 BC and China under the Chou dynasty 1,000 years later. It was probably in China that a posthouse relay system was first developed and was brought to a high state of development under the Mongol emperors. The great Persian Empire of Cyrus in the 6th century BC also employed relays of mounted messengers, served by posthouses. The system was favourably described by the Greek historians Herodotus and Xenophon. The admiration of the Greeks was natural since their political divisions inhibited the growth of a coherent postal system, although each city-state possessed its corps of messengers.

The development of Rome from a small city-state into a vast empire embracing most of the known world brought with it the necessity for reliable and speedy communications with the governors of distant provinces. This need was met by the *cursus publicus*, the most highly developed postal system of the ancient world. The relay stages of the *cursus publicus*, established at convenient intervals along the great roads of the empire, formed an integral part of its complex military and administrative system. The speed with which messengers were able to travel during the peak of the administration was not to be rivalled in Europe until the 19th century: it has been claimed that over 170 miles could be covered in a day and a night. The maintenance of the *cursus publicus* required a high degree of organization; an inspectorial system existed to control its operation and prevent abuse for private ends.

The *cursus publicus*

The fall of the Roman Empire in the west during the 5th century did not completely destroy the *cursus publicus*. Its advantages were evident to the new barbarian rulers; some, such as Theodoric, king of the Ostrogoths, who ruled Italy from AD 493 to 526, are known to have maintained the essentials of the Roman postal system within their own domains. Even in the early 9th century, under the Carolingian Empire, the vestiges of the *cursus publicus* appear to have persisted, and posthouses were maintained. While the service did not follow a regularly organized pattern, it was at least reasonably frequent. The continued decay of the Roman roads, the increasing unwillingness of communities bordering the roads to support the expenses of the system, and the progressive political fragmentation of Europe, however, caused all traces of the Roman postal system to disappear.

The *cursus publicus* fared better in the Eastern Empire because its provinces were eventually absorbed into the Islāmic Empire. The substitution of one centralized imperial regime for another meant that the *cursus publicus* could be incorporated into a similar Arabian postal system based in Baghdad.

The pre-Columbian civilizations of America, responding to the same needs as the imperial states of Asia and Europe, also evolved relay systems, limited to foot messengers. In the Inca Empire, posthouses were maintained at frequent intervals along the remarkable road network, and a like system probably served the Mayan civilization for over 1,000 years.

Growth of business correspondence in the Middle Ages. The end of the Carolingian Empire (987) marked the beginning of two centuries of confusion in Europe, in which it is difficult to trace any postal system worthy of the title. Since the kings of the period were constantly struggling to assert their authority over their unruly feudal vassals, the strong central authority that sustained most postal systems was lacking.

The uncertain political situation did not favour the creation of a regulated postal service, though it necessitated frequent contact between the kings and vassals and among the great princes. They, along with other powerful institutions—the municipalities, the religious orders, and the universities (notably in Paris)—started to maintain corps of messengers to serve their particular needs.

One of the more significant trends of the later Middle Ages was the development of international commerce and, with it, the growth of business correspondence. Many corporations or guilds established messenger systems to allow their members to maintain contacts with customers. Notable among these was the so-called Butcher Post (Metzger Post), which was able to combine the carriage of letters with the constant travelling that the trade required.

The mercantile corporations of Italy, however, provided the most extensive and regular postal system of this period. Of particular importance were the links maintained from the mid-13th century between the great Italian commercial centres like Florence, Genoa, and Siena and six important annual fairs held in the Champagne area of northern France. Two fixed despatches were made to each of these fairs: the first to carry orders and commissions and the second to effect settlements. The service was carefully regulated. Conditions of acceptance, scales of payment, and timetables were laid down; the route was fixed, and hostels were maintained along the route. Since the Champagne Fairs were attended by merchants from all over Europe, the postal system provided a valuable international link.

Early postal links between Italy and France

Italian business interests were also responsible for the only regular extra-European postal link of this period, between Venice and Constantinople. The extent and importance of Venetian business correspondence may be gauged from the fact that in 1320 the King of Persia accorded its couriers the right of free passage throughout his domains.

Russia shared in the general European trend toward the development of postal services in the 13th century. Horses and drivers for the transport of couriers were kept at regular staging posts to provide the so-called carriage express, which gradually developed into an organized system for the exchange of letters.

Growth of the post as a government monopoly. Institutional postal systems that developed during the later Middle Ages also conveyed letters between private persons, with or without official sanction and for a substantial fee in either case. Initially, such letters were relatively few. Outside the institutions with their own postal services, the number of literate people having interests that ranged beyond their own neighbourhoods was small.

In the late 15th century, however, the trend toward improved postal services was reinforced by Gutenberg's printing press (*c.* 1450) and the expansion of education. The growth of demand made letter carrying a profitable business, leading to the rise of private undertakings—the majority, like the Swiss "Stumpelbotten," purely local in scope. Some, however, like the Paar family in Austria, developed postal organizations on a national scale. By far the most famous and extensive of such systems was that built up by the Taxis family, who originally came from Bergamo near Milan, Italy. Under the patronage of the Habsburg emperors, they became the organizers of an extensive network of postal routes linking the imperial possessions. Their system developed throughout the 16th century until it covered most of Europe, using 20,000 couriers to operate a relay system that was speedy, efficient, and highly profitable.

Although the remnants of the Taxis postal system survived in Germany up to 1867, it was essentially out of keeping—like the empire and the petty German states it served—with the main trend of development in Europe, the rise of nation-states with strong central governments. The first reflection of this trend in the postal sphere was the establishment of efficient national systems of relay posts under the control of the state. In France, Louis XI set up a Royal Postal Service in 1477 employing 230 mounted couriers. In England, a "Master of the Posts"

was appointed by Henry VIII in 1516 to maintain a regular postal service along the main roads radiating from London. Neither of these systems was comprehensive, nor were they intended to serve the public. The security and regularity of the service along certain routes, however, inevitably resulted in an increasing amount of unofficial correspondence being carried. After initial attempts to prevent this practice in France, its fiscal advantages were realized, and the carrying of private mails was legalized about 1600. The basis of a real public service was not created until 1627, when fees and timetables were fixed and post offices established in the larger cities. In Britain, a separate public service was set up in 1635 by a royal proclamation . . ."for the settling of the letter-office of England and Scotland." Thomas Witherings, a London merchant, was given the task of organizing regular services to run by day and night along the great post roads.

Legalizing private mails

In both countries, these state systems naturally began to develop into monopolies since such an evolution was seen by rulers as advantageous both to the security and to the revenues of the state. In England, the establishment of state posts along the principal roads was accompanied by the suppression, under the royal monopoly, of private and municipal posts, although "common carriers" were still permitted to convey letters on routes not covered by the royal system.

In France, postal services were declared a state monopoly in 1672, but private undertakings that had established legal rights in this field were allowed to continue. Most private messenger systems, however, were eventually forced out of business by state competition or were bought out. In 1719 the University of Paris, the most important private competitor, gave up its last postal privileges in return for substantial compensation.

There was, nevertheless, still opportunity for private enterprise to succeed by introducing services that were not at that time provided by the state systems. It was in this way that an important step in postal history, the establishment of local collection and delivery services in the great cities of London and Paris, was taken. London was the first city to benefit from an urban service when one William Dockwra set up his "Penny Post" in 1680. Striking features of the scheme were that letters were prepaid and stamped to indicate place of posting and the time they had been sent out for delivery. Deliveries were made almost hourly. Dockwra's scheme was so successful that he was prosecuted for infringing the state monopoly, and his service was closed down in November 1682, only to be reopened by the government. Not until 1760 was a similar local service introduced in Paris. It too was quickly absorbed by the state postal system; but its originator, Claude-Humbert Piarron de Chamousset, was paid compensation. Thus, the state monopolies expanded their scope, happily combining an improved service to the public with greater profitability.

The pace of postal progress in England during the later 18th century was accelerated by remarkable economic growth and a consequent demand for better mail services to the growing commercial and manufacturing centres. The most striking improvements came as the result of an extensive program of road building that began about 1765, paving the way for the era of the stagecoach. These were first used by the post office in 1784 and rapidly superseded the mounted postboys on the main routes. They began by averaging six or seven miles an hour, but continuing improvements in the roads and in the design of the vehicles pushed this up to ten miles an hour in the 1830s. With the stagecoaches making possible a general reorganization of the entire system of mail circulation, letters could be delivered the morning after posting in towns more than 120 miles (193 kilometres) from London. A carefully regulated postal service—unprecedented for its standards of speed, frequency, and security—was evolved during this period.

Use of the stagecoach

Despite the disruptive effects of the Revolutionary and Napoleonic wars, great progress was made throughout Europe in the late 18th and early 19th centuries in improving the speed and regularity of postal service and in providing internal delivery services for most of the larger cities. In the United States too, postal services expanded at a remarkable rate: in 1789 only 75 post offices existed, but 40 years later there were over 8,000.

Rowland Hill's reforms. The publication in 1837 of "*Post Office Reform: Its Importance and Practicability*," by Rowland Hill, a British educator and tax reformer, is justly regarded as one of the most important milestones in postal progress. Based on an exhaustive study of the cost structure of postal operations, it demonstrated conclusively that transit charges were an insignificant factor in the total cost of handling a letter. The then current intricate charging scales based on distance were shown to be irrelevant: they inflated operating costs by requiring a host of clerks to apply them and to prepare complicated interoffice accounts. He also realized that another major item in the current cost structure—the collection of money payments on delivery—was easily avoidable. Hill's solution was a uniform rate of postage, regardless of distance, and prepayment of postage by means of adhesive stamps that the sender could buy at a post office. Hill proposed a basic rate of one penny for each half ounce, calculating the "natural cost of distribution" to be slightly less than this. The cheapest current rate of postage was fourpence, and the average charge 6¼ pence (11.56 cents).

Not surprisingly, Hill's proposals rapidly gained strong support: popular agitation for the "penny post" overcame initial political disinterest, and the uniform rate and a system of prepayment by stamps were introduced in 1840. The originality of Hill's proposal for an adhesive postage stamp has been questioned but is irrelevant in considering the overall merits of his work. The significance of his reforms lies not only in the fact that they brought the post within the means of the mass of the people but also in the less obvious way in which they gave the postal system the technical capacity to deal with the vastly increased demand for postal service that ensued. The radical simplification of postal organization and methods characterizing Hill's reforms are the key to the speed and economy with which modern postal systems in many countries handle tens of millions of letters daily.

The chief features of Hill's system were gradually adopted in varying degrees by other countries throughout the world, first among which were Switzerland and Brazil in 1843.

The introduction of uniform cheap rates of postage for letters was accompanied by the establishment of even lower tariffs for newspapers (carried free in some countries) and for printed matter (*e.g.*, the British "Book Post" of 1848). These reduced rates were perhaps originally intended to favour the spread of education but quickly expanded, under the vigorous pressure of vested interests, to cover all sorts of commercial documents, advertising matter, magazines etc. An inexpensive form of correspondence, the postcard, first introduced by Austria in 1869, was soon adopted by most other countries.

The general postal reforms of the mid-19th century ensured maximum benefit from the technological progress in transport in the great age of the railway and the steamship. These new modes of conveyance permitted a far speedier, more regular, and more reliable mail service, both internally and internationally. Railways in particular had a marked effect on the organization of postal work: instead of merely using trains to carry mailbags more speedily, postal administrations soon introduced the practice of sorting the letters in transit, using specially adapted carriages. This greatly multiplied the advantages of railway conveyance. The first travelling, or railway, post offices ran in 1838 between Birmingham and Liverpool and London and Preston. By the end of the century, Britain, many European countries, the United States, and India had built up a complex network of such services, allowing the delivery of letters the day after posting at distances three or four times as great as had been possible with the stagecoach, exceeding 400 miles (643.74 kilometres) in some cases.

Effect of railways

International postal reform: the Universal Postal Union. The advent of the steamship and the railway had provided the opportunity for speedier international postal

services, and the expansion of commerce ensured a growing demand for such facilities. Unfortunately, serious obstacles to the free exchange of international mails existed. Postal relations between states were the subject of bilateral postal treaties that had multiplied alarmingly during the 19th century. Most large European states were party to at least a dozen of them by the 1860s. Such treaties necessitated the maintenance of detailed accounts between the countries concerned. Owing to the bewildering variety of currencies and units of weight and measurement then in use, the accounts attained a complexity described by a contemporary postmaster general of the United States as "almost beyond belief." Understandably, the users of the post suffered from this chaotic situation and from the high international postage rates that were its natural result.

The first practical step toward reform did not come, however, until May 1863, when the delegates of 15 European and American postal administrations met at the Paris Postal Conference, convening at the suggestion of the U.S. Postmaster General. The Conference established important general principles for the simplification of procedures, which were adopted as a model for subsequent bilateral treaties by the countries concerned.

The final step required the embodiment of these principles in a formal international treaty and the creation of an organization to administer them. An example was set by another conference at Paris two years later, which established the International Telegraph Union. Similar developments in the postal field were unfortunately delayed by the advent of the American Civil War and the Franco-Prussian War.

In 1868, however, a plan for a general postal union was put forward by the director of posts of the North German Confederation. Eventually, an international postal congress met, on September 15, 1874, in Bern. It was attended by the representatives of 22 states, all European except for Egypt and the U.S. On October 9 a "Treaty concerning the Establishment of a General Postal Union" was signed. It was implemented on July 1, 1875, when the General Postal Union came into being. This title was changed in 1878 to the Universal Postal Union (UPU), and the basic treaty was renamed the Universal Postal Convention. The treaty provides a uniform framework of rules and procedures for the exchange of international mails. The Union grew rapidly, increasing its membership to 55 within ten years. By 1914, when China was admitted, it included almost all independent countries.

Activities of the UPU

The scope of the Union's activities also expanded. In addition to its primary role, it gradually extended its functions to cover other international services provided by postal administrations, such as money orders (1878), parcel post (1885), postal checks (Giro; 1920), cash on delivery (1947), and savings bank (1957).

Development of airmail. The most important postal development of the 20th century has been the growth of airmail services. Certain experiments were undertaken before World War I, such as airmail service between Hendon and Windsor in 1911 to mark the coronation of George V and flights between Paris and Bordeaux in 1913. It was not until 1919, however, when the reliability of airplanes had considerably improved, that the first regular international service was introduced—between London and Paris—soon to be followed by other European links. It was over long-distance continental and intercontinental routes, however, that airmail was able to demonstrate its clear superiority over all surface transport. Technical factors delayed progress in opening up longer routes. Although the first American transcontinental airmail flight took place in 1920, a regular service did not begin until 1924. In 1926 a service between Egypt and Karāchi was started and was linked to London in 1929. It was extended to Singapore by 1933 and to Australia in 1934. It was not until 1939, however, that a regular service across the North Atlantic was established.

An important development of the late 1920s was the practice—started by a number of European countries—of forwarding mail by air without a surcharge. This practice had been extended to a number of intercontinental routes just before World War II (*e.g.*, between Britain and almost all parts of the British Empire). Since that war, however, the consistently high costs of airmail conveyance, which increase in direct proportion with the distance flown, have made such arrangements in the international service outside Europe very exceptional.

The aerogram

During World War II, a lightweight air-letter service for correspondence with military personnel abroad was introduced in Britain. Retained as an inexpensive form of airmail, it was widely adopted by other countries. In 1952 it was recognized by the UPU as the aerogram.

Since 1946, airmail services have developed rapidly to provide a network connecting Europe and America with all parts of the world. There are still areas, however, in which airmail services are relatively poor. African countries, for example, while generally enjoying a good airmail service to Europe, often lack direct links with each other.

Despite the immense improvements in postal service that it has made possible and its considerable potential for further development, airmail has not affected the basic organization of national postal systems as profoundly as did the railways. This is unlikely to be true of the other great trend of 20th-century postal development —mechanization and automation of the manual processes involved in handling mail. As this movement can hardly yet be considered part of the history of the post, it will be given separate treatment later in this article.

NATIONAL POSTAL SYSTEMS

The United States. Although the first official reference to overseas mail arrangements (concerning the receipt of overseas mail at Fairbanks' Tavern in Boston) dates back to 1639, little real progress was made in building up a postal system in colonial America until the appointment of Benjamin Franklin, formerly postmaster at Philadelphia, as deputy postmaster general for the American Colonies in 1753. Through diligent personal survey and inspection, he provided a more extensive, frequent, and speedier mail service, both within the Colonies and to England, thereby laying a sound foundation for the postal service in the U.S., of which he fittingly became first postmaster general in 1775.

Postal service expanded rapidly after independence: annual revenue increased from $37,935 in 1790 to $1,707,000 by 1829, when the postmaster general first became a member of the Cabinet. The heavy cost of establishing a postal structure to keep pace with the remarkable economic progress of the country and the accelerating extension of its settled area caused expenditure to rise even faster than revenue. The 1820s saw the beginning of the trend toward annual postal deficits, which commonly exceeded an annual figure of $5,000,000 by the later 19th century.

By 1901, however, this expenditure had produced remarkable results. The accessibility, quality, and range of services provided had improved immeasurably. The number of post offices stood at a peak of 76,945. Postage rates had been considerably reduced with the gradual adoption of Rowland Hill's principles: a single uniform rate regardless of distance was adopted in 1863, after an interim period with two rates since 1845. Postage stamps were introduced in 1847. Free collection services came with the provision of street letter boxes in 1858. A free delivery service was established in 1863, covering 49 cities and employing 440 letter carriers. By 1900 the service was provided at 796 offices by 15,322 carriers. The rural free delivery (RFD) service was introduced in 1896 and village delivery in 1912.

Growth of services

These delivery services have greatly expanded their scope: in 1969 the city and village delivery services were provided from 6,219 offices and employed 158,228 letter carriers. The rural delivery service was maintained by 30,945 carriers. About 80 percent of mail is now delivered by carriers, 18 percent through post office boxes, and only 2 percent at windows or counters. The total cost of carrier delivery services in 1969 was over $2,300,000,000.

The range of services available to the public has also grown steadily since the first supplementary postal service

—registered mail—was introduced in 1855. The major milestones in this progress were postal money order service (1864); international money orders (1867); special delivery (1885); parcel post, with its accessory collect on delivery (C.O.D.) and insurance services (1913); certified mail (1955, to provide proof of posting for items without intrinsic value). In 1911 a postal savings system was inaugurated, reaching a peak of more than 4,000,000 accounts in 1947. A decline to less than 1,000,000 depositors caused the service to be discontinued in 1966. Mail was formally divided into three classes in 1863, and a fourth was added in 1879. First-class, or letter, mail is the basis of the post office monopoly. The other classes were established to regulate the complex and often preferential rate structure applicable to publications, circulars, and samples. Second- and third-class mail is normally regarded as "subsidized," since it pays less than its fully allocated costs.

The post office has played a vital role as a pioneer and major user of all systems of transport as each was developed: the stagecoach, steamboat, canals, and railroads; the short-lived pony express; and, finally, the airlines. It also subsidized their development. A travelling post office system, introduced experimentally in 1862, made the railway mail service the dominant form of mail conveyance for the next century. The gradual reduction of passenger train services during the 1930s led to the birth of a highway post office service in 1941. Both of these services have declined rapidly, however, over the last five years. Railway post office mileage was reduced from 96,400,000 in 1965 to 10,100,000 in 1969, and the number of highway post offices in operation fell from 163 to 64. Conversely, ton-miles of airmail flown have increased from 188,103 to 323,662 over the same period, indicating a significant trend toward air transportation of regular mail without surcharge.

The U.S. Post Office now maintains what is by far the largest postal system in the world, handling almost half the world's volume of postal traffic: 84,800,000,000 pieces in 1970, over 400 items per head of population. Traffic is expected to top 120,000,000,000 by 1981.

To deal with the problem of increasing deficits and to improve the overall management and efficiency of the post office, Congress gave final approval to the Postal Reorganization Act of 1970 (HR 17070) on August 6, 1970. The act transformed the Post Office Department into a government-owned corporation, called the United States Postal Service. Congress no longer retains power to fix postal tariffs (although changes may be vetoed) or to control employees' salaries, and political patronage is virtually eliminated. Government subsidies continue but on a declining basis. The new corporation also has authority to raise capital to modernize its equipment and buildings.

The availability of adequate funds for its mechanization and automation program allowed the post office to benefit fully from its sustained effort in research and development. Already by 1969 over 60 percent of all letter mail was handled by preparation and sorting machines, a trend greatly assisted by the ZIP (Zone Improvement Plan) Code program, which by 1970 had achieved a usage of 84.8 percent.

Great Britain. The development of the British Post Office up to the reforms introduced by Rowland Hill has already been described. After 1840 the volume of postal traffic increased enormously, by 1870 reaching ten times its prereform level. The growth was fostered by the introduction of new facilities, such as registration and postcards, and of preferential rates for books, printed papers, and samples. Financial services were also expanded: a savings bank was established in 1861, and postal orders were introduced in 1881 to supplement the money order service taken over from private interests in 1838. In 1883 a parcel post service was established.

British Post Office social services

The sweeping social reforms of the 20th century have given the post office an additional role as the chief payment agency for social security benefits, beginning with old-age pensions in 1908 and expanding continuously to provide a variety of payments and also to collect large sums for state insurance schemes. The huge volume of these agency services is evident from statistics for 1968–69: over 800,000,000 social security payments were made, totalling over £2,600,000,000, and insurance stamps to the value of £1,244,000,000 were sold. The scale and range of financial transactions has been further boosted by the establishment of a postal check, or Giro, service in 1968, providing improved money transfer arrangements for the settlement of bills, and a current account banking system, which now includes loan facilities.

Another important 20th-century trend has been the gradual recognition of the post office's role as a commercial enterprise rather than as a government revenue department. The slow process of achieving full commercial status ended in October 1969, when the post office became a public corporation. The two great businesses of posts and telecommunications are now directed by a board with full executive authority. Overall political responsibility is vested in a new Ministry of Posts and Telecommunications. The National Savings Bank was transferred to the Treasury.

In the field of mail conveyance, Britain has not been forced to resort to extensive employment of air transport to maintain high-quality service. Although limited use of airmail is made in services between the South of England and Ireland and Scotland, the efficient and extensive network of railway travelling post offices remains the backbone of the circulation system, ensuring a high standard of next-day delivery service. To maintain the quality of this service economically by removing the need to handle nonpriority mail at peak periods, a controversial change to a "two-tier" system of letter classification was started in September 1968. The system abolished the complex preferential rate structure for printed papers and similar material—based on the contents of correspondence—and substituted the more relevant criterion of priority. The sender indicates the urgency of his item as being high (first tier) or lower (second tier) and pays on this basis, which simplifies the work of accepting the mail. Despite initial criticism, the new system has increased the percentage of priority mail given next-day delivery from 92 percent before the change to 94 percent in 1969. Another important measure to improve mail handling and facilitate mechanization has been the gradual introduction of an alphanumeric postal code, which has the unusual advantage of providing for inward sorting at the office of delivery.

France. The chief stages in the pre-Revolutionary development of postal services in France have already been described. With the Revolution, the system of farming or selling the right to operate a postal service was abolished, and the postal service was made a national agency in 1793. By an edict of June 16, 1801, the basic articles of the present postal monopoly were established.

First rural delivery service

In 1829 France became the first country to start a rural delivery service. A uniform internal postal tariff was established in August 1848, and postage stamps were introduced in the same year. Throughout the 19th and early 20th centuries, the scope and range of the services provided by the postal administration continually expanded: money orders (1817), registered letters (1829), postcards (1872), parcel post (1881), savings bank (1881), postal checks (1918). The parcels service is largely operated by railway authorities. France in 1970 had 17,500 post offices, in which a staff totalling 165,000 was employed to provide this extensive range of services.

The development of mail transport has been equally important over this period. A network of railway travelling post offices was built up during the latter 19th century, beginning with the Paris–Rouen service in 1844. Air conveyance was first employed experimentally in 1913 between Paris and Bordeaux, but a regular internal airmail service did not begin until 1933. Since World War II, the French postal administration has been among the pioneers in the field of internal airmail services, creating an autonomous postal fleet of 29 specialized mail transport aircraft to serve a comprehensive network of routes. In the early 1970s, the fleet was modernized. Since the services necessarily operate at night, turbojet aircraft

have been chosen in order to minimize noise. The total weight of mail flown has increased from 2,100,000 kilograms (4,629,700 pounds) in 1948 to 27,400,000 kilograms (60,407,000 pounds) during 1968, now representing over a quarter of the internal letter traffic.

The maintenance of this high standard of service in the face of constant growth in mail volume has posed serious problems. Daily traffic grew from 16,500,000 items in 1950 to 34,000,000 in 1969 (when the annual total exceeded 10,000,000,000). A study of probable future trends indicates that the daily figure will reach 41,500,000 by 1975 and 55,000,000 in 1985. To simplify the problems of handling such vast volumes of mail economically and speedily, the administration introduced a "two-tier" system in January 1969, allowing the customer to choose the priority to be accorded to his letter by paying an appropriate tariff.

West Germany. Postal organization in Germany remained on a relatively small scale until the later 19th century because of numerous sovereign states. The need for a more widely based postal system, to a certain degree, had been met since the 16th century, when the famous Taxis postal service was begun. The fragmented political state of Germany, in fact, allowed the Taxis organization to survive until 1867, when its last privileges were acquired by the postal service of the North German Confederation.

The unification of Germany under Prussia during the second half of the 19th century, culminating in the establishment of the German Empire in January 1871, was followed by the creation of a German postal service in the same year, under a law establishing a state monopoly for conveying letters and newspapers. In 1924 the postal administration was accorded a considerable degree of financial autonomy, allowing it to conduct business on semicommercial lines, taking due account of national economic and social factors in determining rate structure and pattern of service. The post office provides a complete range of normal postal and financial services, operates an extensive postal passenger transport network, and carries out various social security, revenue, and other agency work. In 1968 it maintained 26,333 post offices with a staff of over 275,000.

Establishment of the German postal service

Mail is transported chiefly by the railway, but there also exists an extensive complementary postal road network and an important night airmail service. The air service, started in 1961, carries letters and postcards on an unsurcharged basis.

Owing to labour shortages and the buoyancy of postal traffic, the West German Post Office has followed an energetic policy of mechanization.

Italy. Italy has a long history of organized postal systems, beginning with the *cursus publicus* of the Romans; but the present national postal system was not established until 1862, when the process of unifying the former small sovereign states into the new Kingdom of Italy had been successfully completed.

The new state postal service was given a monopoly for the collection, conveyance, and delivery of letters, printed papers, and newspapers, and a uniform tariff was established. The monopoly was surrendered for printed papers and newspapers in 1873 but was extended to cover parcels up to 20 kilograms (44 pounds) in 1923. A registration service also began in 1862, and postal orders were introduced, extending on a national scale a service that had been made available to the public in the former Kingdom of Sardinia as early as 1845. A Post Office Savings Bank was started in 1875 and now has over 10,000,000 accounts. The postal check service, which began in 1917, is also of considerable importance, carrying out almost 49,000,000 transactions in 1968. In addition to this range of financial services, the post office also acts as an agency for the payment of social security benefits such as state pensions and various other grants.

Railway conveyance for mails, authorized in 1865, has become the chief means of mail transport. Since 1965, however, the annual mileage on postal rail routes has declined, while the use of road transport and of air conveyance has been increasing. The rapid growth of airmail has resulted from the need to provide next-day delivery service between important cities like Milan and Palermo. Even the fastest surface transport takes over 36 hours for transits between north and south.

A program of mechanization and automation is being pursued, facilitated by the general cooperation of the public in using the postal routing code on letters and parcels.

Soviet Union. Since the establishment of the Soviet Union after the October Revolution of 1917, there has been a rapid development of postal services throughout Russia. In 1913 the country was served by only 7,618 official postal and telegraph establishments, four-fifths of which were concentrated in European Russia. In 1970 there were over 79,500 post offices, more than ten times the 1913 figure. The emphasis that has been placed on establishing an efficient postal structure in the Central Asian Republics is evident because postal establishments in these areas have increased over 30 or 40 times. It is also noteworthy that over 70 percent of the post offices are now to be found in rural areas, which were very badly served before 1917.

Central Asian postal service

Greatly improved postal facilities, in combination with the development of education and a high rate of economic growth, have fostered a remarkable increase in the volume of postal traffic. In 1913 the annual total of correspondence was 615,000,000 items; newspapers and magazines, 358,000,000; and parcels, 10,000,000. By 1970 letter post had risen to over 7,600,000,000, newspapers to 31,500,000,000, and parcels to 168,000,000.

To improve mail conveyance in such an extensive country—where the distance covered by postal routes is now almost 2,000,000 miles (3,218,690 kilometres)—greatly increased use of airmail is being made for the transport of both letter mail and newspapers. The amount airlifted has grown from 300,000 short tons (272,160 metric tons) in 1966 to over 350,000 short tons (317,510 metric tons) in 1969. There is also a road network covering over 500,000 miles (804,672 kilometres) and the fleet of postal vehicles exceeds 15,000.

An intensive mechanization program is being implemented, with particular emphasis on a high degree of automation and regulated mail flow at larger postal establishments. At smaller offices, appropriate mechanical aids for laborious and repetitive work are extensively employed. The money-order system has also been extensively automated, featuring a system of accounting machines and electronic computers to process and verify money order operations. Envelope and newspaper vending machines and automatic equipment for registered mail have been installed at offices serving the public.

China. The first use of a posthouse system was in China under the Chou dynasty (*c.* 1122–221 BC). A reference by Confucius in the late 6th century demonstrates that it was already renowned for its efficiency: "The influence of the righteous travels faster than a royal edict by post-station service."

Under the Han dynasty (206 BC–AD 220), posts were extended into conquered lands in Central Asia. Contacts were made with the Roman Empire, and Han historians observed that it maintained a postal system similar to that of China. It was not until the late 13th century, however, that Marco Polo revealed to Europe the scale and speed of the Chinese postal organization under the Mongol emperors. Nothing comparable to it existed in Europe at the time.

The system reached its peak under the Ch'ing dynasty (1644–1911). It was administered by the Post Station Department in Peking, and 16 regional directors supervised two complementary systems: a speedy mounted courier service along the five main routes of the empire, and an extensive network of foot runners. The mounted couriers regularly travelled up to 200 miles a day, occasionally attaining 270 miles on favourable terrain in North China. There were 1,600 posting stations, staffed by 70,000 men, using 40,000 horses. The annual cost in 1902 was estimated to be $3,000,000.

Postal system of the late Ch'ing dynasty

By this time, however, the Post Station Department was in decline; and in 1912 its functions were transferred to

an alternative state system, the Imperial Post, created in 1896. This service was organized on European lines. Unlike the Post Station system, the Imperial Post was not limited to official correspondence. It came into competition, therefore, with private letter-carrying companies, which had proliferated during the 19th century to serve the needs of the growing trading and business community for internal and overseas postal services. More than 300 such companies were registered in the treaty ports alone, with about 60 foreign post offices, maintained by Britain, the U.S., France, Germany, Japan, and Russia. The new state system gradually absorbed the business of the private companies, although the last was not closed until 1934. All foreign post offices had withdrawn by the end of 1922, following the establishment of efficient internal and international services by the Chinese Post Office, which joined the UPU in 1914.

Postal services were not greatly affected by the revolution that overthrew Imperial rule in 1911. Despite the incessant political disturbances that subsequently afflicted republican China, postal traffic and facilities expanded at a remarkable rate, as the Table demonstrates.

Postal Traffic and Facilities in Republican China

year	letter traffic (1,000,000)	post offices
1901	10.5	293
1911	117.0	957
1926	586.0	2,562
1946	1,039.0	3,162

Since the Communist occupation of mainland China in 1949, the postal service in Taiwan has grown rapidly. In 1969 it handled almost 500,000,000 items (compared with less than 30,000,000 in 1946); accordingly, extensive postal mechanization has been installed, including a complete mail-flow system in Taipei.

India. The earliest reference to an official postal system in India was made in the 14th century by the Arab traveller and historian Ibn Baṭṭūṭah, remarking upon the organized official service of mounted couriers and runners. The system was brought to its height during the 16th century under the great Mughal emperor Akbar, with a network of 2,000 miles of post roads.

Two centuries of political turmoil, without a strong central authority, destroyed this courier system. It was not until 1766 that an official post was re-established to serve a new ruling power, being made available for public use in 1774.

Establishment of the Indian Post Office

In 1837 the Imperial Post was established and granted a monopoly to provide efficient postal communications between the seat of government at Calcutta and the principal provincial towns. Within the provincial districts, a complementary local service was maintained. In 1854 the basis of the present Indian Post Office was established when these parallel systems were merged under a director general, and a uniform postage rate was introduced, a step of particular significance in so vast a country.

International postal relations developed rapidly after the establishment of a weekly steamer service between Bombay and England in 1867, with India becoming a member of the Universal Postal Union as early as 1876.

At the same time, the post office began to expand the range of services that it provided to the public: C.O.D. in 1877, an insured service in 1878, money orders in 1880, and savings bank in 1885.

Internal communications were improved to keep pace with the progress made in these other fields. The "bullock train"—the Indian equivalent of the mail coach—gave way to the growing network of railways. A regular travelling post office service was introduced in 1870, and the railway mail service still continues to carry the bulk of the mails, despite the development of airmail.

Although experiments with airmail conveyance started in 1911, a regular inland service did not begin until 1932. Rapid expansion followed in 1949. A complex night airmail network now connects 58 cities, carrying over 22 percent of the mail. Railways are likely to retain much of the traffic as a result of speedier train services.

More traditional forms of transport—the foot runner, horse, mule, camel, bullock cart, and bicycle—still help distribute mail to many of India's 567,338 villages. Over 50 percent of these now have a daily delivery of letters and over 90 percent enjoy at least a weekly delivery.

The number of post offices has been increased from 22,116 at independence (1947) to 105,433 in 1970.

Pakistan. Although Pakistan's existence as an independent sovereign state dates only from 1947, its postal history extends to the official postal systems established by Muslim emperors. Particularly noteworthy was the network of post relays at caravansaries (inns) established by the emperor Shēr Shāh Sūr in the early 16th century. Under British rule in 1852, the province of Sind had the distinction of being the first region in the subcontinent to adopt a uniform letter postage rate.

On becoming independent, Pakistan faced a particularly difficult problem of postal communications because its eastern and western regions were separated by 1,000 miles of Indian territory. The growth of civil aviation services and maritime links overcame the initial difficulty. In 1952 an airmail service was introduced between East and West Pakistan in which letters and postcards were carried without payment of any surcharge. This concessionary airmail system was extended to cover the internal mail service of both regions in 1959. The post office made use of the entire network of the national airline to connect 29 key points throughout the country. The service for printed paper and parcel mail was improved in 1962 by the introduction of direct sea-mail links between Karāchi and Chittagong.

Pakistan airmail service

Postal facilities were greatly expanded throughout the country, the number of post offices being increased from 7,550 in 1952 to a total of 14,200 in 1971. Door-to-door delivery of letters, parcels, and registered items by postmen has long been the standard practice in all towns and is almost universal in rural areas.

The post office is run on a commercial accounting basis but with strong emphasis on the public utility aspect. Postal traffic has grown rapidly. To improve its capacity to deal efficiently with the continuing growth, the postal administration has pursued a carefully considered and moderate policy of mechanization. At some of the busiest offices, machines designed to facilitate the acceptance of registered mail have been introduced. To speed up the handling of mail in sorting offices, electromechanical sorting machines were installed at Karāchi in West Pakistan and at Dacca (then in East Pakistan, now Bangladesh), in 1964 on an experimental basis. Another was installed at Lahore in 1966.

Japan. Although official and private systems of communications had existed in Japan from ancient times, it was not until 1870 that the creation of a comprehensive government-operated postal service was proposed. The idea was put forward by Hisoka Maejima, often called "the Father of the Post" in Japan. It was rapidly accepted by the government, which set up its own service between Tokyo and Osaka on April 20, 1871, extending it throughout the country in July 1872. In 1873 the postal service was proclaimed a monopoly and private courier systems were prohibited, a uniform postal tariff scale was adopted, and postage stamps and postcards were introduced. The first official overseas mail service was established in 1875 to the U.S. International postal relations expanded rapidly, and in 1877 Japan became a member of the Universal Postal Union.

Other landmarks were the introduction of a parcel post service in 1892, express delivery in 1911, airmail in 1929, and the achievement of a special accounting status on semicommercial lines in 1934.

A distinctive feature of the postal scene in Japan is the special New Year's Mail Service, introduced in 1900, which treats the vast number of New Year cards (now over 1,000,000,000) posted between December 15 and 31, separately from normal mail. They are retained at post offices so that all can be delivered together on January 1.

Postal traffic in Japan has been growing at a rate of over

5 percent per annum for several years; to keep pace, the department has greatly expanded the number of post offices; there is now a post office for every 5,000 inhabitants. Manpower expenses account for over 70 percent of the total costs of the postal service. Advanced statistical and work-measurement techniques are therefore employed in all sections of postal work to determine optimum staffing levels.

The shortage of manual labour and the growth in postal business resulting from the country's rapid economic development have also led to the adoption of an extensive mechanization policy. To facilitate this process, a postal code address system was introduced in July 1968. Automatic segregating or culling machines, facing and cancelling machines, and automatic postal code reader-sorters have been specially developed by Japanese firms for the postal department and have now been installed at major post offices throughout the country.

Japan's postal code address system

POSTAL SERVICES IN THE DEVELOPING COUNTRIES

The establishment of efficient and comprehensive postal systems in all developing countries is important on an international plane as well from the purely domestic viewpoint. International postal service depends for its successful maintenance and progressive improvement upon the effective cooperation of all member countries of the UPU.

The internal need for a good postal service is sufficient in itself, however, to justify a high priority. A country-wide network of post offices provides government with many points of contact with its people for implementing administrative programs in such fields as social security, taxation, and public information. When its facilities are properly used, the post office may also become one of the principal employers of labour in a country and help to promote economic growth through its need for buildings, vehicles, and equipment and as a major user of all transport services. The employment potential is evident from the fact that the percentage of the working population engaged in providing postal services is generally about five times as high in developed as in developing countries.

An efficient postal service, as well as promoting national cohesion, provides an essential infrastructure for the expansion of industry and commerce. Postal money transfer and savings services are particularly valuable in developing economies, where banking facilities are limited. In Ghana, for example, the postal savings bank balances are the equivalent of over 7 percent of the gross domestic product.

Postal administrations in developed countries have long appreciated the importance of collaborating in the improvement of postal services throughout the world. In fact, participation in postal technical assistance is an aim embodied in the constitution of the Universal Postal Union. In addition to fostering bilateral assistance between members, the UPU itself has, since 1964, taken part as a specialized agency in the United Nations Development Programme (UNDP). The Vienna Postal Congress of 1964 in particular gave a special impetus to UPU activities in this field.

In view of the rapid expansion of postal business in most developing countries, the training of staff is a most urgent need. The UPU's initial activity has, therefore, been largely devoted to this field: the needs of individual countries are evaluated by itinerant specialists employed by the International Bureau; expert aid is provided for the establishment of national training schools for postal workers and regional centres for middle-grade management staff, such as those set up in Abidjan, Bangkok, and Damascus. Special instructor-training courses are conducted in Britain and France, and seminars are held for higher grade officials, particularly in the developing regions. Specialist aid is also provided, however, to evaluate operational and organizational needs and subsequently to enlist UNDP assistance in meeting them. The missions of experts are thus directed either to the postal services in general or to a certain sector of postal activity. Some larger projects may include overall postal organization as well as postal training. An example is a large-scale project to establish a postal savings bank in Ethiopia and to strengthen the existing postal training school. Fellowships are awarded for overseas study of specialized subjects, such as international mails, philately, and, in particular, instructor training. Most fellowships are paid from UNDP resources, but many fellowships have also been financed by the UPU's own voluntary special fund. Since 1963 over $3,000,000 from multilateral sources has been spent on aid to some 80 recipient countries. Moreover, a considerable amount of aid is given directly by developed countries within the framework of their own bilateral technical assistance arrangements.

UPU's provision of specialist aid

In addition, attention may be called to the important role played by the Restricted Unions in regard to UPU technical assistance.

The necessity for technical cooperation on such an extensive and constantly growing scale is evident when worldwide disparities in standards of postal service are considered. On the average, there is one post office for every 7,000 of the world's population; but this figure cloaks a wide variation between European countries like Norway, Cyprus, and Portugal, with fewer than 1,000 people per post office, and Rwanda with over 300,000. There is a similar divergence in per capita postal traffic, in postal savings, and in other fields.

Many developing countries are not yet able to provide even the minimum scale of postal facilities warranted in their present circumstances. To underline the pressing need for improvement and to provide a yardstick for future progress, the UPU has adopted, for the second Development Decade, certain indicative planning objectives, among which are the following: (1) one post office, on average, to every 40 square kilometres (15 square miles) or 6,000 people; (2) widespread introduction of postal financial services; (3) next-day delivery of priority items within a 500-kilometre (310.68-mile) radius; (4) a progressive increase in the percentage of parcels conveyed by air in the international service. To draw particular attention to the dynamic function of postal services within the national infrastructure, the UPU has also published a valuable "Memorandum on the Role of the Post as a Factor in Economic and Social Development."

UPU planning objectives

THE INTERNATIONAL SYSTEM

International mail has a key role in furthering economic, social, and cultural links between nations. The international postal system is in itself an outstanding example of worldwide organization and mutual trust. A postal administration relies completely upon the postal authorities of other countries to play their parts in ensuring that its foreign mails reach their destination.

International cooperation in this field has been assured by the Universal Postal Union since 1875. It is now the most comprehensive of international organizations, with a membership of 143 comprising 136 sovereign states and seven groups of dependent territories (as of April 1, 1971). The few postal administrations that are not represented, notably that of mainland China, generally follow the rules of the UPU.

These fundamental rules of the international postal service are to be found in the Universal Postal Convention and have been little changed since they were set out in the treaty of Bern. The first basic principle is that all member countries form "a single postal territory for the reciprocal exchange of correspondence." From it is derived the principle of liberty of transit: every member country guaranteeing to respect the inviolability of transit mails, and to forward them by the most rapid transport used for its own mails.

Another important principle is the nonsharing of charges for letter-post items. Since 1875 each country has retained the postage it collects on international mail. Although intermediate countries are paid for transit service, the country in which the mail is delivered receives no payment. This principle was adopted in order to minimize the need for complex international accounts and was justified on the supposition that a letter normally generated a reply. Certain developing countries, however, have found themselves at a considerable disadvantage

under this rule, owing to an excessive imbalance between incoming and outgoing mail. To remedy this, the 1969 Congress of Tokyo has authorized payment to be made in such cases.

Provisions of the Universal Postal Convention

As a further measure of simplification, the convention fixes standard international postal rates, within a limited tolerance, and prescribes weight steps, limits of size, and conditions of acceptance for letter-post items. Disputes between postal administrations, which usually concern allocation of liability for the loss of registered or insured items, are to be settled by arbitration. The convention is completed by two other basic documents: (1) the Final Protocol, which allows member countries to register certain general and specific reservations to the provisions of the convention; and (2) the Detailed Regulations for Implementing the Universal Postal Convention. Apart from these obligatory documents, there are a number of optional agreements concerning supplementary services already mentioned earlier, such as parcel post and cash on delivery. The provision of a registration service is compulsory under the convention.

Mention must finally be made of the constitutive acts of the Union that prescribe its general aims, its organization, its financial structure, and the rules of membership, namely, the constitution and its general regulations.

This comprehensive framework of international regulations is regularly revised to take account of changing circumstances and technical advances. This is the chief function of the Union's quinquennial congress. Between congresses, the continuity of the Union's work is ensured by its elected Executive Council and its permanent office in Bern, the International Bureau. The latter acts as a clearing house for the settlement of international accounts and for the exchange of information between members, especially notifications of important operational and organizational changes. Problems arising in the technical, operational, and economic fields are studied by another permanent organ, the Consultative Committee for Postal Studies.

The UPU Constitution authorizes member countries to establish restricted unions, a provision that enables regional groups such as the Arab and the African Postal Union to conclude agreements aimed at improving postal services between their members by such means as reduced rates of postage or the elimination of transit fees. As such reforms are more easily achieved on a limited regional basis than on a worldwide scale, the restricted unions have a valuable role in the perennial task of the UPU, which is, basically, to improve international postal service by simplifying its organization and reducing its cost.

EMERGING POSTAL TECHNOLOGY

The post office and technological progress in transport. Postal administrations have been among the first to utilize new forms of transport. They have often applied considerable technical skill in maximizing the benefits to be derived from progress in this field, particularly in originating the travelling post office concept and apparatus enabling express trains to pick up and discharge mails without slackening speed. They have also developed their own transport systems to combat traffic congestion in certain busy cities, such as the pneumatic tube used in Paris, New York, and other cities and the automatic underground railway, opened in 1927, which links London's chief mail centres to railway termini.

Looking to the future, various postal administrations and private interests have conducted experiments in the use of rocket-propelled missiles as a means of mail transport. The rocket, however, cannot yet be considered a reliable and practical means of mail conveyance. In the same way, facsimile transmission, which eliminates physical conveyance altogether, is unlikely to prove a practicable general alternative to letters within the foreseeable future. Of more immediate significance for the future is a recent feasibility study into the bulk movement of all classes of mail by air.

Experiments with missilery

Automation of mail handling. During the last 15 years there has been a marked intensification of research and development effort toward applying technology to the handling of mails, especially in countries faced by manpower problems and higher labour costs. The wide variety of projects undertaken in many countries and the progress made have been summarized in several coordinated studies of different aspects of mechanization and automation by the UPU's Consultative Committee for Postal Studies (CCPS).

Actual implementation has generally been rather slower than expected. There have been good reasons for this. Primarily, most postal administrations, as government agencies, are subject to strict control of their capital investment programs. Secondly, mail traffic patterns—with marked peaks of work—make economic utilization of machines difficult: the introduction of measures to counteract this problem takes considerable time. Similarly, the introduction of postal-address codes and the standardization of sizes of envelopes and cards, which are prerequisites for mechanical handling, are relatively slow because of difficulties inherent in the change of procedures.

It is therefore still appropriate, so far as application is concerned, to consider postal technology as "emerging." The trend toward adopting public corporation status for the postal service should, however, alleviate the primary problem of a chronic shortage of capital for investment in comprehensive programs of postal automation.

Materials-handling equipment. Postal systems at present rely heavily on human labour for bulk materials handling and distribution, both at loading bays and between work processes within sorting centres. New mail centres, however, are normally built on factory lines to include all appropriate materials-handling equipment.

Equipment used for loading and unloading sacks of mail, rigid containers, and loose parcels includes mobile belt conveyors, roller conveyors, forklift trucks, mobile and fixed cranes, and table lifts. Handling equipment within buildings includes chain conveyors; horizontal and rising belt conveyors of all types, for the transport of loose letters, packets, and trays of letters (notably used for continuous clearance of public posting boxes); tow conveyors, which allow wheeled containers to be hooked on to a fixed-path underfloor traction system; bucket or pan elevators; and chutes and various other gravity-feed installations.

The use of a wide range of equipment is necessitated by the varied handling characteristics of different types of mail at particular stages. Buffer-storage facilities, in the form of ramps, hoppers, and moving belts, have to be incorporated to compensate for normal postal-traffic fluctuations. The smooth distribution of traffic through the system is often monitored by closed circuit television, which allows effective centralized control. Automatic regulation and recording, using a variety of sensing and counting devices linked to a computer, is the ideal. Modern systems-engineering techniques are thus able to ensure a carefully planned continuous mechanized mail flow with maximum productivity benefits.

Buffer-storage facilities

Segregating, or culling, machines. Mail collected from branch post offices and street letter boxes, although for the most part made up of ordinary letters and cards, also contains packets, newspapers, magazines, and large envelopes. These items, because of their size or shape, cannot be handled on machinery designed for the normal letter and have to be segregated from the majority of standard "machinable" letters. Owing to its varied characteristics, most packet mail has to be manually stamped and sorted, although its movement between work processes may be fully mechanized. So-called packet sorting machines are, in fact, essentially conveyor systems for distributing manually sorted mail.

A commonly adopted type of segregator consists of a laterally inclined rotating drum, into the upper end of which a regulated flow of "mixed" mail is fed from a storage conveyor. Letters within a thickness standard, but of excessive length or breadth, are picked out by various simple mechanical devices installed on the conveyor belt that eventually delivers machinable letters to the storage stacks of the facer–canceller equipment.

Facing and cancelling equipment. Facing is the process of aligning letters so that all will have the address side

facing the canceller, with stamps in a uniform position. The process is normally combined with a separation of the mail into at least two streams, letter and printed-paper rate or first and second class, to allow priority handling for one of the streams.

Facer–canceller machines perform these processes by passing letters through sensing or stamp-detecting units, which identify the presence or absence of a stamp on the side of the envelope facing them, and, when present, its position. Sensing units are also designed to separate mail in the priority class from nonpriority mail by identifying the stamp or commonly used combination of stamps representing the basic postage rate and manipulating selector gates accordingly. This identification is usually achieved by printing distinctive indexes on the stamps in normally invisible, phosphorescent or luminescent inks that are sensitive to ultraviolet radiation emitted by the sensing unit.

Coding and sorting machines. For manual sorting of letters, each operator normally uses a device with between 40 and 50 pigeonholes. This has been found by most administrations to be the optimum arrangement in view of the limited arm span and "memory" of the sorter. The development of various types of postal codes aimed at making the sorting of a coded letter a mechanical process for the operator by dispensing with the need to memorize a sorting plan. Unfortunately, public response to such coding schemes has not been enthusiastic.

For the past decade, therefore, many administrations have concentrated research on using an operator only to impress the postal code on each letter, employing phosphorescent or magnetic ink patterns that can be read by a sensing unit attached to a sorting machine. After the code has been impressed, the letter can be sorted at any subsequent stage by high-speed automatic machines, which are no longer utilized at the pace of a single operator and indeed can take the output of several operators. Furthermore, any second sortation required—even at an intermediate office or where the code includes the necessary information to postmen's routes at the delivery office—does not require further manual operations. Another potential advantage of this method is that letters may be directly encoded by the mail-processing machines used by large-volume mailers.

Optical character recognition. The ultimate aim in automated sorting is a machine that will be able to read certain elements of the address on letters. Research in this field is being carried on in the U.S., Great Britain, the U.S.S.R., Japan, Germany, France, Switzerland, Belgium, The Netherlands, and Australia. The immediate aims of these national research programs vary insofar as the type of character to be recognized is concerned: printed, typewritten, or addressographed characters; stylized handwritten scripts; and even ordinary handwriting. Also, some administrations require the machine to read a purely numeric code, others an alphanumeric code, and others the names of towns or regions. Several different techniques are used for the basic task of pattern matching in identifying the characters. For example, the observed character as a whole may be compared with matrices registered in the memory of the machine. Or the different traits of the character observed—vertical or horizontal strokes, curves, etc.—may be analyzed and their combination successively compared with a series of models registered by the computer.

A successful optical character reader (OCR) could be used either for the direct sorting of mail or for marking the mail with a machine-readable code so that sorting at subsequent stages could be carried out by high-speed automatic machines. The U.S. Post Office has had an alphanumeric optical character reader in operation on live mail since 1965. On the basis of field experience with this model and of technological advances in the computer industry since 1965, an improved Model II optical character reader is under development.

Numerical speech translator. One of the latest lines of research being pursued in the U.S. is the development of equipment that translates three-digit ZIP codes and sorting-code numbers spoken by an operator into instructions for a sorting machine. Since this system would obviate the need for a keyboard, it will leave the operator's hands free, making it particularly valuable in the operation of parcel and sack-sorting machines, where it will also eliminate the need for keyboard training of operators. The testing of the equipment includes determination of the effects of regional speech variations, background noise, and operator speech fatigue.

BIBLIOGRAPHY. Postal services are treated in official publications, such as legislative texts, manuals, working instructions, postal guides, annual reports, and commission reports; much of this material is available to the public. Many postal administrations publish periodical reviews or bulletins informing employees and users of recent events and trends. ROWLAND HILL, *Post Office Reform: Its Importance and Practicability*, 4th ed. (1838), is a work by an expert whose reforms affected subsequent worldwide postal development; G.B. HILL, *The Life of Sir Rowland Hill and the History of Penny Postage*, 2 vol. (1880), provides a fuller account of Sir Rowland's innovations. Detailed studies of specific postal service aspects are A.D. SMITH, *The Development of Rates of Postage* (1917); and A.D. LITTLE, *The Market for Postal Services* (1968). International mail services are treated in G.A. CODDING, *The Universal Postal Union* (1964); and L. ZILLIACUS, *Mail for the World* (1953), a popular treatment; both F. STAFF, *Transatlantic Mail* (1957), and H. ROBINSON, *Carrying British Mails Overseas* (1964), provide detailed treatment of the expansion of international mail conveyance.

General and historical literature includes: H. RACKOW *et al.*, *Handwörterbuch des Postwesens* (1953), a small but comprehensive encyclopaedia; *An Historical Summary of the Post Office* (HMSO 1911); I.G.J. HAMILTON, *An Outline of Postal History* (1910), including information on the history of postal service in India; and H. ROBINSON, *The British Post Office* (1948). Recent studies concerning the United States include: "Towards Postal Excellence," Kappel Commission Report (1968); C.H. SCHEELE, *A Short History of the Mail Service* (1970); and G. CULLINAN, *The Post Office Department* (1968); somewhat dated but still useful are: E.G. DANIEL, *United States Postal Service and Postal Policy* (1941); W. SMITH, *The History of the Post Office in British North America, 1639–1870* (1920); R.A. MCREYNOLDS, *History of the United States Postal Development, 1607–1931* (1937); A.E. SUMMERFIELD, *U.S. Mail: The Story of the U.S. Postal Service* (1960); and W.E. FULLER, *R.F.D.: The Changing Face of Rural America* (1964). The postal history of France is extensively described in E. VAILLE, *Histoire générale des postes françaises*, 6 vol. (1947–55), and *Histoire des postes françaises*, 2 vol. (1946–47). A current and anecdotal publication devoted to the postal system in the Federal Republic of Germany is S. ORTH, *Die Post* (1967). Postal services in developing countries are treated in UPU, *Memorandum on the Role of the Post as a Factor in Economic and Social Development* (1969).

The origins, development, and procedures of the Universal Postal Union are treated in G.A. CODDING, *The Universal Postal Union* (1964); see also the monthly review *Union Postale* (1875–), which is devoted to a wide range of questions concerning the UPU and various general aspects of postal activities.

H. SCHWAIGHOFER, *Postbetriebmechanik*, 2 vol. (1927), treats postal mechanization principles and practices; the *Collection of Postal Studies* (1950–), comprising some hundred UPU reports on postal service modernization, mechanization, economics, transport, and working methods, is available in French with much of the material also in English.

Pottery

Pottery in its widest sense includes all objects made from clay and hardened by fire: earthenware, stoneware, and porcelain. This article deals with pottery as art and craft, covering the kinds of pottery, the processes and techniques of forming and decorating it, and, finally, the history of pottery. Since the subject is so vast, the article will deal only with museum, or art, pottery. Artifacts used for purely utilitarian purposes will be discussed in articles such as AFRICAN PEOPLES, ARTS OF.

This article is divided into the following sections:

I. Kinds, processes, and techniques
  Kinds of pottery
  Forming processes and techniques
  Decorating processes and techniques
  Marking
II. Western pottery
  Ancient Near East and Egypt

Ancient Aegean and Greece
Etruscan and Roman
Islāmic
European: to the end of the 18th century
19th century
20th century
III. East Asian and Southeast Asian pottery
China
Korea
Japan
Thailand and Annam
IV. American Indian pottery
North America
Central America
South America

I. Kinds, processes, and techniques

KINDS OF POTTERY

Clay, the basic material of pottery, has two distinctive characteristics: it is plastic (*i.e.*, it can be molded and will retain the shape imposed upon it); and it hardens on firing to form a brittle but otherwise virtually indestructible material that is not attacked by any of the agents that corrode metals or organic materials. Firing also protects the clay body against the effects of water. If a sun-dried clay vessel is filled with water it will eventually collapse, but if it is heated, chemical changes that begin to take place at about 500° C (900° F) preclude a return to the plastic state no matter how much water is later in contact with it. Clay is refractory substance; it will vitrify only at temperatures of about 1,600° C (2,900° F). If it is mixed with a substance that will vitrify at a lower temperature (about 1,200° C, or 2,200° F) and the mixture is subjected to heat of this order, the clay will hold the object in shape while the other substance vitrifies. This forms a nonporous, opaque body known as stoneware. When feldspar or sooprock (steatite) is added to the clay and exposed to a temperature of from 1,200° C to 1,450° C (2,000° F to 2,650° F), the product becomes translucent and is known as porcelain. In this article, earthenware is used to denote all pottery substances that are not vitrified and are therefore slightly porous and coarser than vitrified materials.

The line of demarcation between the two classes of vitrified materials—stoneware and porcelain—is extremely vague. In the Western world, porcelain is usually defined as a translucent substance—when held to the light most porcelain does have this property—and stoneware is regarded as partially vitrified material that is not translucent. The Chinese, on the other hand, define porcelain as any ceramic material that will give a ringing tone when tapped. None of these definitions is completely satisfactory: for instance, some thinly potted stonewares are slightly translucent if they have been fired at a high temperature, whereas some heavily potted porcelains are opaque. Therefore the application of the terms is often a matter of personal preference and should be regarded as descriptive not definitive.

Earthenware. Earthenware was the first kind of pottery made, dating back about 9,000 years. In the 20th century, it is still widely used.

The earthenware body varies in colour from buff to dark red and from gray to black (see below *Firing*). The body can be covered or decorated with slip (a mixture of clay and water in a creamlike consistency, used for adhesive and casting as well as for decoration), with a clear glaze, or with an opaque tin glaze. Tin-glazed earthenware is usually called maiolica, faience, or delft (see below *Decorative glazing*). If the clear-glazed earthenware body is a cream colour, it is called creamware. Much of the commercial earthenware produced in the second half of the 20th century is heat- and cold-proof and can thus be used for cooking and freezing as well as for serving.

Stoneware. Stoneware is very hard and, although sometimes translucent, usually opaque. The colour of the body varies considerably; it can be red, brown, gray, white, or black.

Fine white stoneware was made in China as early as 1400 BC (Shang dynasty). In Korea, stoneware was first made during the Silla dynasty (57 BC–AD 935); in Japan, during the 13th century (Kamakura period). The first production of stoneware in Europe was in 16th-century Germany. When tea was first imported to Europe from China in the 17th century, each chest was accompanied by a red stoneware pot made at the I-hsing kilns in Kiangsu province. This ware was copied in Germany, the Netherlands, and England. At the end of the 17th century, English potters made a salt-glazed white stoneware that was regarded by them as a substitute for porcelain (see below *Decorative glazing*). In the 18th century, the Englishman Josiah Wedgwood made a black stoneware called basaltes and a white stoneware (coloured with metallic oxides) called jasper. A fine white stoneware, called Ironstone china, was introduced in England early in the 19th century. In the 20th century, stoneware is used mostly by artist-potters, such as Bernard Leach and his followers.

First porcelain

Porcelain. Porcelain was first made in a primitive form in China during the T'ang dynasty (AD 618–907). The kind most familiar in the West was not manufactured until the Yüan dynasty (AD 1279–1368). It was made from kaolin (white china clay) and petuntse (a feldspathic rock), the latter being ground to powder and mixed with the clay. During the firing, which took place at a temperature of about 1,450° C (2,650° F), the petuntse vitrified, while the refractory clay ensured that the vessel retained its shape.

In medieval times isolated specimens of Chinese porcelain found their way to Europe, where they were much prized, principally because of their translucency. European potters made numerous attempts to imitate them, and, since at that time there was no exact body of chemical and physical knowledge whereby the porcelain could be analyzed and then synthesized, experiments proceeded strictly by analogy. The only manufactured translucent substance then known was glass, and it was perhaps inevitable that glass made opaque with tin oxide (the German *Milchglas*, or milk glass, for example) should have been used as a substitute for porcelain. The nature of glass, however, made it impossible to shape it by any of the means used by the potter, and a mixture of clay and ground glass was eventually tried. Porcelain made in this way resembles that of the Chinese only superficially and is always termed soft, or artificial, porcelain. The date and place of the first attempt to make soft porcelain are debatable, but some Middle Eastern pottery of the 12th century was made from glaze material mixed with clay and is occasionally translucent (see below *Islāmic pottery: Egypt*). Much the same formula was employed with a measure of success in Florence about 1575 at workshops under the patronage of Duke Francesco de'Medici. No further attempts of any kind appear to have been made until the mid-17th century, when Claude and François Révérend, Paris importers of Dutch pottery, were granted a monopoly of porcelain manufacture in France. It is not known whether they succeeded in making it or not, but, certainly by the end of the 17th century, porcelain was being made in quantity, this time by a factory at Saint-Cloud, near Paris.

The secret of true, or hard, porcelain similar to that of China was not discovered until about 1707 in Saxony, when Ehrenfried Walter von Tschirnhaus, assisted by an alchemist called Johann Friedrich Böttger, substituted ground feldspathic rock for the ground glass in the soft porcelain formula. Soft porcelain, always regarded as a substitute for hard porcelain, was progressively discontinued because it was uneconomic; kiln wastage was excessive, occasionally rising to nine-tenths of the total.

Hard and soft porcelain

The terms soft and hard porcelain refer to the soft firing (about 1,200° C, or 2,200° F) necessary for the first, and the hard firing (about 1,450° C, or 2,650° F) necessary for the second. By coincidence they apply also to the physical properties of the two substances: for example, soft porcelain can be cut with a file, whereas hard porcelain cannot. This is sometimes used as a test for the nature of the body.

In the course of experiments in England during the 18th century, a type of soft porcelain was made in which bone ash (a calcium phosphate made by roasting the bones of cattle and grinding them to a fine powder) was added to the ground glass. Josiah Spode the Second later added this

bone ash to the true, hard porcelain formula, and the resulting body, known as bone china, has since become the standard English porcelain. Hard porcelain is strong, but its vitreous nature causes it to chip fairly easily and, unless especially treated, it is usually tinged slightly with blue or gray. Bone china is slightly easier to manufacture. It is strong, does not chip easily, and the bone ash confers an ivory-white appearance widely regarded as desirable. Generally, bone china is most popular for table services in England and the United States, while hard porcelain is preferred on the European continent.

FORMING PROCESSES AND TECHNIQUES

Preparation of material. Raw clay consists primarily of true clay particles and undecomposed feldspar mixed with other components of the igneous rocks from which it was derived, usually appreciable quantities of quartz and small quantities of mica, iron oxides, and other substances. The composition and thus the behaviour and plasticity of clays from different sources are therefore slightly different. Except for coarse earthenwares, which can be made from clay as it is found in the earth, pottery is made from special clays plus other materials mixed to achieve the desired results. The mixture is called the clay body, or batch.

To prepare the batch, the ingredients are combined with water and reduced to the desired degree of fineness. The surplus water is then removed.

Shaping the clay. The earliest vessels were modelled by hand, using the finger and thumb, a method employed still by the Japanese to make *raku* teabowls. Flat slabs of clay luted together (using clay slip as an adhesive) were employed to make square or oblong vessels, and the slabs could be formed into a cylinder and provided with a flat base by the same means. Coiled pottery was an early development. Long rolls of clay were coiled in a circle, layer upon layer, until the approximate shape had been attained; the walls of the vessel were then finished by scraping and smoothing. Some remarkably fine early pots were made in this way.

Introduction of the potter's wheel

It is impossible to say when the potter's wheel, which is a difficult tool and needs long apprenticeship, was introduced. A pot cannot be made by hand modelling or coiling without the potter either turning it or moving around it, and, as turning involves the least expenditure of human effort, it would obviously be preferred. The development of the slow, or hand-turned, wheel as an adjunct to pottery manufacture led eventually to the introduction of the kick wheel, rotated by foot, which became the potter's principal tool. The potter throws the clay onto a rapidly rotating disc and shapes his pot by manipulating it with both hands. This is a considerable feat of manual dexterity that leads to much greater exactness and symmetry of form. Perhaps the most skillful of all potters have been the Chinese. Excellent examples of their virtuosity are the double-gourd vases, made from the 16th century onward, which were turned in separate sections and afterward joined together. By the 18th century the wheel was no longer turned by the potter's foot but by small boys, and since the 19th century the motive power has been mechanical.

Jollying, or jiggering, is the mechanical adaptation of wheel throwing and is used where mass production or duplication of the same shape—particularly cups and plates—is required. The jolly, or jigger, was introduced during the 18th century. It is similar to the wheel in appearance except that the head consists of a plaster mold shaped like the inside of an object, such as a plate. As it revolves, the interior of the plate is shaped by pressing the clay against the head, while the exterior, including the footring, is shaped by a profile (a flat piece of metal cut to the contour of the underside of the plate) brought into contact with the clay. Machines that make both cups and plates automatically on this principle were introduced in the 20th century. Small parts, such as cup handles, are made separately by pressing clay into molds and are subsequently attached to the vessel by luting.

Molding

One of the earliest methods of shaping clay was molding. Pots were made by smearing clay around the inside of a basket or coarsely woven sack. The matrix was consumed during firing, leaving the finished pot with the impression of the weave on the exterior. A more advanced method, used by the Greeks and others, is to press the pottery body into molds of fired clay. Though the early molds were comparatively simple, they later became more complex, a tendency best seen in those molds used for the manufacture of pottery figures. The unglazed earthenware figures of Tanagra (Boeotia, central Greece) were first modelled by hand, then molds of whole figures were used, and finally the components—arms, legs, heads, and torsos—were all molded separately. The parts were often regarded as interchangeable, so that a variety of models could be constructed from a limited number of components. No improvement on this method of manufacture had been devised by the 20th century: the European porcelain factories make their figures in precisely the same way.

Plaster of paris molds were introduced into Staffordshire about 1745. They enabled vessels to be cast in slip, for when the slip was poured into the mold the plaster absorbed the water from it, thus leaving a layer of clay on the surface of the mold. When this layer had reached a sufficient strength and thickness, the surplus slip was poured off, the cast removed and fired, and the mold used again. This method is still in common use.

Drying, turning, and firing. Newly shaped articles were formerly allowed to dry slowly in the atmosphere. In 20th century pottery factories, this stage is speeded up by the introduction of automatic dryers, often in the form of hot, dry tunnels through which the ware passes on a conveyor belt.

Turning is the process of finishing the greenware (unfired ware) after it has dried to leather hardness. The technique is used to smooth and finish footrings on wheel-thrown wares or undercut places on molded or jiggered pieces. It is usually done on the potter's wheel or jigger as the ware revolves. Lathe turning, like most hand operations, was tending to disappear in the mid-20th century except on the more ornamental and expensive objects.

The firing process

The earliest vessels, which were sun-dried but not fired, could be used only for storing cereals and similar dry materials. If a sun-dried clay vessel is filled with water it absorbs the liquid, becomes very soft, and eventually collapses; but if it is heated, chemical changes that begin to take place at about 500° C (900° F) preclude a return to the plastic state.

After thorough drying, the pottery is fired in a kiln. In primitive pottery making, the objects were simply stacked in a shallow depression or hole in the ground, and a pyre of wood was built over them. Later, coal- or wood-fired ovens became almost universal. In the 20th century both gas and electricity are used as fuels. Many improvements have been made in the design of intermittent kilns, in which the ware is stacked when cold and then raised to the desired temperature. These kilns are extravagant of fuel, however, and are awkward to fill or empty if they have not had time to cool completely. For these reasons they are being replaced by continuous kilns, the most economical and successful of which is the tunnel kiln. The wares are conveyed slowly from a comparatively cool region at the entrance to the full heat in the centre. As they near the exit after firing, they cool gradually.

The atmosphere in the kiln at the time of firing, as well as the composition of the clay body, determines the colour of the fired earthenware pot. Iron is ubiquitous in earthenware clay, and under the usual firing conditions it oxidizes, giving a colour ranging from buff to dark red according to the amount present. In a reducing atmosphere (*i.e.*, one where a limited supply of air causes the presence of carbon monoxide) the iron gives a colour varying from gray to black, although a dark colour may also occur as a result of the action of smoke. Both of the colours that result from iron in the clay can be seen in the black-topped vases of predynastic Egypt.

DECORATING PROCESSES AND TECHNIQUES

Impressing and stamping. Even the earliest pottery was usually embellished in one way or another. One of

the earliest methods of decoration was impressing the raw clay. Finger marks were sometimes used, as well as impressions from rope (Japanese Jōmon ware of the 1st millennium BC) or from a beater bound with straw (used to shape the pot in conjunction with a pad held inside it). Basketwork patterns are found on pots molded over baskets and are sometimes imitated on pots made by other methods.

Relief decoration

The addition of separately modelled decoration, known as applied ornament (or appliqué), such as knops (ornamental knobs) or the reliefs on Wedgwood jasperware, came somewhat later. The earliest known examples are found on Mediterranean pottery made at the beginning of the 1st millennium. Raised designs are also produced by pressing out the wall of the vessel from inside, as in the Roman pottery known as terra sigillata, a technique that resembles the repoussé method adopted by metalworkers. Relief ornament was also executed—by the Etruscans, for example—by rolling a cylinder with the design recessed in intaglio over the soft clay, the principle being the same as that used to make Babylonian cylinder seals.

Incising, sgraffito, carving, and piercing. The most primitive kind of decoration is that which is incised into the raw clay with a pointed stick or with the thumbnail, chevrons (inverted v's) being a particularly common motif. Incised designs on a dark body are sometimes filled with lime, which effectively accents the decoration. Examples can be seen in some early work from Cyprus and in some comparatively modern work from primitive tribes. Decoration that has been engraved after firing is much less usual, but the skillful and accomplished engraving on one fine Egyptian pot of the predynastic period (*i.e.*, before *c.* 3100 BC) suggests that the practice may have been more frequent than was previously suspected.

Originally, defects of body colour suggested the use of slip, either white or coloured, as a wash over the vessel before firing. A common mode of decoration is to incise a pattern through the slip, revealing the differently coloured body beneath, a technique called sgraffito ("scratched"). Sgraffito ware was produced by Islāmic potters and became common throughout the Middle East. The 18th-century scratched-blue class of English white stoneware is decorated with sgraffito patterns usually touched with blue.

Sgraffito technique

Related to the sgraffito technique is slip carving: the clay body is covered with a thick coating of slip, which is carved out with a knife, leaving a raised design in slip (champlevé technique). Slip carving was done by Islāmic and Chinese potters (Sung dynasty).

Much pierced work—executed by piercing the thrown pot before firing—was done in China during the Ming dynasty (reign of Wan-li). It was sometimes called "demon's work" (*kuei kung*) because of the almost supernatural skill it was supposed to require. English white molded stoneware of the 18th century also has elaborate piercing.

Slip decorating. In addition to sgraffito and carving (see above *Impressing and stamping*), slip can be used for painting, trailing, combining, and inlay. The earliest forms of decoration in ancient Egypt, for example, were animal and scenic motifs painted in white slip on a red body; and in the North American Indian cultures coloured slips provided the material for much of the painted freehand decoration.

Slip, too, is sometimes dotted and trailed in much the same way as a confectioner decorates a cake with icing sugar; the English slipwares of the 17th and 18th centuries are typical of this kind of work. Earthenware washed over with a white slip and covered with a colourless glaze is sometimes difficult to distinguish from ware covered with a tin glaze (see below *Decorative glazing*). In consequence it has sometimes been wrongly called faience. The term for French earthenware covered with a transparent glaze (in imitation of Wedgwood's creamware) is *faience fine*, and in Germany it is called *Steingut*. *Mezza-Maiolica* (Italy) and *Halb fayence* (Germany) refer to slip-covered earthenware with incised decoration.

Slip is also used for combed wares. The marbled effect on Chinese pottery of the T'ang dynasty, for example, was sometimes achieved by mingling, with a comb, slips of contrasting colours after they had been put on the pot.

Combed wares and inlay

The Koreans used slip for their inlay technique called *mishima*. The designs were first incised into the clay, and the incisions were then filled with black and white slip.

Burnishing and polishing. When the clay used in early pottery was exceptionally fine, it was sometimes polished or burnished after firing. Such pottery—dating back to 6500 and 2000 BC—has been excavated in Turkey and the Pan-shan cemetery in Manchuria. Most Inca pottery is red polished ware.

Decorative glazing. Early fired earthenware vessels held water, but, because they were still slightly porous, the liquid percolated slowly to the outside where it evaporated, cooling the contents of the vessel. Thus, the porosity of earthenware was, and still is, sometimes an advantage in hot countries, and the principle is utilized in the 20th century in the construction of domestic milk and butter coolers and some food-storage cupboards.

Porosity, however, had many disadvantages—*e.g.*, the vessels could not be used for storing wine or milk. To overcome the porosity, primitive peoples often applied varnishes of one kind or another. Varnished pots were made, for example, in Fiji. The more advanced technique is glazing. The fired object was covered with a finely ground glass powder often suspended in water and was then fired again. During the firing the fine particles covering the surface fused into an amorphous, glasslike layer, sealing the pores of the clay body.

The art of glazing earthenware for decorative as well as practical purposes followed speedily upon its introduction. On stoneware, hard porcelain, and some soft porcelain, which are fired to the point of vitrification and are therefore nonporous, glazing is used solely for decoration.

Except for tin-glazed wares (see below *Painting*), earthenware glaze was added to the biscuit clay body, which was then fired a second time at a lower temperature. Soft porcelain glaze was always applied in this way. Hard porcelain glaze was usually (and stoneware salt glaze, always) fired at the same time as the raw clay body at the same high temperature.

Kinds of glazes

Basically, there are four principal kinds of glazes: feldspathic, lead, tin, and salt. (Modern technology has produced new glazes that fall into none of these categories while remaining a type of glass.) Feldspathic, lead, and salt glazes are transparent; tin glaze is an opaque white. Hard porcelain takes a feldspathic glaze, soft porcelain usually a kind of lead glaze and can be classified according to the kind of glaze used.

There are two main types of glazed earthenware: the one is covered with a transparent lead glaze, and the other with an opaque white tin glaze.

Tin glaze was no doubt employed in the first place to hide faults of colour in the body, for most clays contain a variable amount of iron that colours the body from buff to dark red. Tin-glazed wares look somewhat as though they have been covered with thick white paint. These wares are often referred to as "tin-enamelled." As noted above, other terms in common use are maiolica, faience, and delft. Unfortunately, these are variously defined by various authorities. The art of tin-glazing was discovered by the Assyrians, who used it to cover courses of decorated brickwork. It was revived in Mesopotamia about the 9th century AD and spread to Moorish Spain, whence it was conveyed to Italy by way of the island of Majorca, or Maiolica. In Italy, tin-glazed earthenware was called maiolica after the place where it was mistakenly thought to have originated. The wares of Italy, particularly those of Faenza, were much prized abroad, and early in the 16th century the technique was imitated in southern France. The term faience, which is applied to French tin-glazed ware, is undoubtedly derived from Faenza. Wares made in Germany, Spain, and Scandinavia are known by the same name. Early in the 17th century a flourishing industry for the manufacture of tin-glazed ware was established at the town of Delft, Holland, and Dutch potters brought the art of tin-glazing to England

together with the name of delft, which now applies to ware manufactured in The Netherlands and England. Some misleading uses of these terms include that of applying maiolica to wares made outside Italy but in the Italian style, and faience to Egyptian blue-glazed ware and certain kinds of Near Eastern earthenware.

Although glazed stoneware does not fall into such definite categories as glazed earthenware, to some extent it can be classified according to the kind of glaze used. The fine Chinese stonewares of the Sung dynasty (AD 960–1279) were covered with a glaze made from feldspar, the same vitrifiable material later used in both the body and glaze of porcelain. Stoneware covered with a lead glaze is sometimes seen, but perhaps the majority of extant glazed wares are salt-glazed. In this process a shovelful of common salt (sodium chloride) is thrown into the kiln when the temperature reaches its maximum. The salt splits into its components, the sodium combining with the silica in the clay to form a smear glaze of sodium silicate, the chlorine escaping through the kiln chimney. Salt glazes have a pitted appearance similar to that of orange peel. A little red lead is sometimes added to the salt, which gives the surface an appearance of being glazed by the more usual means.

Crazed and coloured glaze

Some fusion usually occurs between glaze and body, and it is therefore essential that both should shrink by the same proportion and at the same rate on cooling. If there is a discrepancy, the glaze will either develop a network of fine cracks or will peel off altogether. This crazing of the glaze was sometimes deliberately induced as a decorative device by the Chinese.

One method of applying colour to pottery is to add colouring oxides to the glaze itself. Coloured glazes have been widely used on earthenware, stoneware, and porcelain and have led to the development of special techniques in which patterns were incised, or outlined with clay threads (cloisonné technique), so that differently coloured glazes could be used in the same design without intermingling; for example, in the *lakabi* wares of the Middle East.

Unglazed wares

Earthenware, stoneware, and porcelain are all found in unglazed as well as glazed forms. Wares fired without a glaze are called biscuit. Early earthenware pottery, as discussed above, was unglazed and therefore slightly porous. Of the unglazed stonewares, the most familiar are the Chinese Ming dynasty teapots and similar wares from I-hsing in Kiangsu Province, the red stoneware body made at Meissen in Saxony during the first three decades of the 18th century and revived in modern times, and the ornamental basaltes and jaspers made by Josiah Wedgwood and Sons since the 18th century. Biscuit porcelain was introduced in Europe in the 18th century. It was largely confined to figures, most of which were made at the French factories of Vincennes and Sèvres. Unglazed porcelain must be perfect, for the flaws cannot be concealed with glaze or enamel. The fashion for porcelain biscuit was revived in the 19th century and called Parian ware.

Painting. Painted designs are an early development, some remarkably fine work made before 3000 BC coming from excavations at Ur and elsewhere in Mesopotamia, as well as urns from Pan-shan in Manchuria (Northeast Provinces) that date back to 2000 BC.

The earliest pottery colours appear to have been achieved by using slips stained with various metallic oxides (see above *Slip decoration*). At first these were undoubtedly oxides that occurred naturally in the clay; later they were added from other sources. Until the 19th century, when pottery colours began to be manufactured on an industrial scale, the oxides commonly used were those of tin, cobalt, copper, iron, manganese, and antimony. Tin oxide supplied a useful white, which was also used in making tin glaze (see above *Decorative glazing*) and occasionally for painting. Cobalt blue, ranging in colour from grayish blue to pure sapphire, was widely used in the Far East and Europe for blue-and-white porcelain wares. Cupric oxide gives a distinctive series of blues, cuprous oxide a series of greens, and, in the presence of an excess of carbon monoxide (which the Chinese achieved by throwing wet wood into the kiln), cupric oxide yields a bluish red. This particular colour is known as reduced copper, and the kiln is said to have a reducing atmosphere. (For the effect of this atmosphere on the colour of the biscuit body, see above *Firing*). The colours obtained from ferric iron range from pale yellow to black, the most important being a slightly orange red, referred to as iron red. Ferrous iron yields a green that can be seen at its best on Chinese celadon wares. Manganese gives colours varying from the bright red purple similar to permanganate of potash to a dark purplish brown that can be almost black. The aubergine purple of the Chinese was derived from this oxide. Antimony provides an excellent yellow.

Overglaze and underglaze colours

Pottery colours are used in two ways—under the glaze or over it. Overglaze painting is executed on a fired clay body covered with a fired glaze, underglaze painting, on a fired, unglazed body (which includes a body that has been coated with raw or unfired, glaze material).

Earthenware and stoneware are usually decorated with underglaze colours. After the body is manipulated into the desired shape it is fired. It is then painted, coated with glaze, and fired again. The second firing is at a lower temperature than the first, being just sufficient to fuse the glaze. In the case of most tin-glazed wares the fired object was first coated with the tin glaze, then painted, then fired again. The painting needed exceptional skill, since it was executed on the raw glaze and erasures were impossible. The addition of a transparent lead glaze over the painted decoration needed a third firing. In 18th-century Germany especially tin-glazed wares were decorated with colours applied over the fired glaze, as on porcelain. The wares were sometimes called *Fayence-Porcellaine*.

The body and glaze of most hard porcelain are fired in one operation, since the fusion temperature of body and glaze is roughly the same. Underglaze colours are limited because they must be fired at the same temperature as the body and glaze, which is so high that many colours would "fire away" and disappear. Although the Chinese made some use of copper red, underglaze painting on porcelain is more or less limited to cobalt blue, an extremely stable and reliable colour that yields satisfactory results under both high- and low-temperature firings. On soft porcelain, manganese was sometimes used under the glaze, but examples are rare. All other porcelain colours were painted over the fired glaze and fixed by a second firing that is much lower than the first.

Underglaze pigments are known as high-temperature colours, or colours of the grand feu. Similarly, overglaze colours are known as low-temperature colours, or colours of the petit feu. Other terms for overglaze colours are enamel colours and muffle colours, the latter name being derived from the type of kiln, known as a muffle kiln, in which they are fired. Overglaze colours consist of pigments mixed with glaze material suspended in a medium, such as gum arabic, with an alkaline flux added to lower the melting point below that of the glaze. They were first used in Persia on earthenware (*minai* painting) in the 12th century and perhaps at the same date on Chinese stoneware made at Tz'u-chou.

Lustre painting, gilding, and silvering

Lustre decoration is carried out by applying a colloid suspension of finely powdered gold, silver, platinum, or copper to the glazed and fired object. On a further, gentle firing, gold yields a purplish colour, silver a pale straw colour, platinum retains its natural hue, and copper varies from lemonish yellow to gold and rich brown. Lustre painting was invented by early Islāmic potters.

Pottery may be gilded or silvered. The earliest gilding was done with gold mixed with an oil base. The use of gold ground in honey may be seen on the finest porcelain from Sèvres during the 18th century, as well as on that from Chelsea. Toward the end of the same century gold was applied as an amalgam, the mercury subsequently being volatilized by heating. Silver was used occasionally for the same purposes as gold but with time has nearly always turned black through oxidation.

Transfer printing. The transfer print made from a copper plate was first used in England in the 18th century.

In the 20th century transfers from copper plates are in common use for commercial wares, as are lithographic and other processes, such as silk screen printing, which consists of rubbing the colour through a patterned screen of textile material. Combinations of hand-painted and transfer decorations are often used. The outline or other part of the decoration is put on with a transfer print, then parts of the design, such as leaves, flowers, clothing, or water, are painted in.

MARKING

Most porcelain and much earthenware bears marks or devices for the purpose of identification. Stonewares, apart from those of Wedgwood, are not so often marked. Chinese porcelain marks usually record the dynasty and the name of an emperor, but great caution is necessary before accepting them at their face value. The Chinese frequently used the mark of an earlier reign as a sign of veneration for the products of antiquity and, in recent times, for financial gain.

The majority of European factories adopted a device —for example, the well-known crossed swords of Meissen taken from the electoral arms of Saxony, or the royal monogram on Sèvres porcelain—but these, also, cannot be regarded as a guarantee of authenticity. Not only are false marks added to contemporary forgeries but the smaller 18th-century factories often copied the marks of their more august competitors. If 18th-century European porcelain is signed with the artist's name, it generally means that the painting was done outside the factory. Permission to sign factory work was rarely given.

On earthenware, a factory mark is much less usual than on porcelain. Workmen's marks of one kind or another are frequently seen, but signatures are rare. There are a few on Greek vases.

It is often desirable to identify the provenance and the date of manufacture of specimens of pottery as closely as possible. Not only does such information add to the interest of the specimen in question and increase understanding of the pottery art as a whole but it also often throws fresh light on historical questions or the social habits and technical skills of the time it was made. Since ceramics are not affected by any of the agents that attack metal, wood, or textiles, they are often found virtually unchanged after being buried for thousands of years, while other artifacts from the same period are partially or completely destroyed. For this reason archaeologists use pottery extensively; for example, to trace contacts between peoples, since vessels were often widely distributed in course of trade, either by the people who made them or by such maritime nations as the Phoenicians.

Archaeologists' use of pottery

Pottery making is not universal. It is rarely found among nomadic tribes, since potters must live within reach of their raw materials. Moreover, if there are gourds, skins, and similar natural materials that can be made into vessels without trouble, there is no incentive to make pottery. Yet pottery making is one of the most widespread and oldest of the crafts.

II. Western pottery

ANCIENT NEAR EAST AND EGYPT

Ancient Near East. In the early 1960s, excavations at a Neolithic settlement at Çatalhüyük, on the Anatolian Plateau of Turkey, revealed a variety of crude, soft earthenware estimated to be approximately 9,000 years old. A more advanced variety of handmade pottery, hard-fired and burnished, has proved to be as early as 6500 BC. The use of a red slip covering and molded ornament came a little later.

Handmade pottery has been found at Ur, in Mesopotamia, below the clay termed the Flood deposit. Immediately above the Flood deposit, and therefore dating from a time soon after the Flood (about 3000 BC), was wheel-made decorated pottery of a type usually called Al 'Ubaid. Perhaps the most richly decorated pottery of the Near East, remarkable for its fine painting, comes from Susa (Shushan) in southwest Iran. The motifs are partly geometric, partly stylized but easily recognizable representations of waterfowl and running dogs, usually in friezes. They are generally executed in dark colours on a light ground. Vases, bowls, bowls on feet, and goblets have been found, all dating from about 3200 BC. By 3000 BC pottery was no longer decorated. Earthenware statuettes belong to this period, and a vessel (in the Louvre, Paris) with a long spout based on a copper prototype is the ancestor of many much later variations from this region in both pottery and metal.

Remarkable glazed brick panels have been recovered from the ruins of Khorsabad (Dur Sharrukin), Nimrud (Calakh), Susa, and Babylon. They provide the first instance of the use of tin glaze; although the date of its introduction cannot be certainly determined. A well-known fragment from Nimrud in the British Museum belongs to about 890 BC, and by the 5th century BC extremely large friezes, one of them about 11 yards (10 metres) long, were being erected at Susa. The presence of lead in the blue glazes derived from copper suggests that the lead may have been added deliberately as a flux, and that this glazing technique, like that of tin-glazing, subsequently was forgotten—to be recovered only at a much later date.

First use of tin glaze

Ancient Egypt. In Egypt, pottery was made in great variety in the predynastic period (up to *c.* 3100 BC), and a hard-fired ware of good quality was attained. The earliest forms of decoration were geometrical or stylized animal or scenic motifs painted in white slip on a red body. There is comparatively little variation until the 26th dynasty (*c.* 664–525 BC), when clay was probably imported from Greece. Most artifacts are vessels of one kind or another, although pottery figures of variable quality were made, some of the later examples (after 500 BC) showing signs of Greek influence.

The so-called faience of Egypt is an unfired ware and thus, strictly speaking, falls outside the definition of pottery used in this article. As early as the 1st dynasty, figures, vases, and tiles of this material were covered with a fired glaze that was coloured turquoise and green with copper oxide. Later, the colouring materials common to the Egyptian glassmaker, including cobalt and manganese, were added.

ANCIENT AEGEAN AND GREECE

Neolithic period (c. 6000–3000 BC). The potter's art first reached the Aegean in the Neolithic, or New Stone Age. All Neolithic vases are handmade, and the best are highly polished; in other respects, the various local schools have little in common, since communications were severely limited in this remote period. The main centres of pottery production lay in Thessaly and Crete. Thessalian potters favoured a red monochrome ware but occasionally attempted simple painted decoration consisting of rectilinear patterns, with a vertical or diagonal emphasis. The Neolithic pottery of Crete is remarkable for its finely burnished surface, any decoration usually incised.

Early Bronze Age (c. 3000–2000 BC). On the mainland, the pottery initiative passed from Thessaly to the Peloponnese and Boeotia. Early Bronze Age pottery from these two areas has been classified into Early, Middle, and Late Helladic, each subdivided into stages I, II, and III. Early Helladic wares show how quickly pottery fell under the influence of the new craft of metalworking: the two leading shapes, the sauceboat and the high-spouted jug, both have metal prototypes. Painted ornament is rare before the final stage (Early Helladic III, or EH III); in the central phase (EH II), the surface is coated with a dark pigment formed from a solution of the clay. This type of paint, later much improved by the Athenians (see below *Attic black-figure and red-figure*), remained the normal medium of decoration on all Aegean pottery until the adoption of a true silicate glaze in Byzantine times.

The contemporary wares of the Cyclades are similar, but more use is made of incised ornament; spirals are common motifs, while some vases bear primitive representations of ships. The pottery of Early Minoan Crete bears simple geometrical patterns, at first in dark paint on a light clay ground (EM I–II), and subsequently in white

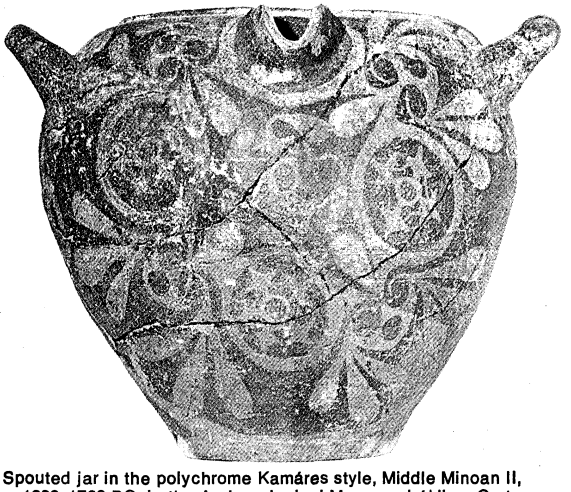
over a coat of dark paint (EM III). The surface of the ware of Vasilikí in eastern Crete (EM II) has a mottled red and black appearance. The commonest Early Minoan shapes are high-spouted jugs and long-spouted drinking jars resembling teapots.

Middle Bronze Age (c. 2000–1580 BC). After the conquest of the mainland by the first Greeks in the Middle Bronze Age, the local schools of pottery developed on widely different lines. The Minyan ware introduced by the newcomers in an unpainted monochrome body thrown on a fast wheel and fired in a reducing kiln to a uniform gray colour that penetrates the biscuit; the surface is then highly polished and feels soapy to the touch. The shapes are all strongly ridged (carinated) and probably derive from metalwork.

Equally characteristic of this period are the mat-painted wares, which are mainly handmade: here rectilinear patterns are applied in dull black or lilac to a porous white surface. This style, although native to the Cyclades, was also widely imitated on the mainland; in the latest stage the ornament falls increasingly under the influence of the polychrome and curvilinear style of Middle Minoan Crete.

Pottery of Crete

By far the most sophisticated pottery of this epoch was made in Crete, contemporaneously with the first palaces at Knossos and Phaistos. The finest ware (Middle Minoan II) is confined to these two royal capitals and to the Kamáres cave sanctuary whence the style derives its name. Over a dark lustrous ground the ornament is added in red and white, the carefully composed designs striking a subtle balance between curvilinear abstract patterns and stylized motifs derived from plant and marine life. The decoration sometimes takes the form of appliqué molded ornament or barbotine (made of slip) knobs. By the time of MM II the use of the fast wheel had become general, imparting a new crispness to the profiles. Among the commonest shapes are carinated cups (often of eggshell thinness), small, round jars with bridge-spouts, and large storage jars (pithoi). In the course of MM III the fashion for polychrome schemes gradually died out, but at the very end of the period (MM III B) a new naturalistic style was born, inspired by the floral and marine frescoes on the walls of the second palaces. The wide distribution of MM pottery illustrates the vigour of Cretan commercial enterprise; several Minoan emporia were founded in the Aegean Islands, while exports also reached Cyprus, Egypt, and the Levant.



Spouted jar in the polychrome Kamáres style, Middle Minoan II, *c.* 1900–1700 BC. In the Archaeological Museum, Iráklion, Crete. Height 24.1 cm.
Hirmer Fotoarchiv, Munchen

Late Bronze Age (c. 1580–1100 BC). Aegean civilization now reached new heights of prosperity, displayed in the luxurious life of the Minoan palaces and the splendid treasures of the shaft graves at Mycenae. Potters were much influenced by work in richer and more spectacular media: many of their shapes can be traced to originals in gold and bronze found in Cretan palaces and Mycenaean tombs. Their decoration owes much to contemporary wall painting, although the motifs are always carefully adapted to the shape of the vase.

With the spread of Minoan culture around the shores of the Aegean, Cretan potters exercised a profound influence on the other local schools, and for the first two centuries of this period the vases of the mainland (known as Late Helladic or Mycenaean) are closely related to Minoan models. In the 16th century BC (LM I A), Cretan potters reversed their colour scheme, returning to dark-on-light decoration. Their repertoire includes some abstract motifs (*e.g.*, running spirals and vertical ripples) but is mainly derived from nature, a continuation of the figurative style of MM III B: flowers, grasses, and olive sprays are drawn with charm and spontaneity. After 1500 BC (LM I B) marine creatures are much in evidence, rendered with considerable realism: in a setting of coral and seaweed may be found argonauts, starfish, dolphins, and, above all, the octopus, wrapping his tentacles round the vase. On the palace style amphorae of the late 15th century BC (LM II), however, there is a reaction against this extreme naturalism: plants and marine life continue, but in a more stylized and symmetrical form.

After the destruction of Knossos in *c.* 1400 BC, the artistic initiative passed to Mycenae and remained there until the end of the Bronze Age. In the 14th and 13th centuries BC (LH III A and B), Mycenaen vases were widely exported, not only to Egypt and the Levant but also as far west as Italy and Sicily. In the interests of commerce, pottery was mass-produced, and the Mycenaean colonies on Rhodes and Cyprus were as prolific as the mainland. Some shapes, like the stirrup-vase, were imported for their contents of oil and unguents; others, such as the tall stemmed goblets, were prized for the excellence of their form. Yet, in spite of their high technical standards, the decoration shows a lack of invention. In the absence of any new ideas, the old floral and marine motifs were subjected to an ever-increasing degree of stylization: the flowers degenerate into chevrons and dashes, the octopus into wavy lines. At the same time there is a new tendency to concentrate the decoration into a single focal zone, in anticipation of later Greek pottery. A few large jars bear crude representations of human figures in chariot scenes, probably derived from palace frescoes. (No less schematic are the painted female figurines found in tombs and shrines of this period.) In the pottery of the 12th century BC, which saw the collapse of Mycenaean civilization (LH III C), there is an abrupt decline in quality as well as in artistic imagination.

Proto-geometric and geometric styles

Early Iron Age (c. 1100–c. 725 BC). Pottery was the first art to recover its standards after the Dorian invasion and the overthrow of Mycenae. Athens escaped these disasters and in the ensuing dark age became the chief source of ceramic ideas. For a short time Mycenaean motifs survived in debased form but on new shapes. This Submycenaean ware soon gave place to the style known as Protogeometric (*c.* 1100–900 BC) by a natural process of evolution that converted the decaying Mycenaean ornament into regular geometrical patterns; thus, the slovenly spirals were transformed into neat sets of concentric circles, always drawn with a compass fitted with a multiple brush. These circles are the hallmark of Protogeometric decoration, which, like the latest Mycenaean, is confined to the handle zone; in the final stage the rest of the surface is covered with a thick black paint remarkable for its high lustre. Many shapes were inherited from Submycenaean, but all were tautened and vastly improved: the drinking vessels rest on high conical feet, while the closed vases have graceful ovoid bodies. After its invention in Attica, the Protogeometric style spread to other parts of the Aegean world.

In the early 9th century BC Athenian potters introduced the full Geometric style by abandoning circular for rectilinear ornament, the key meander assuming the leading role. At first decoration was restricted to a small reserved area surrounded by the lustrous dark paint; later, as the style approached maturity, more decorated zones were added, until the potter achieved a harmonious balance

between light and dark. In the 8th century, after nearly 400 years of abstract decoration, living creatures appear once again, although their style is hardly less angular than the geometric ornament that supports them. Geometric pottery reached its fullest development in the gigantic amphorae and kraters that served as grave monuments in the Athenian Dipylon cemetery; here a funerary scene, showing the corpse on the bier surrounded by mourners, occupies the main panel, while other friezes contain chariot processions, battles on land and sea, rows of animals, and linear geometric designs. The creators of these monumental vases established a continuous tradition of figured painting that persisted on Greek pottery until the end of the Classical period; the immediate consequence of their innovation was a loss of interest in purely abstract design, which became increasingly perfunctory on the latest Geometric vases.

Period of Oriental influence (c. 725–c. 600 BC). After several centuries of isolation, the renewal of contact with the Near East provided a welcome stimulus to the Greek potter. In art, as well as in commerce, it was Corinth that now led the way. Unlike the Athenians, Corinthian potters specialized in small vases and especially in the tiny aryballos, or scent bottle, which found a ready market all around the Mediterranean. Hence arose the miniature style called Proto-Corinthian, which borrowed much of its repertoire from the fauna and flora of Syrophoenician art. Processions of animals, both real and legendary, are placed in the main friezes, while lotus flowers and palmettes serve as subsidiary ornament. When human beings appear, mythical scenes can often be recognized, reflecting the early diffusion of Homeric epic poetry. It was on Proto-Corinthian vases that the technique known as black-figure was first applied: the figures were first drawn in black silhouette and were then marked with incised detail; further touches were added in purple and white.

First use of black-figure technique

Other notable Orientalizing styles arose in Attica, the Cyclades, Laconia, and Rhodes, regional differences in pottery becoming more clearly marked as the Hellenic city-states grew into self-conscious political units. The Athenians still did their best work on large funerary vases. At first they cultivated a wild and grandiose manner in which the figures of men and animals were elaborated in outline; later, incised ornament introduced from Corinth imposed a salutary discipline. Cycladic potters also attempted the grand manner; Laconian work, on the other hand, is confined to a small scale and owes comparatively little to Oriental influence. The Rhodians rarely progressed beyond animal friezes drawn in outline; their style is known as "wild goat", after their favourite quadruped.

Attic black-figure and red-figure. *Archaic period (c. 750–c. 480 BC).* By *c.* 550 BC Athens had once again become the principal centre of pottery manufacture in Greece, having ousted its Corinthian rivals from the overseas markets. Its success is at least partially due to a sudden improvement in technique, for its potters had learned how to obtain the familiar orange-red surface of their vases by mixing a proportion of ruddle, or red ochre, with their clay. As the main medium of decoration, the Athenians perfected a shiny black pigment that was more lustrous than anything that had been hitherto achieved.

Difference between black- and red-figure techniques

In these centuries most of the more important vases were painted either in the black-figure or in the slightly later red-figure technique, so that some explanation of the essential difference is necessary. The red-figure style can be compared with a photographic print, the black-figure with a negative. The latter figures were painted in silhouette in glossy black pigment on the orange-red polished surface. Details were indicated by incised lines and by the occasional use of white and purple, the female figure, especially, being painted in white. Decoration on the red-figure vases was first outlined in black; the surface outside of the outline was then completely covered by the black pigment, leaving the figures reserved in red. Details were added in black, and in dilutions of the black pigment that appear as brown; purple is occasionally found at first but dies out in mature red-figure work. The use of white was revived on the gaudier vases of the 4th century, where yellow brown, gold, and even blue are sometimes used. The forms of Attic black- and red-figure, in the course of centuries, were limited to certain well-defined types, such as the amphora, kylix, krater, and hydria.



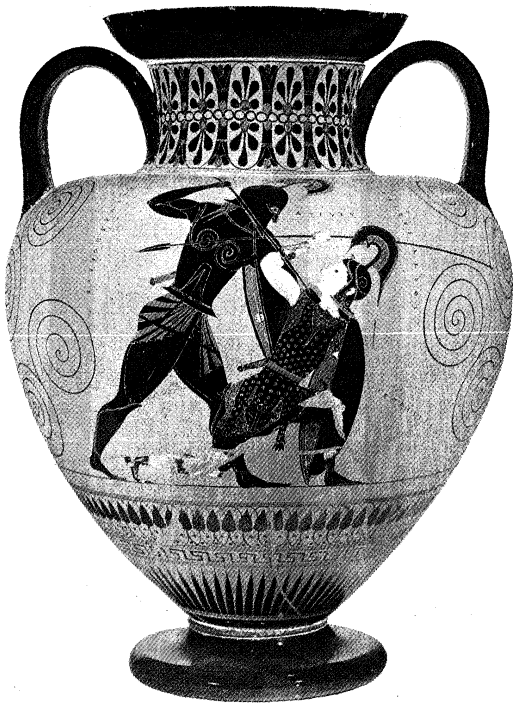
Reveling satyrs, Attic psykter (wine cooler) in the red-figure style, signed by Douris, *c.* 480 BC. In the British Museum. Height 28.6 cm.
By courtesy of the trustees of the British Museum

The practice of signing vases, already begun in the 7th century, became more common in the 6th. The signatures record either the potter or the painter or in some cases both. The inscription on the celebrated François vase in the Museo Archeologico in Florence—"Ergotimos made me; Kleitias painted me"—supplies the first positive evidence that, with only occasional exceptions, the two functions had become separate. When the name of a recognizable painter is not known from an inscription, it has become the fashion to name him after the potter with whom he usually worked: thus, the "Amasis painter" is the habitual colleague of Amasis the potter.

The Attic black-figure style was well developed by the beginning of the 6th century. Among the most favoured subjects were the labours of Hercules, Theseus, and the revels of Dionysus with his attendant train of satyrs and maenads. The finest Attic black-figure vases were made between 550 and 520 BC, the figures being rendered in a mature Archaic style much influenced by contemporary developments in sculpture. This is the generation of Exekias, the greatest master of the technique. He excelled in painting and in finely engraved detail; he also succeeded, where others had failed, in endowing his figures with mood and emotion, as well as the capacity for action. With Exekias the possibilities of black-figure were virtually exhausted, and after the introduction of red-figure (*c.* 530 BC) it is not surprising that the best artists soon turned to this new technique, which allowed a greater freedom of expression and more naturalistic treatment of the human body. After *c.* 500 BC the only important vases in black-figure are the amphoras presented to victors at the Panathenaic Festival; these have a figure of Athena standing between two pillars and are usually inscribed "I am one of the prizes from Athens."

Early red-figure artists

The early red-figure artists were not slow to exploit the advantages of the new system. Benefitting from the experience of relief sculptors, they had mastered the problems of foreshortening by the end of the 6th century; but since they still avoided any suggestion of depth in their grouping, they were able to convey the illusion of a third dimension without doing violence to the two-dimensional surface of the vase. The most successful work was done in the final years of the Archaic period (*c.* 500–*c.* 480 BC) when the style of the figures, with their formal and elaborate patterns of drapery, was still decorative rather than naturalistic. Monotony was avoided through the use of a



Achilles slaying Penthesilea, Attic amphora in the black-figure style, signed by Exekias, *c.* 530 BC. In the British Museum. Height 41.3 cm.
By courtesy of the trustees of the British Museum

wide variety of poses and simple devices for rendering character and mood. Besides the old heroic and Dionysiac themes, many scenes from daily life (especially orgiastic banquets) were now being used.

Classical period (c. 480–c. 330 BC). This period saw a progressive decline in Attic vase painting. Because of the limitations imposed by the pot surface, the vase painter could no longer keep pace with the rapid advance toward naturalism in the major arts; the occasional attempts at perspective and depth of grouping simply detracted from the shape of the vessel (a mistake repeated in some painting on Italian maiolica in the late 16th century AD). Furthermore, in contrast with the earlier wares, much of the later Attic vase painting shows a saccharine sentimentality and triviality in both the choice of subject and its treatment. Distinguished exceptions are the funerary lekythoi of the late 5th century, decorated in subdued mat colours on a white background. The figures on these vases, isolated and statuesque, share the serenity and restraint of the Parthenon sculptures and suggest something of the grandeur of classical free painting, nearly all of which is now lost. In the 4th century, the figured decoration of pottery had become a degenerate art, and by *c.* 320 BC it had died out in Attica.

In addition to their black- and red-figure vases, the Athenians manufactured plain black-painted wares in great quantity; these follow the shapes of the figured pottery.

Hellenistic period (c. 330–c. 30 BC). After the end of red-figure, Greek pottery is undistinguished. Painted decoration is virtually limited to festoons of ivy, laurel, and a vine in white or yellow over a black ground; the black pigment loses its lustrous sheen and assumes a dull metallic texture. A class of hemispherical bowls, known as Megarian, was made in molds and bears relief decoration in imitation of metal bowls. More remarkable are the contemporary terra-cotta figurines; among the most accomplished are the draped women from Tanagra in Boeotia, whose artistic value is sometimes marred by excessive sentimentality.

ETRUSCAN AND ROMAN

Etruria. At the beginning of the Iron Age (*c.* 900 BC), the most characteristic vessel of the Villanovan culture is the cremation urn. It is usually biconical in shape but sometimes takes the form of a primitive hut, decorated with quasi-architectural ornament in relief.

The first pottery of importance is the Etruscan ware called *bucchero*, which was fired in a reducing kiln. The earliest examples of the 8th century BC, for which the wheel was rarely used, were decorated with incised or engraved geometric patterns. By the 6th century lively and stylized birds and animals were engraved, modelled, or applied in friezes or in conjunction with such geometric patterns as re-entrant (coiling inward) spirals. Later, relief ornament was often executed by rolling a cylinder with design recessed in intaglio over the soft clay, the principle being the same as that used to make Babylonian cylinder seals. Vases with covers in the form of a human head, with arms slipped through fixed ring handles, were made for funerary purposes until about the mid-6th century.

In the late Archaic period the Etruscans excelled in life-size terra-cotta sculptures, of which the outstanding examples are the menacing figure of Apollo, from his temple at Veii, and the large sarcophagi from Caere, with couples of banqueters reclining on the lid. Figures, heads, and busts continued to be produced in the Hellenistic period.

Greek influence

Proto-Corinthian ware was copied with great exactness by Greek colonists as early as 700 BC at Cumae, near Naples. The Etruscans soon learned to use the Greek black pigment, and stylized human and animal figures appear in red, black, and white on a light clay or on the *bucchero* surface. Copies of the black-figure vases were soon so accomplished that it is not always easy to tell exactly where a specimen was made. The red-figure class, however, is rarely difficult to separate from Greek work. The decoration is much more complex and elaborate, and the reverse is often carelessly executed. (Long after the red-figure style had fallen into disuse in Greece, it lingered on in Italy, particularly in the south.)

Roman Empire. The characteristic and most widely dispersed type of pottery of the Roman Empire was the red, polished Arretine ware, so called because manufacture was at first concentrated at Arretium (modern Arezzo). It is sometimes also misleadingly termed Samian ware, from a supposed connection with the island of Samos. The body was generally formed in a mold and was frequently decorated with raised designs. These were

By courtesy of the trustees of the British Museum



Etruscan amphora of *bucchero* ware decorated with a frieze of horsemen in relief, 6th century BC. In the British Museum. Height 52.1 cm.

achieved by using a mold that had itself been impressed with several stamps arranged in the desired pattern. This decorative technique—which gave the ware yet another name, terra sigillata (clay impressed with designs)—was borrowed from metalwork. The patterns, too, were often influenced by metalwork and include floral and foliate motifs, mythological scenes, and scenes from daily life. The potteries at Arretium, which were organized on factory lines, operated between about 30 BC and AD 30. Their products were highly prized and widely exported.

Lead glazing perhaps originated or was rediscovered (the Assyrians having used it) in Egypt. Certainly it was established in the Near East by the 1st century BC. The glazes were generally stained with copper to yield a greenish colour and were sometimes used over relief decoration which, like the designs on Arrentine ware, betrays the influence of metalwork. The technique reached Italy and France by the 1st century AD.

Of the other varieties of Roman pottery, lamps made either in a buff or a dark gray clay are exceptionally common and usually have an impressed or molded design. A few depicting Christian motifs or gladiatorial combats are prized a little more highly than most specimens, but, generally, they are of little value, monetarily or otherwise. Molded terra-cotta plaques with reliefs of mythological and other subjects borrowed from Greece were often used to decorate buildings.

ISLAMIC

Influence of Islāmic pottery on Europe

In quality, the Islāmic pottery of Syria, Egypt, Mesopotamia, Persia, Afghanistan, and Anatolia rivals even the wares of the Far East, and its influence on the development of European pottery was more profound than that of any other region except China. In China, too, Islāmic innovations were adopted, while the Islāmic potter, in his turn, owes an incalculable debt to the Chinese.

Near and Middle Eastern pottery was at its best between the 9th and 13th centuries, and its history is closely linked to the fortunes of the caliphate (the dominion of the temporal and spiritual head of Islām). Each dynasty was surrounded in its capital by a wealthy and beauty-loving court that patronized artists and artisans. When one dynasty fell and another established itself elsewhere, it seems that the finest potters emigrated to the new capital, carrying with them their special, and often secret, skills. At first the principal centres of manufacture were Baghdad, al-Fusṭāṭ (old Cairo), and Samarkand; later they shifted to Raqqah on the Euphrates and to Rāy (Rhagae) and Kāshān, both in northern Iran.

Most of the extant pottery has been excavated and consequently is fragmentary. Little made before the 14th century has survived above ground; and tombs, often rich depositories of undamaged wares in other regions of the world, are fruitless because Muslims did not bury pottery with their dead. Only one or two discoveries of undamaged wares have been made: for example, at Gurgan, Iran, entire specimens were found carefully packed in large earthenware jars. They had probably formed part of the stock of merchants, who buried them and fled before the invading Mongols in 1221. Because of deterioration through burial, much Islāmic pottery (like Roman and Near Eastern glass) is iridescent.

Umayyad. There is little pottery of merit from the period of the Umayyad caliphate (661–750). At this time the capital was at Damascus, and the chief interest of the pottery lies in its mingled Mediterranean and Middle Eastern derivation; for example, attempts were made to synthesize the formal repetitive style derived from the ancient Babylonian and Assyrian civilizations with naturalistic ornament in the Greco-Roman style. When the ʿAbbāsids overthrew the Umayyads and moved the capital to Baghdad, the European influence on ornament waned. Good use continued to be made of Western techniques, however, particularly of lead glazes that had been employed by Greek and Roman potters since the 3rd century BC.

ʿAbbāsid. An event that had a profound effect on the development of the Middle Eastern pottery was the presentation of a number of T'ang porcelain bowls to the caliph Hārūn ar-Rashīd about AD 800 (see below *China: T'ang dynasty*). Shortly after this, the first fine pottery was produced in Baghdad and elsewhere in the caliphate. Thus, it seems possible that it was through the example of the Chinese that pottery came to be regarded as an artistic medium instead of a purely utilitarian one. This supposition is borne out by the fact that T'ang wares were in great demand and were imported in large quantities after this date: they and early Islāmic imitations, particularly of the dappled T'ang glazes, have been found in various parts of Mesopotamia and as far apart as Egypt and eastern Persia. Unlike their contemporaries in China, however, Islāmic potters aimed primarily at richness of colour and decoration rather than beautiful shapes and textures. Nearly all their pottery is glazed and is painted with elegant, rather stylized motifs. Floral and foliate ornaments predominate, although complex geometrical patterns are also characteristic. In theory there was a religious ban, formulated in the Ḥadīth (traditions of the Prophet), on all representations of animal life, which were thought to encourage idolatry. In practice, particularly in Persia, the limitation was often disregarded except in the decoration of mosques. The animal figures on pottery are spirited and rhythmical, while the human ones tend to be stiff, resembling those in contemporary miniatures. Arabic calligraphy was commonly and effectively used as an element of design.

Islāmic technical innovations

The Islāmic potters were responsible for a number of important technical innovations, the most influential of which was the rediscovery of tin glaze in the 9th century AD. Though tin was first used by the Assyrians and according to some authorities was discovered as early as 1100 BC, it had fallen into disuse. The ʿAbbāsid potters first used it in an attempt to imitate the texture of T'ang wares, but soon it became the vehicle for characteristically Middle Eastern decoration. From Mesopotamia and Persia the technique was later taken to Moorish Spain and then to Italy and other parts of Europe, where it was employed for a number of important wares—maiolica, faience, and delft (see below *European: to the end of the 18th century*).

Like that of tin glazing, the technique of lustre painting was perfected (and probably invented) by Islāmic potters. Again, like tin-glazing, it later passed to Muslim Spain but not to the Far East. Lustre on pottery probably was first used to cover entire vessels, thus simulating vessels made of precious metals that were proscribed by sumptuary laws laid down in the Ḥadīth, which sought to preserve the earlier simplicity of Muslim life. The metallic pigments employed in lustre painting were probably silver and copper in combination, although an occasional ruby glint suggests that gold may sometimes have been included. After firing, the painting may be dull yellow, golden brown, or olive, tinged with green or red.

Extensive use was made of slip. Wares such as the early Gabrī type of the 11th century and later have a reddish body washed over with white slip. Designs were executed by scratching through the slip to the body beneath (sgraffito). On some later specimens the background was cut away to leave a raised design in white slip or the design was incised through the white slip, as in the earlier specimens, and then was itself covered with green and brown glazes. The usual motifs are large floral forms, animals, and bold inscriptions. Sgraffito ware became common throughout the Middle East and appears in Egypt and Syria in the 13th century.

Many fragments of Chinese pottery and porcelain have been found at the site of Sāmarrāʾ, on the Tigris, where the ʿAbbāsids built their summer palaces in the 9th century (see below *China: T'ang dynasty*). Among the native wares are some made in a buff body decorated in relief under a green glaze; others with monochrome green, white, and yellow glazes or with glazes in imitation of a well-known type of T'ang decoration; and those painted with cobalt blue (perhaps the earliest use of underglaze blue) and further embellished with lustre of various colours.

Sāmānid. To the northeast, beyond the Oxus (modern Amu Darya) River, the Sāmānid dynasty (874–999) be-

came practically independent of the caliphate at Baghdad and fostered a national artistic and literary revival. Sāmānid pottery, which has been found chiefly at Samarkand and Nīshāpūr, differs from the pottery of more westerly regions in technique and style. The best pieces have a reddish body covered with a white, vivid red-brown, or purplish-black slip that was then painted and fired under a lead glaze. The function of the slip, besides providing colour, was to prevent the pigments of the painting from running when the lead glaze was applied. The colours used in painting were the same as those of the slips, with the addition of yellowish green and browns. The designs often consist of the angular Arabic Kūfic characters or stylized birds and floral motifs. The shapes are plain—usually either plates or rather shallow bowls—and the total effect is both bold and elegant.

Egyptian. Egyptian pottery of the Islāmic period was at its best during the Fāṭimid dynasty (969–1171). Wares were at first coarser than those of Mesopotamia because of the poor quality of local materials, and the shapes were less refined, since Chinese influence was absent. Lustre painting (probably introduced in mid-10th century) was nevertheless, excellent in quality. A typical feature is the painting on the backs of dishes, a practice derived from Baghdad and later copied by the Moorish potters of Spain. Signed specimens of lustre ware and tin-glazed wares are known, the best coming from a potter named Sa'd.

Painting on the backs of dishes

Toward the end of the period a much whiter type of ware, with a compact body, came into use and thereafter became common throughout the Middle East. Another widespread group of wares, popular until the 14th century, has decoration carved and incised into the body and is covered with transparent glazes. The patterns suggest the influence of some of the Sung wares of China.

Mesopotamia and Persia, 11th to 15th century. In the 11th century the Seljuq Turks overran Persia and Mesopotamia, and their ascendancy lasted until the advent of the Mongols during the 13th century. As the Seljuqs had no capital, the most flourishing cities during this time were those on the trade routes. In the 12th century very fine pottery was made in the new white body recently developed in Egypt; it was decorated with bold carving, occasional piercing, and translucent glaze. Most of these wares are said to have been found at Rāy near Teheran, where many other beautiful wares have been excavated. Wares with a sandy body and a clear glaze were painted with a golden-brown lustre, often in conjunction with blue. These seem not to have been made after the city was sacked by Genghis Khan in 1220. Especially associated with Rāy are examples of *minai* painting of uncommon quality. The *minai* technique, a Persian discovery of the 12th century, was a method of decoration in which colours were painted onto a glazed and fired bowl and then fixed by refiring the bowl at a comparatively low temperature. The advantage of the process was that many colours that would not have withstood the heat of the first firing could now be used. The technique may perhaps have influenced the rare examples of overglaze decoration on late Sung or Yüan wares from Tz'u-chou, although it did not come into common use in China until the early part of the 15th century (see below *China: Ming Dynasty*).

At Rāy the glaze is cream or turquoise, and the *minai* palette included blue, turquoise, purple, red, green, and white, with the addition of gold leaf. All these colours, except the blue, are mat in appearance, and the style strongly recalls that of Persian manuscript illumination of the 13th century.

Another technique employed at Rāy was the use of silhouette decoration, a kind of sgraffito. The pot was covered with a thick black or blue and black slip, and the design was carved out with a knife. The glazes were applied without colour or stained with copper to yield a brilliant turquoise.

Raqqah was a prosperous trading city until it was sacked by the Mongols in 1259. Most of its pottery, which can be dated between the 9th and 14th centuries, is rougher and the designs bolder than those of Rāy. The body is white, inclining to buff, and is covered with a siliceous glaze. Some of the Raqqah fragments are painted with a brownish lustre. Others have designs in relief, sometimes covered with an opaque turquoise glaze or with a bluish-green translucent glaze. In the 12th and early 13th centuries bold designs were executed in black under pale-blue glazes and, more frequently, in blue and black under a clear glaze. Occasionally the glazes were stained purple with manganese.

Kāshān tiles

Kāshān is chiefly famous for its tiles, in fact the words *kāshī* or *kāshānī* ("of Kashan"), are commonly used as synonyms for tile (and have been incorrectly applied to tilework from India). Lustre-painted tiles had been made since at least the 9th century and were used mostly on the walls of mosques and public buildings. Those of Kāshān, particularly in the 13th and 14th centuries, are distinguished by their fine workmanship, brilliance, and intricacy of design. In shape they are square, rectangular, or of interlocking cross or star shapes, each carrying a small part of the total design. The relief inscriptions are frequently picked out with blue pigment.

Also associated with Kāshān are the *lakabi* ("painted") wares made in the 12th century. The term, a misnomer, refers to a variation of the sgraffito silhouette technique mentioned above: an incised design was decorated with different coloured glazes (blue, yellow, purple, and green), which were kept apart by intervening threads of clay. Although a number of *lakabi* wares were also made at Raqqah, the technique was soon abandoned at both places, as the glazes always tended to run out of their compartments during firing, giving a smudged effect.

Both the original site of Solṭānābād and the nature of the wares that may have been made there are extremely uncertain. Principally associated with it are wares decorated with relief molding under a turquoise or dark-blue glaze or painted in black slip under a clear turquoise glaze. They date from the second half of the 13th century onward. Toward the end of the 12th century the glaze material was frequently mixed with the white-burning clay then in use. In the more highly fired specimens the product is not unlike a primitive soft porcelain, and occasional specimens are slightly translucent. These wares probably inspired the attempts to make porcelain at Florence (see below *European pottery to the end of the 18th century*). Neither stoneware nor true porcelain was ever made in Persia.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.



Persian tankard with decoration cut through a black slip and covered with a turquoise glaze, 12th century, found at Soltānābād (modern Arāk, Iran). In the Victoria and Albert Museum. Height 12.7 cm.

After the Mongol conquests of the 13th century the production of pottery practically ceased, except at Kāshān. A slow revival began about 1295, and, although pottery in the Near and Middle East never again reached its former height, some fine wares were made at Solṭānābād in the 14th century. Good use was made of the rich sombre colours beloved by the Mongols, particularly dark blues, grays, and blacks.

Later Persian. Since the whole of Central Asia now lay under the Mongol domination, overland trade with

China greatly increased. By the 15th century Chinese influence, particularly that of Ming blue-and-white, was predominant, and the older styles were tending to die out (see below *China*: *Ming dynasty*). A group of blue-and-white wares belonging to the 15th and early 16th century are known as Kubachi wares because large numbers of them survived above ground in this town in the Caucasus. They have a very soft body, a brilliant crackled glaze, and rhythmical and spontaneous designs. The later Kubachi blue-and-white is closer to the Chinese originals.

Polychrome appears about 1550, and the palette includes a red related to, though lighter than, the Armenian bole introduced about the same time in Turkey (see below *Turkish pottery*). The best polychrome painting was done on tiles. Tabriz has been suggested as the real centre of manufacture, but although it seems likely that Tabriz was a manufacturing town in view of its tiled mosques and the fact that Tabriz potters were famous abroad (and indeed were either invited or carried off to Turkey on two occasions), no kiln sites have been found there.

One of the later kiln sites in Persia is Kerman, which was the leading pottery centre in the 17th century. Its wares are characterized by a very strong bright blue and a wavy, rather bubbly, glaze. Pseudo-Chinese marks were frequently added to the blue and white. The most usual colours on Kerman polychrome wares are blue, green, browns, and a bright red similar to Armenian bole. The quality of production declined considerably during the 18th century.

Lustre painting

Lustre painting, which had almost ceased in the 13th century, was revived during the second half of the 17th century and perhaps lasted into the 18th century. Its place of manufacture is not known. Most of the objects decorated in this manner are small bottles or spittoons, and their cramped designs are timid and fussy. The lustre is warm brown, often with a strong red tinge, and was sometimes used in conjunction with blue glaze. Another early technique revived at the same time was piercing, formerly practiced in the Seljuq era. There are a number of delicate pierced white wares covered with a colourless glaze, which were imitated in China during the reign of Ch'ien-lung. Pierced pottery and porcelain of this kind was often known in Europe as Gombroon ware, the name of the port (now Bandar ʿAbbās) from whence it was shipped.

Chinese celadon was imitated, not very successfully, from the 14th century. In the 16th century other monochrome glazes were produced at Kerman and elsewhere. These and the celadon were frequently decorated with painted or incised ornament—the former a practice quite foreign to Chinese Sung dynasty wares.

During the 18th century most of the pottery produced in Persia was inferior blue-and-white. In the 19th century the standard declined still further with the adoption of the Chinese-inspired *famille rose* palette (see below *China*: *Ch'ing dynasty*), and only a group of wares made at Teheran between 1860 and 1890 can command any respect. Some excellent peasant pottery with a buff body and lead glaze was made in Turkistan, however.

Syrian. The potters from al-Fusṭāṭ and Raqqah may have migrated to Damascus after their potteries were destroyed by the Mongols, for lustre painting continued in Syria throughout the 13th and 14th centuries after it had ceased elsewhere in the Middle East. The lustre ranges in colour from silver to yellow and dull brown and is often used in conjunction with a blue glaze on big, heavy jars and albarellos (a jar with an incurving waist, used for dry drugs and ointments). Characteristic are gold designs arranged in panels with much use of inscriptions and heraldic devices. The body material is coarse and grayish, and the glaze sometimes has a wide crackle. Lustre painting fell into disuse in Syria about 1400 and might have died out altogether had not the secret meantime been carried from Egypt to Spain (see below *European pottery to the end of the 18th century*). The commonest type of Syrian pottery in the 14th century is a blue-and-black style similar in shape and design to the lustre ware. Rather uncertainly drawn animals appear on some of the vessels.

The earliest known Middle Eastern copies of Chinese blue-and-white were made in Syria at the end of the 14th century. Blue-and-white became commoner on both vessels and tiles in the first half of the next century. Later, the potteries seem to have fallen into disuse until the new mosque built in Damascus by the Turkish ruler Süleyman I (the Magnificent) in the mid-16th century provided a fresh impetus for the industry. The polychrome tiles of the 16th century at first have designs with a hard black outline; later, a more flowing foliate style was developed. A soft purple replaces the Armenian bole of Iznik (see below *Turkish pottery*). Vessels and tiles, gradually declining in quality, continued to be made in Damascus until the end of the 18th century.

Turkish. A branch of the Seljuq Turks occupied Anatolia from 1078 to 1300 and was succeeded by the Ottoman Turks, who first extended their lands westward, conquering Byzantium in 1453 and in the 16th century becoming masters of much of southeastern Europe and the lands lying to the east and south of the Mediterranean. The first notable pottery wares from Turkish lands were the tiles and bricks covered with coloured glazes made in Anatolia for architectural purposes in the 13th century. Mosques in particular were decorated in this way. (Persian influence in decoration suggests the presence of potters from that region.) The art of tilework apparently died out after 1300 and was not reintroduced until about 1415, when Persian craftsmen were brought from Tabriz to decorate the mosques at Bursa and Edirne. Apart from tilework, pottery appears to have received little encouragement until the late 15th century, by which time the chief centre of production was firmly established at İznik (earlier called Nicaea).

The great era of Turkish pottery

The great era of Turkish pottery (*c.* 1500–*c.* 1580) coincides with the expansion of Ottoman power. Decoration was at first influenced by 15th-century Ming blue-and-white porcelain. The earlier designs were probably taken at second hand from Persian sources, since a distinctly Persian flavour is usually evident. This is indicated by the intricacy of the designs and their arrangement in bands, and by the shapes of some of the vessels, which suggest the influence of metalwork. At one time the wares in this style, which lasted until about 1525, were thought to come from Kütahya in central Anatolia and are still sometimes known by that name.

At this and later periods the body of Iznik pottery was soft and sandy. It was made from grayish-white clay covered with a thin slip that was usually white, although occasionally red or blue was used as a ground on later wares. Decoration was carried out in underglaze colours under a transparent siliceous glaze. The commonest shapes are flat dishes, but jugs, dishes with a high foot, and bowls are also found. Cylindrical vessels with small rectangular handles set halfway down are flower vases, not tankards, as one might think. A rare form is a pottery version of a mosque lamp.

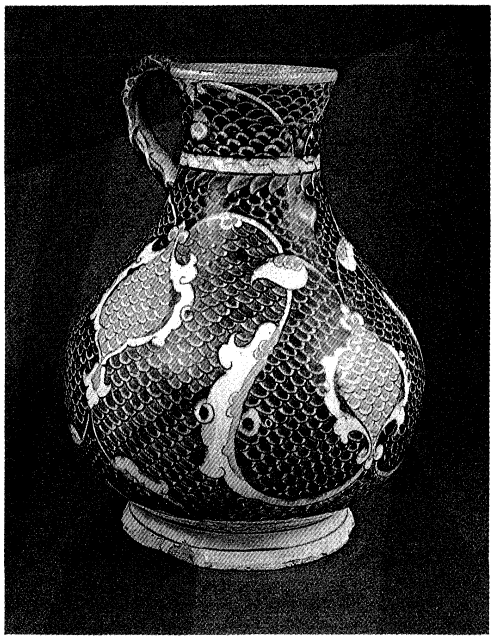
During the next period (*c.* 1525–50), some wares of which have been erroneously attributed to Damascus, Iznik pottery was at its finest. Ming blue-and-white was now copied directly; for example, the central motif of grapes on a dish in the Victoria and Albert Museum, London, is an almost exact imitation of a well-known mid-15th century Chinese motif. On the same dish is a characteristic border pattern, which was called the Ammonite scroll border because it was thought to resemble the coiled shell of the fossil ammonite but which is certainly a debased version of the Ming Rock of Ages pattern. This scroll border appears often; a slightly later and even more debased version, which incorporates large S-shaped scrolls, is sometimes known as the dollar pattern.

The palette was gradually expanded to include turquoise, sage green, olive green, purple, and black. Most of the blue and turquoise specimens are painted with flowers. The Chinese flora motifs were almost entirely replaced by tulips, poppies, carnations, roses, and hyacinths in the form of fairly symmetrical sprays springing from a single point. The earliest flowers are often rather more stylized than the later, perhaps because the representation of living things was prohibited by Qur'ānic

(Koranic) tradition. Even on comparatively late examples, floral designs are sometimes stylized to the point of abstraction, suggesting that decorators might have suited their patterns to the religious susceptibilities of their customers. An effective abstract pattern is formed from a series of overlapping scales that are usually carefully

Scale pattern

By courtesy of the Victoria and Albert Museum, London; photograph, Wilfrid Walter



Tin enamelled Turkish jug decorated with the characteristic scale pattern, Iznik (Anatolia), Ottoman period, *c.* 1575. In the Victoria and Albert Museum, London. Height 27.9 cm.

drawn. The same ground was later employed in Italy on maiolica and at the Berlin porcelain factory and may have indirectly inspired the series of wares with scale grounds made at Worcester, England.

After about 1550 Iznik pottery enters its third stage. The most notable technical innovation is the use of Armenian bole (sealing-wax red), a thick pigment that stands out in slight relief from the surface of the vessel.

The other great change is that tiles, which had previously been made in small numbers, became all important and remained so until the early 17th century. They were used to provide lavish decoration for the new mosques built at Constantinople by Süleyman I. Once again potters were brought from Tabriz to begin the work. Much use is made of copper green and the new red, the colours very brilliant on the glossy white ground. The tiles, usually square, make up flowing repeating patterns or long high pictures with elaborate borders.

On pottery, symmetrical sprays of flowers continued to be used as decoration until about 1600. Paintings of animals and birds are found occasionally, probably executed by Persian workmen since their resemblance to Persian wares is strong. The rare specimens with human figures were probably painted by Greeks or Armenians for export to the West. Turkish sailing vessels sometimes appear as a decorative motif.

In the 17th century the quality of Iznik wares declined, and by 1800 manufacture had ceased. At Kütahya, pottery making had begun by 1608 and continued into the middle of the 20th century. The wares, though inferior, have some resemblance to those of Iznik with the addition of a yellow pigment.

EUROPEAN: TO THE END OF THE 18TH CENTURY

European wares made before the 19th century fall into six main categories: lead-glazed earthenware, tin-glazed earthenware, stoneware, soft porcelain, hard porcelain, and bone china.

Lead-glazed earthenware was made from medieval times onward and owes little to outside influences. The body is generally reddish buff in colour; the glazes are yellow, brown, purplish, or green. The wares are usually vigorous in form but often crudely finished. Lead-glazed wares fell out of favour when tin glaze became widely known toward the end of the 15th century, but they returned to popularity with the advent of Wedgwood's creamware shortly after the middle of the 18th century. The body of this later lead-glazed earthenware is drab white or cream, the glaze clear and transparent like glass, and the forms precise.

The first important tin-glazed wares came from Italy during the Renaissance, and these colourful examples of the painter's art exerted a profound influence on later work elsewhere. Manufacture spread rapidly, first to France, then to Germany, Holland, England, and Scandinavia. Under the name of maiolica, faience, or delft, it enjoyed immense popularity until the advent of Wedgwood's creamware, after which the fashion for tin-glazed ware declined rapidly.

Stoneware is first commonly seen in Germany during the 16th century; its manufacture was developed in England during the 18th century, culminating in the unglazed ornamental jaspers and basaltes of Wedgwood.

Two other types of ware, less common than those already discussed, are slipware and lustreware. Slip was applied both as a covering over an earthenware body and in the form of decoration, for example on the sgraffito wares of Italy (which owe a good deal to similar wares from Byzantium) and the dotted and trailed slips of 17th- and 18th-century England. Lustre pigments were used in Spain, where they are the principal decoration on the magnificent series of wares referred to as Hispano-Moresque; in Italy, where they supplement other modes of decoration; and much later, in England—although in the last case they are no longer artistically important.

The manufacture of soft porcelain was essayed in 16th-century Italy under the patronage of Francesco de' Medici, grand duke of Tuscany. Similar attempts were made elsewhere in Italy about the same time, and manufacture is supposed to have been continued at Pisa and at Candiana, near Padua (Padova). The first production of soft porcelain on a considerable scale did not take place, however, until toward the end of the 17th century in France.

Porcelain and bone china

In Saxony about 1675 Ehrenfried Walter von Tschirnhaus started experiments to make porcelain from clay mixed with fusible rock. Almost certainly he had made hard porcelain by the end of the century, but manufacture did not become a practical commercial proposition until the year of his death, in 1708. Experiments were continued by his assistant, an alchemist named Johann Friedrich Böttger, who is sometimes credited with von Tschirnhaus' discovery. The factory was established at Meissen about 1710, and the first porcelain sales of any consequence took place at the Leipzig Fair in 1713.

Later, at the end of the 18th century, Josiah Spode the Second added bone ash to the hard porcelain formula to make bone china.

Byzantium. In AD 330 Byzantium became the imperial capital of the Roman Empire and was renamed Constantinople. The term Byzantine, however, is applied to the period that ended in 1453, when Constantinople was captured by the Ottoman Turks (and renamed Istanbul).

Since it was not a Christian custom to bury pottery with the dead, few wares survive, and chronology is difficult. Most of the surviving wares fall into two classes: one is a red-bodied type, sometimes with stamped relief decoration under a clear glaze; the other, a sgraffito type with human figures, animals, birds, monograms, foliate designs, the Greek cross, and the like, engraved through a white slip and covered with yellow and green glazes. The latter is the commonest type after the 12th century. Both styles were fairly widespread and have been recovered in fragmentary form from excavations at Istanbul, and in Greece, Cyprus, and the Crimea.

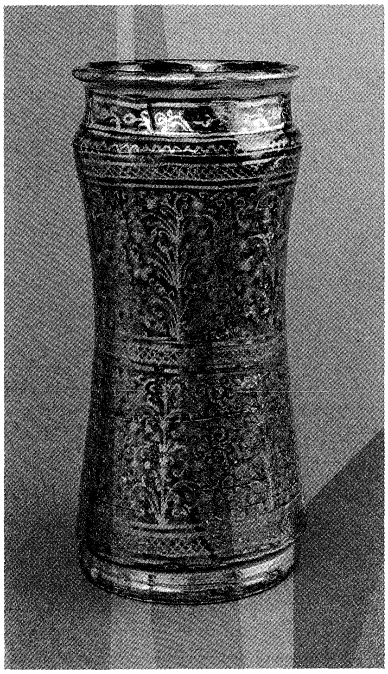
Spain. The earthenware of Spain falls into two classes: lustreware and painted tin glazed ware.

Lustreware. The lustre technique spread to Moorish Spain by way of Egypt, but it is impossible to say exactly when it arrived.

Hispano-Moresque pottery

The body of Hispano-Moresque pottery is usually of fairly coarse clay, which has burned to a pinkish buff, covered with a tin glaze containing lead in varying proportions. The lustre, added overglaze, varies in colour from golden to a pale straw, and a coppery lustre almost invariably indicates at least a 17th century date. Many dishes were additionally painted in blue and, less often, with manganese.

By courtesy of the trustees of the British Museum



Hispano-Moresque albarello painted with lustre on a blue ground, Valencia, *c.* 1460. In the British Museum. Height 27.6 cm.

Most surviving wares of the early period are dishes of various shapes. Less common are albarellos, waisted drug jars based on a Middle Eastern form. Vases based on the old Iberian Amphora but with two massive wing handles (the Alhambra type) are very rare. The decoration on wares of this early period is predominantly Moorish. Fine specimens of this kind are unlikely to be later than 1525. Subsequently, Spanish artists repeated the Moorish designs, but these often degenerate in their hands; and Arabic and the Kūfic script, frequently used by Moorish potters, becomes meaningless. The early designs are, for the most part, plant forms and arabesques, both the vine leaf and the bryony leaf being used. A little later there are magnificently drawn animals in heraldic form, principally lions and eagles. Still later there are deer and antelope, which may owe something to Persian sources. Dishes with coats of arms of noble families surrounded by vine- and bryony-leaf ornament are unusually fine. Many of them were made in Valencia and the neighbouring village of Manises for Italian families. A feature of many of the dishes is the lustre decoration on the reverse. Although often no more than a series of concentric circles, occasionally there are superb eagles and other animals found on dishes from Valencia that are even finer than the obverse designs.

In the 17th century much lustred pottery was made for the cheaper markets and for export to England. The painting is executed in a lustre pigment of deep coppery hue. While this ware is not important in comparison with the early wares, it is often decorative.

Other tin-glazed ware. Although the influence of Valenician lustre pottery on later Italian maiolica is obvious, the wares of Paterna, near Valencia, were hardly less influential in the 14th century. They were decorated in green and manganese, often with motifs taken from Moorish sources; this combination of colours is to be seen in early Italian pottery from Orvieto and elsewhere.

Much tin-glazed pottery of excellent quality was made at Talavera de la Reina, in New Castile, during the 17th and 18th centuries. The palette is characteristic of much Spanish tin-glazed ware; green and manganese play a distinctive part, frequently combined with touches of orange-red and gray. The *istoriato* style of Urbino (see below *Italy*) was copied here, and the Italian painter and engraver Antonio Tempesta (1555–1630) provided a source of inspiration for some of the painting. Alcora, in Valencia, made much faience of excellent quality during the 18th century.

Kinds of tilework

Tilework was particularly common in Spain from the earliest period; according to one proverb, only a really poor man had "a house without tiles." At first tilework was made with a typically Persian technique by which thin slabs of tin-glazed pottery were sawn into pieces and embedded in a kind of mortar (tile mosaic). The *cuerda seca* method of making tiles followed about 1500: outlines were drawn on the surface in manganese mixed with a greasy substance that prevented the coloured glazes used from mingling. Tiles made by the *cuenca* technique had deeply impressed patterns the compartments thus formed being filled with coloured glazes. Tiles were also decorated with lustre pigments.

Porcelain. The early porcelain made at Buen Retiro, near Madrid, in the 1760s, had been justly compared to that of Saint-Cloud. The quality of the ware was good, and some skillful figure modelling was done by Giuseppe Gricci, who had previously worked at Capodimonte (see below *Italy*).

Italy. The pottery of Italy is extremely important not only in itself but for its subsequent influence in other European countries. Indeed, its influence may have spread even farther afield: a few specimens of Ming porcelain have motifs that may have been inspired by it.

There are two well-defined classes of Italian earthenware: maiolica, or tin-glazed ware, and pottery decorated in the sgraffito technique.

Maiolica. Tin-glazing was introduced in the 13th century from the Middle East through the Muslim civilization in southern Spain, wares being shipped from there to Italy by Majorcan traders. The term maiolica was at first applied to this Hispano-Moresque lustreware, but in the 16th century it came to denote all tin-glazed ware.

Italian maiolica is principally noteworthy for its painted decoration, which excelled in technical competence anything produced in Europe since classical times. The painting was executed in several colours on the dry but unfired tin glaze. Great skill was needed, since the surface absorbed the colour as blotting paper absorbs ink, and erasures were therefore impossible. The best wares were given a final coating of clear lead glaze called *coperta.* The range of colours was comparatively limited: cobalt blue, copper green, manganese purple, antimony yellow, and iron red formed the basic palette, while white was provided by the tin-glaze material. When white was used for painting, it was applied onto a bluish-white glaze (*bianco sopra bianco*, or "white on white"), or on a light-blue (*berettino*), or dark-blue ground.

Lustre pigments

Lustre pigments were introduced from Spain. The lustre of Italian wares is often the golden-yellow colour derived from silver, and sometimes it is ruby, suggesting the use of gold. The silver lustre often developed a nacreous effect known as mother-of-pearl (*madre perle*).

The forms of maiolica are few and comparatively simple. For the most part they were dictated by the need for a surface on which the painter could exercise his skill; thus, dishes form the greater part of surviving wares. It is doubtful whether most maiolica was ever intended for general use. Dishes, for instance, were displayed on sideboards and buffets far more often than they were placed on the table. Gaily coloured drug jars were a fashionable decoration for pharmacies and include the albarello shape, copied from Spain, for dry drugs, and a spouted jar for wet drugs. Ewers (pitchers) with a trefoil (leaf-shaped) spout, derived from the Greek oenochoe, were made, as well as the massive jars representative of Florentine work of the 15th century.

The earliest maiolica, beginning in the 13th century, is decorated in green and manganese purple in imitation of

the Spanish Paterna ware. Much work of this kind was done at Orvieto, in Umbria, where the characteristic form was a jug with a disproportionately large pouring lip. Orvieto ware has almost become a generic term for anything in this style, although similar vessels were made at Florence, Siena, and elsewhere. It was current in the 14th century and continued in the 15th century, when other colours were added to the palette. The decorative motifs —masks, animals, and foliage—are Gothic, with some traces of Eastern influence.

From Florence came a series of wares painted in a dark, inky, impasto (or very thick) blue. These, too, have Gothic ornament, particularly oak leaves, which came into use sometime before 1450. Heraldic animals also appear on some specimens. This kind of decoration was obviously inspired by Spanish pottery, and a few examples are hardly more than copies. Soon after 1500, Florentine production was concentrated in the castle of Caffagiolo, in Tuscany, and came under the patronage of the Medici family, whose arms appear frequently. A notable addition to the palette here was a bright red pigment, a most difficult colour to attain and one not often used.

Gothic ornament was gradually displaced by classical motifs, such as grotesques, trophies, and the like, which, early in the 16th century, themselves gave way to the *istoriato* style. This style, no doubt inspired by the achievements of contemporary painting, imitates the easel picture closely. Its realism, including the use of perspective, is quite unlike any previous ceramic decoration. The subjects were often classical, but biblical subjects, some taken from the woodcuts of Bernard Salomon (*c.* 1506–61), are frequently represented. Maiolica was often called Raffaelle ware, a tribute to the influence of the painter Raphael (1483–1520), although he, in fact, never made any designs for pottery. In particular the maiolica painters copied his *grotteschi* (grotesques), motifs adapted from those rediscovered in the grottoes of the Golden House of Nero soon after 1500 and so-called in consequence. They are usually fantastic combinations of human, animal, and plant forms. The works of Albrecht Dürer and Andrea Mantegna were also borrowed, often through engravings made by Jacopo Ripanda (Jacopa da Bologna and Marcantonio Raimondi); some examples are almost exact copies, others are freer interpretations. The paintings sometimes occupy the centre of the dish with a border of formal ornament surrounding them, but in many instances, notably those from Urbino, they cover the entire surface. It is often impossible to regard the pottery body as anything more than a support for the painting, its pictorial or narrative subject having been executed with little or no consideration for the nature of the object it decorated. Although pottery decoration is rarely successful unless it is designed to enhance, or at least not to detract from, the shape of the body, an exception must be made for some of these colourful wares: at their best they are highly ornamental.

Istoriato style

The *istoriato* style probably developed at Faenza (Emilia) in about 1500. One of the earliest and most important centres of production, it had been manufacturing maiolica since before 1450. Almost as early are some examples from Caffagiolo. Castel Durante adopted the same style, and it is particularly associated with the name of Nicola Pellipario (died *c.* 1542), the greatest of the maiolica painters. He also painted grotesques similar to those of Deruta, in Umbria, which are rather more stylized than the grotesques introduced later in the 16th century at Urbino that are humourous and full of movement. The former are often used as a surround to an interior medallion in the *istoriato* style. Urbino was probably the largest centre for the manufacture of maiolica at the time. The industry there was under the patronage of the Della Rovere family, whose name, meaning "oak tree," led to the adoption of the oak-leaf motif in wreathed form.

The Deruta factory

Among the early factories, that of Deruta (which may have been under the patronage of Cesare Borgia) is of considerable importance. Maiolica has been made there from medieval times, and manufacture continues in the mid-20th century. Deruta potters about 1500 were the first to use lustre pigment, which was of a pale-yellow tone, and they also adopted the Spanish practice of painting designs on the reverse of dishes. They also often covered only the obverse with tin glaze and applied a lead glaze to the reverse—again, a typically Spanish practice. The best work was done before 1540.



"Death of the Virgin," maiolica plaque painted in the *istoriato* style by an unknown Italian artist often referred to as the Master of the Death of the Virgin, Faenza, Italy, *c.* 1510. The composition is taken from Martin Schongauer's engraving of the same subject. In the British Museum. Diameter 26.7 cm.
By courtesy of the trustees of the British Museum

The use of lustre pigments at Gubbio, in Umbria, probably started soon after it began at Deruta. The quality of the work was such that maiolica was sent from Castel Durante, Faenza, and even from Deruta itself for this additional embellishment. An interesting series of dishes is that painted with the portraits of young women, often with the addition of a terse and appreciative comment such as "*Bella*," which were made at Deruta, and also apparently at Gubbio.

Maiolica was manufactured in Venice between the 16th and 18th centuries. As might be expected in an important seaport with worldwide trade, its maiolica often shows Eastern influence. The designs of Iznik were sometimes copied (as they were, in fact, on other Italian wares of the period), and imitations of Chinese porcelain of the Ming period gave rise to a style known as *alla porcellana* ("in the manner of porcelain").

Of the later potteries, that of Castelli, near Naples, did excellent work from the 16th century onward, although its later wares tend to become pedestrian. *Istoriato* painting was revived there in the 17th century in a palette paler in tone than that of early work in this style. Much maiolica survives from Savona, in Liguria, a good deal of which is painted in blue in Oriental styles.

Although hardly to be classified as pottery, sculptured reliefs were made by Luca della Robbia (died 1482) in terra cotta and covered with maiolica glazes. He was followed by his nephew, Andrea, and the latter's sons. Giorgio Vasari's suggestion that Luca invented the maiolica glaze is erroneous.

Sgraffito wares. Sgraffito wares are comparatively rare. The technique was derived from Byzantine sources by way of Cyprus which was under Venetian rule from 1472 to 1570. Manufacture was confined to northern Italy, the largest centre being at Bologna. The body of the sgraffito ware was covered with a slip of contrasting colour, the decoration then being scratched through to the body beneath and the whole covered with a lead glaze, which has a yellowish tone. Often the incised designs were first embellished with underglaze colours (blue, green, purple, brown, and yellow) that tended to run during firing. This technique died out finally at the end of the 18th century, but some important work of the kind was done in the late 15th and 16th centuries.

Porcelain. There are only about 50 surviving pieces of the soft-paste porcelain made in Florence at the time of the Medicis, and little is known of its actual production. The earliest definite date for manufacture is 1581. Painting is nearly always in blue with manganese outlines. Most decorative motifs are derived from China, Persia, or Turkey, and the forms usually copy those of Urbino maiolica.

No hard porcelain was made in Italy until Francesco and Giuseppe Vezzi's factory was established in Venice in 1720. It made fine hard porcelain the body of which has a slightly smoky colour. The style is Baroque, and the palette is notable for a brownish red. Another factory, that of Geminiano Cozzi, started in 1764, was the one where most Venetian porcelain was made. Cozzi worked in the Meissen and Sèvres styles and produced some good figures.

Ginori and Capodimonte

The porcelain factory at Doccia, near Florence, still in production in the 1970's, was founded by Marchese Carlo Ginori in 1735. Coffeepots in the Baroque style, sometimes painted with coats of arms, are characteristic of the early period. Equally fine figures were made during the 18th century. Porcelain with figure subjects in low relief was made only at Doccia, although it has been repeatedly and erroneously attributed to the soft-porcelain factory established in the royal palace of Capodimonte by Charles III of Naples in 1743. As well as extremely well painted service ware, Capodimonte is renowned for its figures. The factory was transferred to Buen Retiro, near Madrid, in 1759, when Charles became king of Spain (see above *Spain*).

France and Belgium. The medieval pottery of France is difficult to date and classify with accuracy, but lead glaze was in common use by the 13th century at the latest. Proficient sgraffito decoration was done at Beauvaisis (Oise) and at La Chapelle-aux-Pots (Charente-Inferieure).

Palissy rustic ware

Lead-glazed wares of the 16th century. Bernard Palissy began to experiment with coloured glazes about 1539 and, after much difficulty, succeeded in producing his rustic wares in 1548. For the most part these are large dishes made with wavy centres intended to represent a stream, with realistically modelled lizards, snakes, and insects such as dragonflies grouped thereon. They are decorated on the obverse with blue, green, manganese purple, and brown glazes of excellent quality, while the back is covered with a glaze mottled in brown, blue, and purple. Palissy later turned his attention to classical and biblical subjects, which he molded in relief. After his death in 1589, work in his style was continued at the Avon pottery, near Fontainebleau.

Almost contemporary with Palissy's rustic ware is a type of pottery made in the style of the metalwork of the period. It was made at Saint-Porchaire and is sometimes called, erroneously, *Henri Deux* ware, or *faience d'Oiron.* The body is ivory white and covered with a thin glaze. Before firing, designs were impressed into the clay with metal stamps like those used by bookbinders, and the impressions were then filled with slips of contrasting colours. This technique resembles the *mishima* technique of decoration in Korea (see below *Korea*).

Faience, or tin-glazed ware. The technique and the designs of Italian maiolica influenced the development, in the early 16th century, of French faience. There were Italian potters at Lyon in 1512, and, by the end of the 16th century, painting in the manner of Urbino was well established there. Faience was also made at Rouen, probably as early as 1526, and at Nevers toward the end of the 16th century.

A new factory, established at Rouen about 1656 by Edme Poterat, introduced a decoration of *lambrequins,* ornament with a jagged or scalloped outline based on drapery, scrollwork, lacework ornament, and the like. *Lambrequins* were extremely popular and were copied at other porcelain and faience factories. The faience of Nevers, too, is extremely important and shows the Baroque style at its best. In the second half of the 17th century the porcelain of both China and Japan became increasingly well known in Europe, and many designs were borrowed from Chinese sources by potters at Nevers and elsewhere.

The factory of Moustiers in the Basses-Alpes was founded by Pierre Clérissy in 1679. During the early period frequent use was made of the engravings of Antonio Tempesta (1555–1630) as well as biblical scenes. Later came a series of dishes decorated with designs after Jean I Bérain (1637–1711), whose work greatly influenced French decorative art at the time. These designs usually include grotesques, baldacchini (canopies), vases of flowers, and the like, linked together by strapwork in a typically Baroque manner.

In 1709, when Louis XIV and his court melted down their silver to help pay for the War of the Spanish Succession, the nobility looked for a less expensive medium to replace it. In consequence, faience gained in popularity and importance. A great deal was manufactured in the region of Marseilles, the factory of the Veuve Perrin being particularly noted for overglaze painting in the Rococo style. Perhaps the most influential factory was that of Strasbourg, in Alsace (which had officially become part of France in 1697), founded by C.F. Hannong in 1709. The wares—painted in blue, in other faience colours, and in overglaze colours—were much copied elsewhere. Overglaze colours were introduced about 1740, their first recorded use in France. (For the first use in Europe, see below *Germany and Austria.*) Brilliant *indianische Blumen* (flower motifs that were really Japanese in origin but that were thought to be Indian because the decorated porcelain was imported by the East India companies) were painted in a palette that included a carmine similar to the Chinese overglaze rose ("purple of Cassius"). A characteristic copper green was also used. *Deutsche Blumen* ("German flowers") were introduced, perhaps by A.F. von Löwenfinck, about 1750, and inspired similar painting elsewhere. Figures by J.W. Lanz, who also worked in porcelain here and at Frankenthal, are to be seen. Much work was done in the fashionable Rococo style, including objects, such as clock cases and wall cisterns, and tureens in the form of fruit and vegetables. Both faience and porcelain in a variety of decorative forms were used for the banqueting table. Such table decoration, which in the 17th century had been supplied by confectioners who worked in sugar, had become very fashionable in Europe.

Faience and porcelain as table decoration

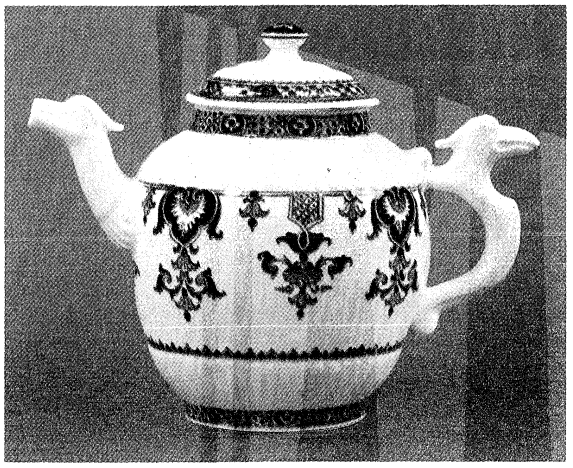
The wares of Niderviller, in Lorraine, were much influenced by those of Strasbourg. The later figures were probably modelled by the sculptor Charles Gabriel Sauvage, called Lemire (1741–1827), and some were sometimes taken from models by Paul-Louis Cyfflé (1724–1806). At Lunéville, not far away, Cyfflé worked in a pleasant but sentimental vein and used a semiporcelain biscuit body known as *terre-de-Lorraine,* which was intended to resemble the biscuit porcelain of Sèvres. The work of both Sauvage and Cyfflé is extremely skillful.

Faience was made at Tournai (now in Belgium) and at Brussels during the 17th century. Their styles were mainly derivative, but Brussels made some excellent tureens in forms such as poultry, vegetables, and fruits during the Rococo period.

After 1800 most French pottery factories concentrated on the manufacture of faience-fire (creamware).

Porcelain. In the second half of the 17th century much interest was taken in both faience and porcelain, although the technique of making soft-paste porcelain (*pâté tendre*) had yet to be mastered, and the secret of hard-paste porcelain manufacture was not discovered until the 18th century.

A factory at Saint-Cloud, founded by Pierre Chicaneau in the 1670s, made faience and a soft-paste porcelain that were yellowish in tone and heavily potted. Much use was made of molded decoration, which included sprigs of prunus blossom copied from the *blanc de Chine* of Tehua (see below *China: Ming dynasty*). Particularly common was a molded pattern of overlapping scales. Most examples are small, but there are some large jardinières (flowerpot holders) that are extremely handsome. The early painted wares were decorated in underglaze blue with typically Baroque patterns, including the *lambrequins* introduced at Rouen. Motifs derived from the



Porcelain teapot with decoration in underglaze blue, Saint-Cloud, France, *c.* 1720. In the Museum of Fine Arts, Boston. Height 13.9 cm.
By courtesy of the Museum of Fine Arts, Boston

designs of Jean Bérain are also to be seen. Polychrome specimens, some of which were decorated in the style of Kakiemon, (see below *Japan: Edo period*), date from about 1730.

At Chantilly, the first soft-paste porcelain was decorated almost entirely in the Kakiemon style, and the body was invariably covered with a tin-glaze. The Japanese period ended about 1740. For some years thereafter simple Meissen styles were copied, in particular the German flowers. In 1753 an edict in support of the newly established factory at Vincennes forbade all other factories to manufacture porcelain or to decorate faience in polychrome; much Chantilly porcelain of the later period, therefore, is creamy white, decorated only with slight flower sprigs in blue underglaze. A transparent glaze was introduced in 1751 and replaced the very unusual practice of covering porcelain with a tin-glaze.

A factory at the Rue de Charonne, in Paris, was started by François Barbin in 1735 and removed to Mennecy in 1748. The early productions were in the manner of Saint-Cloud and Rouen. Later, some excellent flower painting was done, and figure modelling was excellent in quality. Small porcelain boxes from Mennecy, often in the form of animals, are much sought in the 20th century.

Vincennes and Sèvres

The most important of the French factories was established at Vincennes about 1738 and removed to a new building at Sèvres in 1756. Louis XV was a large shareholder in the original company and the factory eventually passed to the crown in 1759. It became state property in 1793, and has so remained.

The factory did not succeed in its attempts to make a practicable soft-paste porcelain until 1745. Much of the work at Vincennes consisted of naturalistic flowers with bronze stalks and leaves, sometimes in vases elaborately mounted in gilt bronze by the court goldsmith, Claude Thomas Duplessis, and others. Meissen was also copied for a short period, but the factory soon evolved its own style, which remained partly dependent on the use of high quality gilt-bronze mounts. A few glazed and painted figures were made, but these gave place in 1751 to figures of biscuit porcelain. In 1757 the sculptor Étienne-Maurice Falconet was appointed to take charge of modelling, a position he retained until 1766. Designs by the painter François Boucher were frequently used by Falconet and others; Boucher's influence is particularly strong during the lifetime of Louis XV's mistress, Mme de Pompadour, who took much interest in the factory. Later, some excellent work in this medium was done by the sculptors Augustin Pajou and Louis-Simon Boizot.

Both at Vincennes and Sèvres much use was made of coloured grounds in conjunction with white panels, which were used for decorative painting of the highest quality. These panels were surrounded by rich and elaborate raised gilding, which was engraved and chased (toolen). The most usual ground colours were a dark underglaze blue (*gros bleu*) and a brighter, overglaze (*blue de Roi*); also used were turquoise blue, yellow, green, and *rose Pompadour* (often miscalled *rose du Barry* in England).

The porcelain of Sèvres was made to harmonize with the exotic and luxurious style of interior decoration that characterized French court circles. The soft-paste body was of superb quality; and, because the extremely fusible glaze partly remelted in the enamelling kiln, the colours sank into the glaze in a way hardly seen elsewhere.

Hard porcelain of Sèvres

The factory prosecuted the search for the ingredients of hard porcelain with vigour. They were found after a prolonged search at Saint-Yrieix-la-Perche, near Limoges, in 1769. The new body was first manufactured soon after 1770, although for some years it was only used for biscuit figures. Later, it was employed for dishes and vases decorated in a severe but luxurious classical style. In 1800 the manufacture of soft porcelain was discontinued.

A large number of smaller factories making hard porcelain sprang up, chiefly in and around Paris, in the second half of the 18th century. Some were patronized by members of the royal family, including Louis XVI's wife, Marie Antoinette. A number of provincial factories were also engaged in the same manufacture.

The Tournai factory, in Belgium, which began to make porcelain in 1751, enjoyed the patronage of the empress of Austria, Maria Theresa. Here, and in the associated factory at Saint-Amand-les-Eaux, the work of Sèvres was imitated on a considerable scale.

Germany and Austria. While Germany is principally noted for its superb porcelain, the stoneware of the Rhineland is no less noteworthy. A great deal of faience was also made, though this was less important.

The earliest type of ware made in Germany (*c.* 1350) was the *Hafnergeschirr* ("stove maker vessel"). Originally the term referred to tiles, molded in relief and usually covered with a green glaze, which were built up into the large and elaborate stoves needed to make mid-European winters tolerable. Jugs and other vessels made by these stove makers, however, came to be called Hafner ware by extension when their manufacture began about the mid-16th century. The work of Paul Preuning of Nürnberg is



Sèvres vase and cover decorated in reserved panels by Morin, France, 1780. Made for presentation to King Gustav III of Sweden. In the Victoria and Albert Museum, London. Height 49.5 cm.
By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.

Hafner ware

an example of this kind of ware. He decorated his pottery with coloured glazes kept apart by threads of clay (the cloisonné technique). In Silesian Hafner ware, on the other hand, the design is cut out with a knife, the incisions preventing the coloured glazes from mingling. The earliest German stove tiles are lead glazed. Tin glazes came into use about 1500.

After these beginnings, German pottery developed in two distinct classes: stoneware and tin-glazed earthenware.

Stoneware. The stoneware (*Steinzeug*) came mainly from the Rhineland and, in particular from Cologne, Westerwald, Siegburg, and Raeren (the latter now in Belgium). Manufacture probably began in Cologne about 1540. The body of the stoneware is extremely hard and varies from almost white (Siegburg) to bluish gray

Gerald Reitlinger, London; photograph, Wilfred Walter



Gray stoneware jug decorated with cobalt and salt glaze, from the Westerwald region, Germany, early 17th century. English silver mounts are dated 1652. Height 24.1 cm.

(Westerwald); a brown glaze over a drab body is also to be seen (Raeren). The surface is glazed with salt—no more than a smear glaze, pitted slightly, like orange peel. A smooth, though still very thin, glaze was achieved by mixing the salt with red lead. Particularly popular at Cologne in the late 16th century was the "bearded-man jug" (*Bartmannkrug*), a round-bellied jug with the mask of a bearded man applied in relief to the neck. This type was sometimes called a "Bellarmine" in England; the mask was thought to be a satire on the hated Cardinal Robert Bellarmine (Bellarmino), but there is no authority for this assumption. In England, where they were imported in large quantities, they were also known as graybeards. The term tigerware was also used for the mottled brown glaze over a grayish body.

Some of the earliest German stoneware is notable for its remarkably fine relief decoration in the Gothic style. Oak-leaf and vine-leaf motifs were common, as were coats of arms on medallions. The applied relief and stamped decoration was, at times, most elaborate, and the thin glaze lent it additional sharpness and clarity. Reliefs of biblical subjects appear on tall, tapering tankards (*Schnellen*), which were provided with pewter or silver mounts. The *Doppelfrieskrüge* were jugs with two molded friezes (usually portraying classical subjects) around the middle. They and the tankards were made in Raeren brownware by Jan Emens, surnamed Mennicken, in the last quarter of the 16th century. Emens also worked in the gray body that was used at Raeren at the turn of the century, employing blue pigment to enhance the decoration. At a later date, blue and manganese pigments were used together, and this practice continued throughout the 17th century. Figures were sometimes set in a frame reminiscent of Gothic architectural arcades, and inscriptions of one kind or another are fairly frequent.

The style of the stonewares gradually fell into line with the prevailing Baroque style, particularly toward the end of the 17th century. At Kreussen, in Bavaria, a grayish-red stoneware was covered with a brown glaze, and the molded decoration was often crudely picked out with opaque overglaze colours that had a tin-glazed base. The earliest dated specimen is 1622, which was the first time overglaze colours had been used on pottery in Europe. The technique, learned from Bohemian glass enamellers, was to have some influence in France as well as in Germany.

Germany stoneware was extremely popular abroad; during the 17th century, for instance, Sieburg even exported to Japan.

Meissen stoneware

An extremely important type of stoneware was first made shortly before 1710 at a factory at Meissen that was under the patronage of Augustus the Strong, elector of Saxony and king of Poland. It was discovered by E.W. von Tschirnhaus (1651–1708) and J.F. Böttger (1682–1719) during their researches into the secret of porcelain manufacture. It usually varies from red to dark brown and is the hardest substance of its kind known. An almost black variety was termed *Eisenporzellan* ("iron porcelain"), and a black glaze was devised by Böttger to cover specimens of defective colour. Decoration is usually effected by means of applied reliefs, although the black-glazed specimens were sometimes decorated with lacquer colours, as well as with gold and silver. Silvering was not uncommon and was also practiced in other German centres during the early part of the 18th century on both stoneware and porcelain.

A particular feature of Meissen stoneware is the incised decoration done by lapidaries on the engraving wheel. Many specimens were engraved with coats of arms, and grinding into facets (the *Muscheln* pattern) was also practiced. The same methods were used to give a plain surface a high polish. Metal mounts, common Rhenish stoneware, also were sometimes accompanied by insetting precious and semiprecious stones.

Because of the vogue for porcelain, stoneware manufacture declined and was finally abandoned about 1730.

Tin-glazed ware. Faience factories were so numerous that it is only possible to mention the most important of them. Perhaps the earliest tin-glazed wares other than stove tiles are the amusing jugs in the form of owls (with detachable heads to be used as cups) that came from Brixen (Bressanone), in the Tirol. Their shape and style no doubt inspired the later owl and bear jugs, with similarly versatile heads, made in England during the 18th century. These owl jugs (*Eulenkrüge*) were, at first, used as prizes in archery contests and were sometimes repeated in Rhenish stoneware.

The first manufacture of faience on a considerable scale took place at Nürnberg, and some dishes in the Italian style still survive. Much more is known, however, of the productions of Kreussen, which is chiefly of interest for its blue-and-white faience jugs. The outline of flowers painted in blue is almost cross-sectional in style and terminates in a small spiral—hence the name spiral family.

A factory of Hanau, near Frankfurt am Main, was started in 1661 and remained in operation until 1806. Many of the early wares were decorated with Chinese motifs. A type of jug with a long narrow neck, the *Enghalskrug*, was made in Hanau. Some have a globular body (sometimes copied in China and Japan in blue painted porcelain); others, a spirally fluted body and a twisted handle. Pewter or, less often, silver covers were common. The painting includes coats of arms, landscapes, and biblical subjects. Groups of dots amid strewn flowers (*Streublumen*) are characteristic. Realistically painted German flowers appear shortly before mid-18th century. Most painting is in blue, manganese, and the other less often used faience colours. Overglaze colours do not seem to have been used.

A factory in Frankfurt am Main itself was founded in 1666. Imitations of Chinese motifs as well as biblical subjects were very popular. The blue is brilliant, and the surface usually suggests the use of a transparent overglaze. Narrownecked jugs were commonly made and are some-

times difficult to distinguish from those of Hanau. This centre closed about 1740. At Nürnberg a later factory was established about 1712, continuing until about 1840. Most of the subjects used at Frankfurt and Hanau were repeated at Nürnberg, as well as designs based on the the Rococo engravings of J.E. Nilson (1721–88), which were also popular at many of the porcelain factories. The Rococo style, which spread from France to Germany about the second quarter of the 18th century, is reflected both in the forms and the decoration.

Wares of Bayreuth

The wares of Bayreuth are particularly interesting. Early products were painted with a misty blue, but overglaze colours were speedily adopted. "Leaf and strapwork" (*Laud-und-Bandelwerk*) was a much used type of motif, and excellent work was done by A.F. von Löwenfinck (who is known particularly for his work on porcelain) and Joseph Philipp Danhofer. Perhaps the finest 18th-century faience was made by the factory at Höchst, near Mainz, which also manufactured porcelain. Decoration was usually in overglaze colours, and landscapes, figure subjects, German flowers, and chinoiseries (European delineations of the Chinese scene with a strong element of fantasy) are of a much higher quality than elsewhere. Faience thus decorated with colours applied over the glaze, as on porcelain, was termed *Fayence-Porcellaine* during the 18th century.

An important aspect of both faience and porcelain decoration in Germany is the work of the studio painters, or *Hausmaler*, who brought undecorated faience and porcelain from the factories and painted it at home, firing the decoration in small muffle kilns. For this reason, their work was done in overglaze pigments. At first they mostly used the *Schwarzlot* technique—decoration in a black, linear style that was nearly always based on line engravings. Faience thus decorated dates from about 1660 and is the work of Johann Schaper (died 1670), who had been a Nürnberg glass painter, J.L. Faber, and others. Polychrome enamel decoration was developed by another glass painter, Abraham Helmhack (1654–1724), who mastered the technique as early as 1690, many years before it was adopted by the factories. The more important studio painters are Johann Aufenwerth and Bartholomäus Seuter of Augsburg, J.F. Metzsch of Bayreuth, the Bohemians Daniel and Ignaz Preussler, and Ignaz Bottengruber of Breslau. The work of the latter is particularly esteemed.

Toward the end of the 18th century a number of German factories, including some already making faience, made lead-glazed earthenware (*Steingut*) in imitation of Wedgwood, while a factory at Königsberg (now Kaliningrad) imitated Wedgwood's black basaltes body.

Porcelain. The earliest hard porcelain, produced by the factory at Meissen, is smoky in tone, but some improvements were made in 1715 and others in the following decade. Many early specimens were painted with a limited range of overglaze colours of good quality, including a pale violet lustre derived from gold that remained in use until about 1730. In 1720 a painter from Vienna, Johann Gregor Höroldt, was appointed chief painter (*Obermaler*) to the factory; he was responsible for introducing a new and much more brilliant palette, as well as some ground colours (*Fond-Porzellan*). The earliest ground colour to be noted is a coffee brown termed *Kapuzinerbraun,* which was invented by the kilnmaster Samuel Stölzel. The use of blue underglaze proved difficult, and little work of the kind was done. Overglaze painting, on the other hand, was of fine quality and includes topographical subjects, figure subjects based either on harlequin, pierrot, and other characters of the Italian comedy or on the style of the painter Jean-Antoine Watteau and his followers, and flowers in the Oriental style (called *indianische Blumen*) as well as native flowers (*deutsche Blumen*) taken from books of botanical illustrations. A series of harbour scenes from engravings of Italian ports were mostly executed by C.F. Herold (cousin to the *Obermaler*) and J.G. Heintze. Perhaps the most important early wares are the chinoiseries, which appear in great variety. The first work of the kind, much of it painted by the *Hausmaler* Bartholomäus Seuter, is in gold silhouette followed by polychrome painting after designs by the *Obermaler.* The figures are painted in three-quarter length. *Indianische Blumen* motifs were used, and Arita decorations, particularly those of Kakiemon (see below *Japan: Edo period*), were closely copied.

Figure modelling

Little figure modelling was done until about 1727, when the sculptor Johann Gottlob Kirchner was appointed *Modellmeister* and asked to make some colossal figures of animals for the Japanische Palais, the building that housed Augustus the Strong's porcelain collection. Because the medium was unsuited to work of this kind, most of the surviving examples are spectacular and magnificent failures. After the death of Augustus the Strong in 1733 large-scale modelling was practically discontinued, and the new *Modellmeister,* Johann Joachim Kändler, turned his attention to small figures suitable for decorating the dining table.

Assisted by other modellers, Kändler soon made the figures of Meissen fashionable throughout Europe. The first important Rococo work in porcelain appears in Saxony after 1737 when Kändler started to make the Swan service—perhaps the best known of all porcelain services. It is decorated with such motifs as swans, nereides, and tritons. Rococo Meissen was widely sought.

Meissen was the most influential European factory until the beginning of the Seven Years' War in 1756, when it was taken by the Prussians. From then until 1763 it was operated by nominees of Frederick the Great, who virtually looted the factory. By the end of the war, leadership had passed to Sèvres, and the work of Meissen for the next 50 years is much less important than formerly. The transitional Louis XVI style of c. 1763–74, the Neoclassical style was increasingly used, and is typified by the figure modelling of Michel Victor Acier, who came to the factory to share the position of *Modellmeister* with Kändler in 1764. From 1774–1814 the designs of Sèvres and of Wedgwood (Wedgwoodarbeit) were copied.

Few marks have been so consistently abused as that of the crossed swords of Meissen. Since the 18th century, it has been added to all kinds of unlikely specimens.

The other German factories of the period were, for the most part, established with the aid of runaway workmen from Meissen and Vienna, where Claudius Innocentius du Paquier had started a factory in 1719 with the aid of two men who were themselves from Meissen. Early Vienna hard-porcelain wares are highly prized. Much use was made of leaf and strapwork patterns, and excellent work was done in black monochrome (*Schwarzlot*). The factory passed to the state in 1744, and its later work is competent without being distinguished. Between 1784 and 1805 it became noted for elaborate gilding and coloured grounds, with minutely detailed painting, after Angelica Kauffmann and others, in reserved white medallions.

The Vienna factory provided a number of wandering arcanists (men who possessed the *arcanum,* or "secret," of hard-porcelain manufacture), two of whom helped to establish the Höchst factory, which began manufacture about 1752. This factory is principally noted for excellent figures in the Neoclassical style by Johann Peter Melchior and for the work of Simon Feilner.

A factory in Berlin, started in 1761 and acquired by Frederick in 1763 when he relinquished his hold on Meissen, produced wares with painted decoration of high quality. The decoration made much use of mosaic patterns—detailed diapers (small repeated motifs connecting with one another or growing out of one another with continuously flowing or straight lines) painted over a coloured ground. A large service made in 1819 for presentation to the Duke of Wellington and decorated with scenes from his battles is now in Apsley House, London.

Nymphenburg figures

There is much interest in the figure modelling of Franz Anton Bustelli, who worked at Nymphenburg, a suburb of Munich. The factory, which is still in operation, was started about 1753. Bustelli became *Modellmeister* in 1754 and retained the position until his death in 1763. His magnificent series of figures based on the Italian comedy are the most important expression of Rococo in German porcelain. The painted wares of the factory were also of fine quality.

Some excellent figures were made at Fürstenberg, where hard porcelain was first manufactured in 1753, and at Frankenthal by such notable modellers as J.W. Lanz, the cousins J.F. and K.G. Lücke, and Konrad Linck. Ludwigsburg, started in 1758, produced porcelain that was grayish in colour and more suitable for figure modelling than for service ware. The figures of artisans by an artist known as the Modeller der Volkstypen (modeller of folk-types) are original and pleasing, and the sculptor, Wilhelm Beyer, did good work in the Neoclassical style.

The Netherlands. During the 17th century, red stoneware was made by Ary de Milde of Delft and others in imitation of the wares of I-hsing (see below *China: Ming dynasty*). Creamware was manufactured at several places at the end of the 18th century. Most Dutch pottery of the period, however, is tin glazed.

Italian potters had settled in Antwerp by 1525, and surviving examples of tin-glazed ware from this period are in the Italian style. Manufacture was concentrated to a great extent in Delft soon after the beginning of the 17th century. By about 1650 the large brewing industry began to decline, and the old buildings were taken over by potters who retained such names as The Three Golden Ash-Barrels, The Four Roman Heroes, and The Double Jug for their potteries. The craftworkers of the town were organized into the Guild of St. Luke, which exercised a considerable amount of control over apprenticeships and established a school of design.

In the 17th century the Dutch East India Company, chartered in 1602, imported Chinese and Japanese wares in great quantities, and the taste for Eastern decoration rapidly ousted Italian fashions. For the greater part of the 17th century decoration was in blue, and Chinese porcelain was closely imitated. In wares of the best quality this imitation is so exact that, without a fairly close inspection, it is possible to mistake them for the originals. Western decorations—biblical and genre scenes, landscapes and seascapes—were carried out in styles similar to Dutch paintings of the period. Tilework was frequently undertaken; many individual tiles have survived, although large panels made up of many tiles are very rarely complete. Blue painting was followed by the use of the usual underglaze faience colours, the outline (known as *trek*) being first drawn with blue or manganese and then filled in. Before firing, the object was covered with an additional transparent lead glaze known as *kwaart,* which made the surface more brilliant. Red was a difficult colour; often when it was to be used, an unpainted space was left during the first firing, and the red was applied afterward and fired at a lower temperature. Gilding is found on the finer specimens and required a further firing. Overglaze colours were introduced by Zacharias Dextra about 1720, and the Chinese *famille rose* patterns were frequently imitated. Among the rarer and more showy examples of delft may be numbered the *Delft dorée,* on which gilding is lavish, and the *Delft noir,* which has a black ground (suggested by Chinese lacquer work) in conjunction with polychrome decoration. Work of this kind is often attributed to Adriaen Pijnacker.

Delft tiles

Marks on Dutch delft are extremely unreliable, for many later copies were given the earlier marks of important potteries, especially during the 19th century.

Britain. The medieval pottery of England was affected little by outside influences. Moreover, poor communications prevented the industry from concentrating in any one place; most wares, therefore, are made of local clay by local craftsmen. The potters worked alone or in extremely small groups, and their tools were few and simple. The clay used for the body ranges from buff to red, or, when fired in a reducing atmosphere, from gray to almost black. As with much Japanese pottery, little effort was made to disguise the method by which the vessel was formed, so that pronounced ridges are frequently visible. Both relief and inlaid decoration are found, especially on tiles, and brushed slip was also used to add simple patterns.

Unglazed ware was common, especially in the early period, but a soft lead glaze came into more general use later, the knowledge probably being derived from France. The early glaze varied between yellow and brown according to the iron content of the clay, although a group having a particularly rich brown glaze was made by first washing the pot with slip containing manganese. The use of copper oxide to give a rich green of variable colour dates from the 13th century. During that period, the green, buff, and brown glazes were used in conjunction. Cistercian wares, made in the monasteries before their dissolution in 1536–39, are more precisely finished. They have a dark-brown glaze over a stoneware body and are sometimes decorated with white slip or incised. By far the greatest number of surviving specimens are jugs and vessels for storing liquids; since they have almost always been excavated, a reasonably perfect specimen is a rarity.

Tin-glazed ware. Lead-glazed wares tended to die out after tin glaze reached England via the Netherlands about 1550. At first it was called gallyware, but, with the rise of the Dutch manufacturing centre at Delft, the ware came to be called delft. Its popularity was due to the fact that it could be painted in bright colours. The earliest surviving examples are the Malling jugs, so called because an early specimen of the kind was preserved in the church at West Malling, Kent. These were almost certainly made in London. The colour varies from turquoise to black; a variety with a blue ground flecked with orange was probably suggested by the tigerware from the Rhineland. The jugs usually have silver or pewter mounts. Similar mounts, often of English manufacture, are to be seen on Rhenish jugs imported into England and occasionally on Turkish jugs of about the same period.

Malling jugs

By 1628 a flourishing factory had been established at Southwark, London. Influenced by some Chinese blue-and-white porcelain of the Ming reign of Wan-li (1573–1620), some surviving specimens are decorated in blue, with birds amid floral and foliate motifs. Almost contemporary are some large dishes painted in polychrome colours. The earliest (1600), which is in the London Museum, bears the following couplet: "The rose is red the leaves are grene/God Save Elizabeth our Queene." The dish has a border of blue dashes and is a forerunner of the so-called blue-dash chargers that were popular later in the century. These were decorated with biblical scenes (Adam and Eve being a special favourite), crude portraits of the kings of England, ships, armorial bear-

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.



Tin-glazed earthenware wall tile with medallion portraits of William and Mary, Delft, "Greek 'A'," factory of Adrianus Kocks, c. 1694. William III of England commissioned these large tiles for Queen Mary's dairy at Hampton Court. In the Victoria and Albert Museum, London. Height 62 cm.

ings, and the like. The influence of Italian maiolica and Chinese porcelain can be seen in the border designs.

Many wine bottles are extant, often with the name of the wine (Sack, Claret, etc.) painted in blue and a date. Others are more elaborately decorated, a few are in polychrome.

Toward the end of the 17th century service ware became more frequent (although tea ware was now scarce). Blue-and-white was still made in large quantities, but a polychrome palette was more in evidence, and the influence of Dutch potters is often obvious.

Chinese influence, which had been particularly strong in the early part of the 18th century, tended to persist, particularly at Bristol. The Rococo style was used to a limited extent. Later, some not very successful attempts were made to utilize The Neoclassical forms. Overglaze colours on tin-glazed wares appear after mid-century. These colours, a special pallet now called Fazackerly colours, were probably used only at Liverpool.

The main centres of production of tin-glazed ware were in London (Southwark and Lambeth), Bristol, and Liverpool, although there were smaller potteries elsewhere. One of them—Wincanton in Somerset—made frequent use of manganese, which produces purple and purplish-black colours. The tin glaze fell into disuse about the turn of the 18th century, its place having been taken by Wedgwood's creamware. (In the mid-20th century manufacture has been successfully revived at Rye, Sussex.)

17th-century slipware. Wares decorated with dotted and trailed slip were made at Wrotham, Kent, and in London during the first half of the 17th century. Wrotham is noted principally for drinking mugs with two or more handles, known as tygs; and London for dishes with such pious exhortations as "Fast and Pray," obviously inspired by the Puritans. Manufacture was also started in Staffordshire, and many surviving examples were signed by the potter in slip. The work of Thomas Toft is particularly valued. The best work of this kind was done before the end of the 17th century, and although it may fairly be described as peasant ware, many of the earlier specimens are vigorously decorated and amusing. Manufacture continued until the end of the 18th century.

Stoneware. The popularity of Rhineland stonewares in England, as well as that of the newly imported Chinese stoneware teapots from I-hsing kilns (see below *China*: *Ming dynasty*), led to attempts to imitate both kinds. The first patent for making copies of porcelain and Cologne ware known to have been exercised was awarded to John Dwight (*c.* 1637–1703) of Fulham in 1671. In addition to German stoneware, he made a brown-glazed stoneware decorated with stamped ornament that was continued at Fulham after his death and has been extensively reproduced since. He probably never made any porcelain, but he mentions red china, which can only refer to imitations of the I-hsing stoneware.

The brothers John Philip and David Elers, of German origin, made red stoneware at a factory in Staffordshire. It is difficult to separate their work from that of Dwight (at Fulham), on the one hand, and that of their Staffordshire imitators, on the other. Most wares are decorated with stamped reliefs, the Chinese prunus blossom being comparatively common. The tendency to utilize patterns from silverwork, which is apparent on some examples, may be connected with the fact that the Elers had been silversmiths. The Elers' migration to Staffordshire perhaps can be regarded as the starting point for the large modern industry that has grown up in that area. Certainly from this time onward Staffordshire wares tend to lose their peasant character and to approach a factory-made precision that was to be general by the end of the 18th century.

Salt-glazed stoneware

The earlier red- and brown-glazed stonewares were replaced about 1690 by a salt-glazed stoneware that was regarded as an acceptable substitute for porcelain. It varies in colour from drab to off-white, the glaze on later specimens often having a richer, more glassy appearance due to the addition of red lead to the salt. One of the earliest varieties is decorated with reliefs stamped from pads of clay that were applied to the surface.

18th-century developments. The "scratched-blue" class of white stoneware dates from about 1730 and is decorated with incised (sgraffito) patterns, usually touched with blue. Decoration is floral, and inscriptions and dates are fairly frequent. Its manufacture continued until about 1775.

From the 1730s molded patterns in relief were popular, the clay being pressed into molds of metal, wood, or fired clay. The introduction of plaster of paris molds around 1745 gave much greater scope and led to the development of intricate shapes in the finer varieties of white stoneware. The patterns greatly increased in sharpness and elaborate piercing is to be seen.

Transfer printing was first used about 1755, possibly at Liverpool, which produced wares of all kinds, including tiles, using this decorative technique.

The earliest use of overglaze colours belongs to the same period—previously, white wares had been sent to Holland for decoration. The Englishman who first mastered the technique was William Duesbury. Established as a decorator in London by 1751, he concentrated on painting porcelain, but he also seems to have overglaze-painted stoneware from Staffordshire. Some extant brilliantly painted figures are probably from his studio. A little earlier than Duesbury's overglaze-painted figures are the uncoloured pew groups, which consist of two or three figures seated on a high-backed settle or pew, modelled in a primitive and amusing fashion. A rich blue overglaze ground, often called Littler's blue after William Littler, who is thought to have invented it, was much used on the salt-glazed stoneware, as well as the porcelain, made at Longton Hall, a factory that operated in Staffordshire from about 1750 to 1760 and that was also associated with Sittler.

John Astbury is particularly associated with a type of brown-glazed ware decorated with stamped pads of white clay. Some of the earliest Staffordshire figures in brown and white clay covered with a lead glaze have been attributed to Astbury.

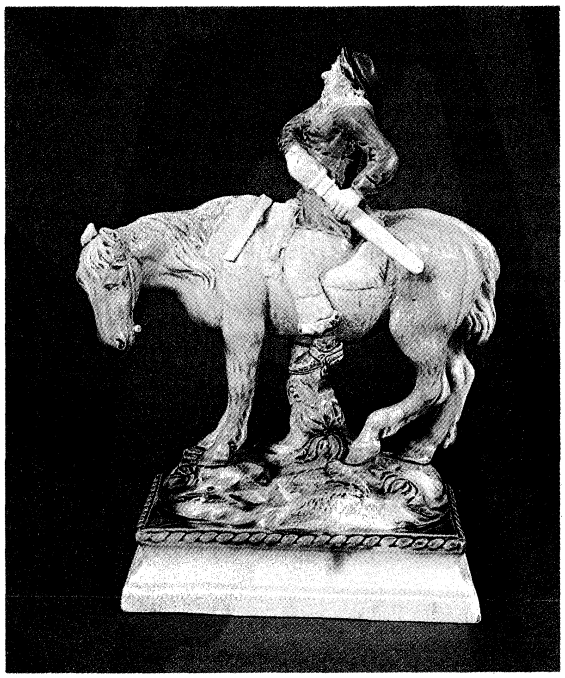
Agateware

Thomas Whieldon (1719–95) of Fenton Low, Staffordshire, manufactured agateware—that is, ware made by combining differently coloured clays or by combing together different colours of slip. In the former method the clays were usually laid in slabs, one on the other, and beaten out to form a homogeneous mass in which the colours were inextricably mingled. Agatewares seem to have been made in Staffordshire between 1725 and 1750, the earlier specimens being salt glazed, while the later ones were covered with a colourless lead glaze. Whieldon is most famous for his use of coloured glazes that were mingled to give a clouded or tortoiseshell effect and were used on an earthenware body, sometimes over molded decoration. A few naïvely modelled figures with this type of glaze are attributed to him. From 1754 to 1759 he was in partnership with Josiah Wedgwood, who developed the fine green and yellow glazes to decorate molded wares in the form of pineapples, cauliflowers, and the like.

Coloured glazes were also used by Ralph Wood I (1715–72) of Burslem, Staffordshire, for decorating an excellently modelled series of figures in a creamware (lead-glazed earthenware) body, the finest, perhaps, a mounted Hudibras in the Victoria and Albert Museum. Many of these figures are attributed to the modeller Jean Voyez, who was much influenced by the work of P.L. Cyfflé at Lunéville (see above *France*). Ralph Wood I is also noted for the typical English Toby jug (first made soon after 1700), which is a beer jug in the form of a man, usually seated and holding a pipe and a mug, the hat (where present) forming a detachable lid. Very popular, they continued in production for many years. Enoch Wood, another member of the family, joined Ralph Wood II in partnership as Enoch Wood & Co., which lasted until 1790. They made most of the wares current in Staffordshire at the time, as well as some excellent figures decorated with overglaze colours.

Work of Wedgwood

Josiah Wedgwood (1730–95), the most famous of all the Staffordshire potters and the most important exponent of Neoclassicism in the field of pottery, is chiefly celebrated for his fine jasper and black basaltes stone-



Mounted Hudibras, creamware decorated with coloured glazes by Ralph Wood, Staffordshire, *c.* 1765. In the Victoria and Albert Museum, London. Height 29.8 cm.

By courtesy of the Victoria and Albert Museum, London; photograph, Wilfrid Walter

wares, but his creamware was undoubtedly the more influential in the 18th century. It was well finished and clean in appearance, with simple decoration in good taste, often in the popular Neoclassical style. His wares appealed particularly to the rising bourgeois class, both in England and abroad, and porcelain and faience factories suffered severely from competition with him. Surviving factories switched to the manufacture of creamware (*faience fine* or *faience anglaise*), and the use of tin glaze almost died out.

Wedgwood secured the patronage of Queen Charlotte (wife of King George III) for his creamware in 1765 and renamed it Queen's ware. Much of it was transfer printed by John Sadler and Guy Green at Liverpool. Evidence of its popularity and importance is provided by the enormous service of 952 pieces made for Catherine the Great's palace of La Grenouillière, in St. Petersburg.

The basaltes ware, also called black porcelain or Egyptian ware, was a type of stoneware introduced about 1768. Like the jasper that followed, it was used almost entirely for ornamental work—vases, ewers, candlesticks, plaques, medallions, and tea and coffee ware. Some of it was painted in what Wedgwood called encaustic enamel in imitation of Greek red- and black-figure vases, but most of the decoration was either molded and applied or incised by turning on a lathe.

Jasperware

Jasper, introduced about 1775, is a fine-grained white unglazed stoneware, slightly translucent when thinly potted or fired above the normal temperature. Undoubtedly it was inspired by the biscuit porcelain of Sèvres. Its name derives from the fact that it resembles the natural stone in hardness. At first the body was stained blue (with applied decoration in white). Other colours, such as sage green, lilac, black, and yellow, followed speedily. Like basaltes, jasper was used mainly for ornamental wares, but perhaps the most interesting products are the portrait medallions of contemporary notables. Vases do not appear to have been made until after 1780. In 1790 Wedgwood produced the first copies of the Portland vase, a magnificent Roman cameo glass vase of dark-blue glass decorated with white figures, at that time owned by the Duke of Portland but now in the British Museum. The vase was reproduced in later years, particularly in Victorian times both by Wedgwood in jasper and by Northwood in glass. Wedgwood's jasperwares were imitated in biscuit porcelain at Sèvres, and Meissen produced a glazed version called Wedgwoodarbeiten. Less influential was the red stoneware (*rosso antico*), which sometimes had an enamelled decoration of classical subjects, and caneware, a buff stoneware.

Lustre pigments introduced into England toward the end of the 18th century, were used in a manner quite different from the earlier styles of other countries (see below *Spain and Islāmic pottery*). To simulate silverwork, wares were completely covered with platinum lustre, which remains unchanged in colour after firing (silver itself yields a pale straw colour); the amount of metal used was extremely small. Such wares were known as poor man's silver. Wares were also painted or stencilled with lustre patterns. The most valuable type commercially were the resist lustres, which have a lustred background and the pattern reserved in white. They were made by painting or stencilling the pattern on the glaze with shellac, which resisted the subsequent application of the metallic pigment. Silver lustre was rarely used, but gold lustre, which gives variable colours from pink to purple, was fairly common. (Copper, whose colour remains more or less unchanged in its lustre form, was used throughout the 19th century for common wares.)

Poor man's silver

Porcelain. A factory for porcelain manufacture, using a soft-paste body similar to that of Saint-Cloud, was established in Chelsea, London, about 1743 by Charles Gouyn and Nicolas Sprimont, the latter a silversmith. The rare surviving specimens include jugs molded in the form of goats and further decorated with an applied bee, obviously based on a silver prototype that no longer exists. (Extant examples of the latter are 19th-century forgeries.) These goat and bee jugs are often marked with an incised triangle, which was then the mark in use. About 1750 a new body was adopted, together with the familiar mark of an anchor, which was raised on a small medallion until about 1752, painted in red until about 1756, and executed in gold thereafter. The work of the Chelsea factory was extensively influenced by Meissen until about 1756, the styles of Sèvres superseding it in the gold-anchor period. Wares marked either with the raised or the red anchor are the most highly valued; the painting of these is excellent in quality. Some of the best wares were painted by an Irish miniaturist, Jeffrey Hamet O'Neal. The gold-anchor-marked wares are noted for rich gilding and some



Jasperware vase moulded with "the crowning of a kitarist" in white relief against a pale blue background, impressed Wedgwood, 1786. In the British Museum. Height 43.2 cm.

By courtesy of the trustees of the British Museum

fine coloured grounds that, on occasion, rivalled those of Sèvres. The figures in the later Rococo style are generally inferior to those of the earlier red anchor period. Some Chelsea porcelain from 1760 onward was painted in the

By courtesy of the Metropolitan Museum of Art, New York, collection of Irwin Untermyer



"The Music Lesson," enamelled soft-paste porcelain, English, Chelsea, gold anchor period, *c.* 1765. In the Metropolitan Museum of Art, New York. Height 38.1 cm.

studio of James Giles of Clerkenwell. The factory was bought by William Duesbury of Derby (see below) in 1770 and entered a phase known as the Chelsea-Derby period. The Neoclassical style was introduced together with the figure in biscuit porcelain made fashionable by Sèvres. It closed finally in 1784.

A group of figures, the best known examples of which are those portraying a girl in a swing, were made in the 1750s—possibly at Chelsea but more probably at a short-lived factory staffed by workmen who had seceded from Chelsea. A class of figures characterized by an apparent retraction of the glaze from the base—dry-edged figures—are attributed to a factory established at Derby about 1750. This enterprise apparently petered out and another factory in Derby was started in 1756 by Duesbury (who was later to buy the Chelsea factory). It advertised itself as the second Dresden and is noted toward the end of the century for the excellence of its painting by Zachariah Borman, William Billingsley, and others.

Longton Hall in Staffordshire made figures and a good deal of service ware molded in the form of leaves. A rich blue ground (Littler's blue) was used on porcelain and salt-glazed wares alike. Its wares are rare and much sought.

The Bow factory (London) was started as early as 1744 with the aid of clay brought from Virginia by the American settler Andrew (André) Duché, who had discovered the secret of manufacture quite independently some years before (see below *Colonial America*). An amusing and primitive class of Bow figures was executed by an anonymous artist known as the Muses Modeller, because the most typical figures portray the Muses. Generally speaking, Bow wares are unsophisticated, and the factory obviously catered to prosperous tradesmen, a market ignored by Chelsea. An important technical innovation took place at Bow in 1750, when calcined bones were added to the porcelain body. This was the first major departure from the French soft-porcelain formula, which was fundamentally a mixture of clay and ground glass. Bone ash was added to soft porcelain by Chelsea about 1755, by Lowestoft (which mainly copied Bow styles) in 1758, and by Duesbury to Derby porcelain in 1770, when he purchased the Chelsea factory. About 1800 at his factory at Stoke-upon-Trent, Staffordshire, Josiah Spode the Second added calcined bones to the hard-porcelain formula to produce the standard English bone-china body (see below *19th century*).

Development of English bone china

Another variation on the original soft-porcelain body was introduced at a factory in Bristol started by Benjamin Lund about 1748. Clay was mixed with a fusible rock called steatite (hydrous magnesium silicate), the principle being similar to that used in the manufacture of hard porcelain. This factory was transfered to Worcester in 1752 and still manufactures fine porcelain. In the 18th century, scale grounds, which consisted of patterns of overlapping scales in various colours, were particularly popular. Transfers taken from engraved plates were also extensively used for decoration. After 1783 wares show a progressive decline in taste. A second factory was established at Worcester by Robert Chamberlain in 1786 (see below *19th century*).

William Cookworthy discovered the secret of hard porcelain independently after many years of experiment. In 1768 he opened a factory at Plymouth (which was transferred to Bristol in 1770) that made figures in the style of Bow and Longton Hall. Richard Champion acquired the patent for hard porcelain in 1772 and manufactured tableware Neoclassical in style and excellent in quality. The patent was bought by a syndicate that established a factory at New Hall, Staffordshire, in 1782 and made a humble variety of wares for about 40 years.

Scandinavia. The faience industry spread to Scandinavia mainly because of migratory workmen from Germany. A number of factories in Denmark, Norway, and Sweden during the 18th century made faience and creamware in the English manner. A distinctive Scandinavian production was that of bowls, made in the shape of a mitre, for a kind of punch called bishop. The most important factories are those of Rörstrand and Marieberg (Koja) in Sweden. A typical Rococo concept to come from Marieberg is a vase standing at the top of a winding flight of steps. Called a terrace vase it is often decorated with a rabbit or some other animal.

In 1774 a factory at Copenhagen directed by Louis Fournier, a modeller from Vincennes and Chantilly, began the manufacture of true porcelain. The factory was acquired in 1779 by King Christian VII of Denmark and Norway. In 1789 the factory started work on an enormous service, originally intended for Catherine the Great, each piece of which was painted with a detailed picture of a Danish flower. This service, the "Flora Danica," is now in Rosenborg Palace, Copenhagen. Numerous skillfully made figures were also produced. The factory continues to produce fine porcelain.

Service for Catherine the Great

Switzerland and Russia. A factory started near Zürich in 1763 and directed by Adam Spengler made both faience and porcelain and, after 1790, creamware. Delicate figures, some modelled by J.V. Sonnenschein from Ludwigsburg, and good-quality service ware were produced.

The factory of St. Petersburg was established about 1745. Later production was on a fairly large scale, and the work of Sèvres and Meissen was freely copied. Some good original work was also done, and well-modelled figures of Russian peasants were made toward the end of the century. Even better figures were made at a factory in Moscow founded about 1765 by an Englishman named Francis Gardner. Many factories at Moscow and elsewhere in Russia were established during the 19th century.

Colonial America. There is little detailed information about the pottery made by the early European settlers in North America. Most of it was manufactured locally for local needs and from the clays that were nearest to hand. Since most of these contained iron in varying quantities, the pottery body burned to colours between buff and red. Until kilns capable of reaching a high temperature were constructed, manufacture was limited to earthenware. Lead glazes were commonly used. Slips, both as a wash and as trailed decoration, were employed, and sgraffito decoration is known. Most of this pottery

was made for practical rather than decorative purposes. A few potteries were established in the 17th century in Virginia, Massachusetts, and New Jersey; and in eastern Pennsylvania, German settlers started work as early as 1735 making slip-painted and sgraffito earthenware in their own traditions.

Perhaps the most important development in colonial America took place in Savannah, Georgia, where Andrew Duché started a pottery about 1730. He interested himself in the manufacture of porcelain and discovered the china clay and feldspathic rock necessary to its manufacture. By 1741 he appears to have made a successful true porcelain but failed to gain adequate financial assistance to develop it. He therefore travelled to London, arriving in 1744, and tried to sell the secret to the founders of the Bow factory in London. Their interest is certain, since the patent specification subsequently filed specifically mentions *unaker*, said to be the Cherokee name for china clay. Duché returned to Virginia by way of Plymouth and there spoke with William Cookworthy, later to be the first manufacturer of true porcelain in England. It is still not known to what extent Duché actually manufactured porcelain; but since the *Bristol Journal* for November 24, 1764, refers to the import of some specimens of porcelain said to have been made in Georgia there is little doubt that the first porcelain to be made in an English-speaking country came from North America. The Cherokee clay was shipped to England from time to time during the 18th century. Wedgwood imported several tons of it to use in the development of the jasper body.

Hard porcelain in America

By 1765 potteries were being established on a sufficient scale to warrant an attempt to recruit workmen from Staffordshire. Wedgwood wrote at the time: "They had a agent amongst us hiring a number of our hands for establishment of new Pottworks in South Carolina."

The manufacture of tin-glazed ware began in Mexico soon after the Spanish Conquest in the first half of the 16th century. Spanish styles predominated, especially that of Talavera, but Chinese influence occurs in the 18th century. The wares became a kind of inspired folk pottery in the 19th century.

19TH CENTURY

There is a fundamental difference between work done before the Industrial Revolution, the effect of which began to be felt in the pottery industry before 1800, and that done subsequently. A student of the older wares, particularly those of the East, may find much of the later work difficult to accept because of its machine finish. When an object is made by hand it is never exactly the same as any other object, nor are the processes by which it has been formed and decorated disguised. Consider, for example, a Sung dynasty pot or a specimen of Japanese tea ceremony ware, whose imperfections of finish by factory standards are an integral part of their beauty and character, or the glaze of a *Kuan* vase, which would lose its individuality if it possessed the smooth finish of a factory-made specimen. The technical precision of the 19th century, which made its products indistinguishable from one another, and the careful concealment of the means by which the end had been achieved, were both unprecedented and deleterious. Style and craftsmanship degenerated steadily in the factories. The situation was aggravated by the Great Exhibition of 1851, which encouraged manufacturers throughout Europe to vie with each other in producing wares displaying virtuosity unhampered by questions of taste; for example, from as far afield as St. Petersburg, hitherto outside the mainstream of European development, came some particularly colossal and hideous vases in a debased Neoclassical style—which were described by a contemporary writer as "second to few of the productions of Dresden and Sèvres for beauty of outline and perfection of finish."

Those who bought these wares—as well as those who produced them—contributed to the degeneration of taste. Before the advent of mass communications in the 20th century, new fashions originated in the wealthiest stratum of society (which was usually also the most cultivated) and filtered downward. As a result of the political and economic effects of the Seven Years' War (1756–63), combined with the beginning of the Industrial Revolution, the European bourgeoisie prospered, and their wealth enabled them to become patrons and arbiters of taste. Primarily interested in the arts as a method of display or as status symbols, they demanded an excess of intricate and expensive ornament. In the Far East the same process of degeneration began at the same time, at least partly as a result of the large number of export orders received. This pernicious influence was kept at bay for awhile by the emperor Ch'ien-lung, who stigmatized the English as cultural barbarians, but became much more pronounced in the 19th century. Similar tendencies may be seen in Japanese pottery after 1853, when many factories worked almost entirely in styles demanded by their customers in the West.

Britain. *Porcelain.* The Neoclassical style, which had been popular during the middle years of the 18th century, gradually lost its earlier simplicity. In France, the rise of Napoleon brought in its train the ostentatious Empire style (copied, for the most part, from the decorative art of imperial Rome), which had much influence in England during the Regency period (1811–20). It is noticeable on the porcelain vases made at such factories as Worcester, Derby, and Rockingham. They were often decorated with well-painted topographical subjects that were no longer confined by frames but ran around the vase as a continuous landscape. Flower painting was often of excellent quality and was much influenced by the work of William Billingsley, a flower painter who worked at Derby toward the end of the 18th century.

At Worcester a factory established by Robert Chamberlain in 1786 produced porcelain decorated in a debased Japanese style. Because of their gaudy colour—iron red and underglaze blue coupled with lavish gilding—some Japan patterns are called thunder-and-lightning patterns. Similar Japan patterns were being employed at Derby and at an older Worcester factory, although much of the work of the latter was more restrained. Some of the best painting at the old factory was executed by Thomas Baxter, who used marine shells as a subject.

It has been said, unfairly, that Josiah Wedgwood, by developing the factory system, was largely responsible for the degradation of the pottery art; but Wedgwood wares have usually been in good taste even if they have not always been particularly adventurous. A far more malign influence was that of John Rose of Coalport (Salop). Rose admired the work of Sèvres and imitated it, buying or borrowing specimens to copy and using such ground colours as the *rose Pompadour.* He was one of the first English exponents of the revived Rococo style, which appeared about 1825, and made much porcelain encrusted with applied flowers. His work has been erroneously regarded as a close copy of old Sèvres. Coalport flower painting, however, is very fine in quality and much in the style of Billingsley, who actually worked at the factory for some years.

Spode and Minton

Josiah Spode the Second, who invented the standard English bone china about 1800, at first made good use of it. Some of his later wares, however, became increasingly pretentious copies of French styles, with highly coloured grounds, lavish gilding, and an excess of applied ornament. The factory was taken over in 1833 by William Taylor Copeland, Jr. and was amalgamated with Wedgwood in 1964.

The firm of Minton's was founded at Stoke-upon-Trent in 1793 by Thomas Minton, a Caughley engraver said to have devised for Spode the Broseley Blue Dragon and Willow patterns that are still in use. Like Coalport, the factory was much occupied in copying the work of Sèvres. From 1848 to 1895 they employed a Frenchman, Joseph-François-Léon Arnoux, as art director, and under his tutelage French artists were brought to England; for example, the sculptor Albert Carrier-Belleuse and also Marc-Louis Solon, who was responsible for introducing *pâte-sur-pâte* decoration into England (see below *The European continent*).

The Derby tradition of fine painting was carried into the

19th century, during which time the flower designs became somewhat overblown, although landscapes remained on a high level. The sets of so-called Campaña vases (more properly Campagna), distantly derived from Italianate copies of the Greek krater, were often decorated with landscapes by the brothers Robert and John Brewer and others. The Brewers were pupils of the topographical painter Paul Sandby.

About 1840 Parian ware, an imitation of Sèvres biscuit porcelain, was introduced by Copeland & Garrett (formerly Spode), and a great many figures, some of them extremely large, were made in this medium. Most of them are either sentimental subjects or quasi-erotic nudes, which were popular subjects of Victorian art. Parian ware had some success in America, where it was manufactured by Norton and Fenton.

Stoneware and earthenware. Production of earthenware and stoneware for the cheaper market continued on an ever-increasing scale. Lustre decoration, which had been revived in the preceding century, was used more frequently than before. A type of stoneware obviously inspired by Wedgwood's jasperware was made at Castleford, Yorkshire. Ironstone china, a type of opaque stoneware sometimes called opaque porcelain, was introduced early in the 19th century. Pseudo-Chinese and Japan patterns were frequently used to decorate it.

Ironstone china

By 1830 new underglaze colours had been pressed into service for transfer printing. These new colours were particularly used by Ridgway & Co. of Hanley, Staffordshire. Transfer-printed earthenware in blue, which became increasingly popular after 1810, was soon being produced in enormous quantities. It was much used by Spode, who often employed American subjects for wares exported to the United States. Polychrome transfer printing, essayed tentatively at Liverpool during the 1760s, was also mastered.

Earthenware figures were made in large quantities in Staffordshire and elsewhere, the best associated with Enoch Wood. They were intended as chimney ornaments, and the subjects range from bullbaiting to sentimental shepherdesses. Many of them are copied more or less directly from Derby porcelain figures, and they are a sad but accurate reflection of the times during which they were made.

The Great Exhibition of 1851 completed the degeneration started by the revival of the Rococo style. Technical progress allowed the manufacturers ever-increasing elaborations with which they bludgeoned the few remaining sensibilities of their customers. Past styles were indiscriminately and ignorantly copied. Minton's, for example, made an earthenware decorated with coloured glazes that they miscalled maiolica. It was used not only for decorative wares but for domestic articles—such as umbrella stands—and for architectural purposes.

The Paris exhibitions of 1867 and 1878 brought Japanese pottery and porcelain once more to the attention of European manufacturers, but it was not the superb porcelain of Arita that had had so much influence in the previous century. This time the Japanese exported cream-coloured earthenware with a closely crackled surface and lavish painting of poor quality, judging that it would appeal to Western taste. It became extremely popular under the name of Satsuma and was copied avidly at Worcester and elsewhere (see below *Japan: 19th and 20th centuries*).

By 1860 a few people had become profoundly disturbed by the level to which popular taste had sunk. Among them was the English poet and designer William Morris, who founded a firm of interior decorators and manufacturers in 1861. One of his pupils, William de Morgan, started a pottery at Fulham (London) in 1888 that made dishes and tiles inspired by Persian, Hispano-Moresque, and Italian wares. De Morgan used brilliant blues and greens and a coppery red lustre. His designs are a great improvement on those of the factories, although they, too, are derivative.

After about 1860 Doultons of Lambeth (London) copied 18th-century brown stoneware, making small figures and repeating earlier designs. The incised decoration by Hannah Barlow is both pleasant and competent. From a Fulham pottery owned by the Martin brothers came grotesque and often amusing stoneware vases that were sometimes decorated with coloured slips.

The European continent. In the 19th century Meissen and Sèvres continued to be the two principal factories and leaders of fashion, although at both places, as elsewhere, artistic standards declined considerably.

Meissen and Sèvres in the 19th century

In the first half of the 19th century Meissen adopted the revived Rococo style, and a large export trade with England was renewed. This was the period of the sentimental Dresden shepherdess, formerly much admired in England and the United States. Later productions include large and ornate candelabra, overdecorated mirror frames, clock cases, and the like, as well as vases and tureens based on the old Rococo models.

From about 1870 styles altered somewhat and are afterward referred to as those of *die Neuzeit* ("the New Period"). Some of the figures and groups illustrating contemporary subjects throw an amusing sidelight on manners and customs of the time.

At Sèvres, as a result of Napoleon's campaign in Egypt and the newly aroused interest in that country, the Empire style of the first decades of the 19th century incorporated many Egyptian motifs, which were somewhat incongruously translated into porcelain. Also, many porcelain plaques with minutely detailed overglaze painting in imitation of easel pictures were executed.

Technical improvements include the introduction, in about 1855, of *pâte-sur-pâte*, a process later popular in England, particularly at Minton's. The design was painted in white slip onto a surface of coloured, lightly fired clay. After each coat of slip dried another superimposed upon it, until the desired degree of relief had been attained. Finally, it was scraped, smoothed, and incised by metal tools and the whole object glazed and fired.

The sculptor Auguste Rodin was employed at Sèvres for a short time but does not seem to have left any enduring marks of his presence. Artistically speaking, Sèvres porcelain has not been very distinguished since the 18th century.

The Royal Porcelain works at Copenhagen has made a great deal of porcelain with simple patterns in underglaze blue derived from Chinese sources by way of Meissen. Molded fluted shapes are characteristic. Production of the well-known biscuit figures after the sculptor Bertel Thorvaldsen (1768–1844) began in 1867. The factory later introduced a slightly amber-coloured biscuit that was used for figure modelling. Painting on a grayish-toned crackled glaze led to experiments with celadons since, technically, the two have much in common (see below *China: Han Dynasty*). Other glazes inspired by early Chinese work followed. The firm of Bing and Gröndahl was established in 1853 and has done excellent and imaginative work.

Royal Copenhagen

The United States. Although Andrew Duché had succeeded in making porcelain as early as 1741, the first man to produce porcelain in any quantity was William Ellis Tucker of Philadelphia. At first he was a decorator of whiteware but started to manufacture both creamware and bone china about 1826. Judge Joseph Hemphill became a partner in 1832, and workmen were imported from Europe. Copies of Sèvres porcelain and other European wares were made about this time in a fine white porcelain body. The first factory at Bennington, Vermont, founded by Capt. John Norton in 1793, made domestic wares, including salt-glazed stoneware. The factory was removed to Bennington Village by his son, Judge Luman Norton, in 1831, and creamware and a brown-glazed ware were produced. In 1839 the factory became Norton and Fenton, and about 1845 the manufacture of Parian ware began. This unglazed near-white porcelain named after Parian marble had been made first in England by Copeland & Garrett (see above *Britain*). John Harrison of Copeland's was hired by Norton and Fenton and brought with him a number of molds. An ironstone china called graniteware or white granite was also made.

The East Liverpool, Ohio, industry was established in 1838 by James Bennett, an English potter. The first prod-

ucts made there were Rockingham and yellow-glazed ware. In the decade following the Civil War, William Bloor, Isaac W. Knowles, and others introduced the production of whiteware. By the last decade of the 19th century, production had grown until it was the largest pottery-producing area in the world.

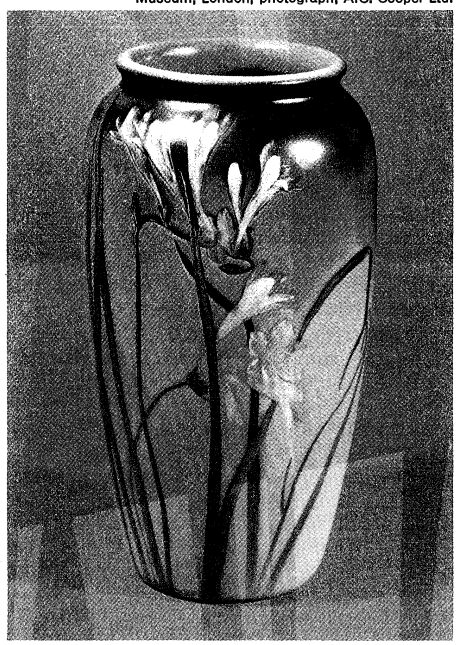
At about the same time, Zanesville, Ohio, was also developing as a pottery centre. First production was salt-glazed and slip-decorated stoneware. At a later date much artware was produced in Zanesville plants operated by Samuel Weller, J.B. Owens, George Young, and others. This artware established the basis for a sizable modern interest in collecting. Another important centre during the 19th century was at Trenton, New Jersey, where the first factory was established in 1852. Connected with it was William Bloor, who had some responsibility for putting the industry on a successful footing in East Liverpool. Trenton, like East Liverpool, produced fine, skillfully decorated whiteware.

Study of pottery manufacture at universities

A close study of the technical side of manufacture was not undertaken until Edward Orton, Jr., succeeded in getting support for the establishment of a department of ceramics at Ohio State University in Columbus in 1894. The New York State College of Ceramics at Alfred, New York, was started soon afterward, with Charles F. Binns as its director. Binns was a member of an English family connected with the manufacture of porcelain at Worcester and Derby during the 19th century and had himself held a supervisory position at Worcester. Similar departments were added to other universities soon afterward, and in 1898 Orton took the lead in forming the American Ceramic Society. In this way knowledge was put on a more scientific basis, and the trained potters who soon became available to the industry were responsible for many technical improvements. Nevertheless, the artistic direction of the factories did not reach a high standard.

Toward the end of the century it became fashionable for American women to study the art of painting on European pottery, and the Cincinnati Art Pottery Company was founded in 1879 to promote sound pottery design. As a result of its work, the Rookwood Pottery was established in 1880 by Maria Longworth Storer. Rookwood wares show a distinct Japanese influence and have excellent red and yellowish-brown glazes.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.



Vase painted in coloured slips under glaze, Rockwood pottery, Cincinnati, Ohio, *c.*1900. In the Victoria and Albert Museum, London. Height 26.4 cm.

20TH CENTURY

Pottery factories. At the beginning of the 20th century the Wedgwood factory, whose work has always remained at a high level, extended its already considerable business in the United States, and a service of nearly 1,300 pieces was supplied to the White House during the presidency of Theodore Roosevelt (1901–09). In 1940 the factory began to move to its present site at Barlaston, Staffordshire, when the historic site at Etruria, Staffordshire, was progressively abandoned; the new plant is one of the most modern in the world.

Designs of Dorothy Doughty

A new development in decorative porcelain is to be seen in the designs of Dorothy Doughty for the Worcester Royal Porcelain Company, in England, and in those of Edward Marshall Boehm, at Trenton, New Jersey. Characteristic of this kind of work are the American birds of Dorothy Doughty issued in limited editions by the Worcester Company. They are especially remarkable for technical advances in preparing the article for firing, which allow the material to be treated with much greater freedom than hitherto. Porcelain becomes very soft when it reaches the point of vitrification, but, using an elaborate series of props to support free-floating parts, the Worcester technicians succeeded in firing designs that would have been completely impossible earlier. Associated with these models are exact reproductions of natural flowers that also excell in complexity and verisimilitude anything made in the past.

In the early part of the 20th century, Bernard Moore experimented with Chinese glazes, producing some successful flambé and *sang-de-boeuf* glazes on a stoneware body at his small factory in Stoke-upon-Trent (see below *China*: *Ch'ing dynasty*). He worked in association with William Burton of Pilkington pottery in Manchester, which made experimental decorative ware of all kinds.

After World War I, figure modelling worthy of the old Meissen tradition was done by Paul Scheurich, Max Esser, Paul Börner, and others. The early red stoneware was also revived. This renaissance was halted temporarily by World War II, but production had been resumed by 1950. In the 1970s the wares exported from the German Democratic Republic into western Europe were excellent in quality.

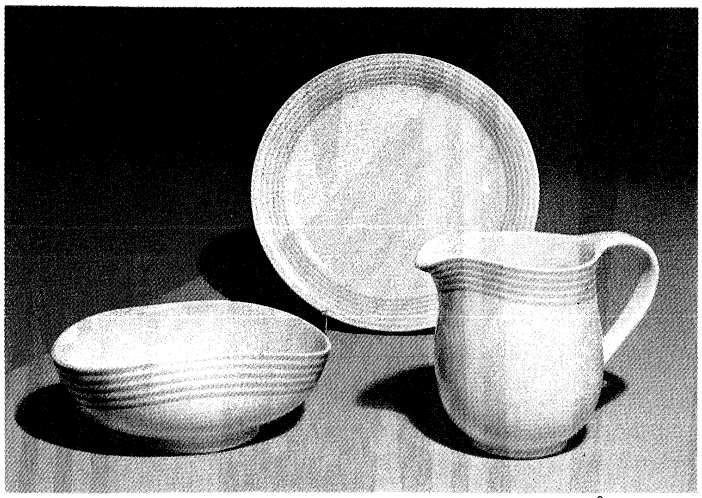
A factory that has preserved its traditional reputation for fine porcelain is Nymphenburg, at Munich, now the Staatliche Porzellan-Manufaktur Nymphenburg. At the beginning of the 20th century it began to use a wider range of underglaze colours with the aid of colour chemists from Sèvres and, about the same time, reissued some of the old figures and services of Bustelli and Auliczek (appropriately marked). Attention was soon turned to services of fine quality in the modern idiom, and excellent figures by Resl Lechner and others were produced. Lechner succeeded in adapting the 18th-century styles to 20th-century purposes in a manner that is an object lesson to those manufacturers who insist, even today, on adding the scrolls and flourishes of the Rococo.

Such factories as Rörstrand and Gustavsberg in Sweden and Arabia Oy in Finland have achieved a growing reputation for excellent design in the modern idiom. The emphasis on form in present-day pottery is to a great extent due to the import of Chinese wares of the Sung dynasty (see below) during the 1920s.

The pottery of the United States in the 1970s bears comparison with that of any other country, and standards are constantly improving. Technically, the U.S. is perhaps ahead of the rest of the world. The growing appreciation of good pottery design has led national, state, and local governments to sponsor pottery making as an art. There is also a pottery experimental station in the Tennessee Valley.

Government-sponsored pottery making

The artist-potter. The artist-potter has had an important influence on modern design from the time that Bernard Leach (1887–) established the St. Ives Pottery in 1920. Leach spent many of his early years in the Far East and learned the art of making *raku* and stoneware in Japan (see below *Japan*: *Kamakura and Muromachi periods*). He began working at a time when interest in early Chinese wares had greatly increased, and much of his work is obviously influenced by the work of Tz'u-chou (see below *China: Sung dynasty*) as well as that of Japan. It is, nevertheless, strongly individual. One of Leach's



"Grey Bands," porcelain service designed by Wilhelm Kåge for the Gustavsberg factory, Sweden, 1944. In the Victoria and Albert Museum, London. Height of jug 14 cm.

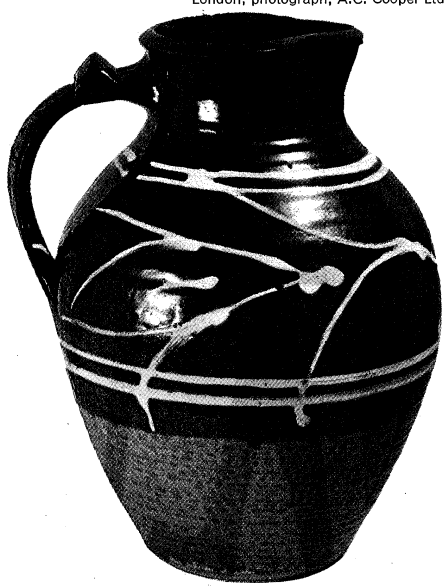
By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.

pupils, Michael Cardew, has done good work in stoneware, which he often decorates with vigorous patterns drawn with a pleasing economy of outline. William Staite Murray, at one time the head of the ceramic department of the Royal College of Art, has made some important and interesting stoneware and has influenced many younger potters. Excellent work has been done by continental potters working in England, among them Lucie Rie from Vienna and Hans Coper from Germany. Amusing figures have come from Marion Morris, who was trained in Budapest.

The artist-potters on the Continent tend to be less conservative than their English counterparts, and many new and interesting developments have occurred. The modern fashion for abstraction in art is particularly favourable to development, since the potter understood its principles long before the 20th-century painters and sculptors came to it. By the 1970s many art schools included pottery making in their curriculum, and students were rapidly increasing in numbers.

Somewhat outside the mainstream of pottery tradition, and a markedly individual production related to their work in other media, is the pottery of such well-known artists as Pierre-Auguste Renoir, Paul Gauguin, Joan Miró, Henri Matisse, Ernst Barlach, and Pablo Picasso. A good many of these wares are unique, although some have been repeated by factory production methods.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.



Slipware jug with clear, honey-coloured glaze by Michael Cardew, Winchcombe, Gloucestershire, *c.* 1938. In the Victoria and Albert Museum, London. Height 29.5 cm.

III. East Asian and Southeast Asian pottery

CHINA

Nowhere in the world has pottery assumed such importance as in China, and the influence of Chinese porcelain on later European pottery has been profound.

It is difficult to give much practical assistance on the question of Chinese marks. Most of the Chinese marks give the name of the dynasty and that of the emperor; however, many of them have been used so inconsequentially that, unless the period can also be assigned with reasonable certainty by other means, it is better to disregard them. The dating of Chinese pottery is further complicated by the fact that there were traditional and persisting types that overlapped; quite often, therefore, dynastic labels cannot be regarded as anything more than an indication of the affinities of the particular object under discussion.

Chinese symbolic decoration

Chinese decoration is usually symbolic and often exploits the double meaning of certain words; for instance, the Chinese word for bat, *fu*, also means "happiness." Five bats represent the Five Blessings—longevity, wealth, serenity, virtue, and an easy death. Longevity is symbolized by such things as the stork, the pine, and the tortoise, the *ling chih* fungus, and the bamboo, all reputed to enjoy long life. The character *shou*, which also denotes longevity, is used in a variety of ornamental forms. Together, the peach and the bat represent *fu-shou*, long life and happiness. The "Buddha's hand" citron, a fruit with fingerlike appendages, is a symbol of wealth, and each month and season is represented by a flower or plant. The *pa kua*, consisting of eight sets of three lines, broken and unbroken in different combinations, represent natural forces. They are often seen in conjunction with the Yin-Yang symbol, which represents the female–male principle, and which has been well described by the pottery scholar R.L. Hobson as resembling "two tadpoles interlocked." The dragon generally is a mild and beneficent creature. It is a symbol of the emperor, just as the phoenix-like creature (*feng-huang*) symbolizes the empress.

There are three principal religious systems in China: Confucianism, Taoism, and Buddhism. Taoist figures, in particular, appear frequently on porcelain as decoration. The most important, Lao-tzu, has a large and protuberant forehead. He is usually accompanied by the "eight immortals" (*pa hsien*), and these are sometimes modelled as sets of figures. The eight horses of the emperor Mu Wang (Chou dynasty) are also frequently represented. The Buddhist goddess Kuan-yin and the eighteen lohan, disciples of Buddha, were also modelled. The "eight Buddhist emblems" appear fairly frequently, as do the "eight precious things" and a collection of instruments and implements used in the arts and known as the "hundred antiques." The "lions of Buddha" (often miscalled dogs) are frequently represented, as is the kylin (properly the *ch'i-lin*), which is a composite animal, not unlike a unicorn, that has a fierce appearance but gentle disposition.

Most of these symbols were not used in pottery decoration before the Ming dynasty, although both the dragon and a phoenix-like creature (probably the Chinese pheasant), as well as some floral motifs, are earlier. The *lei-wen*, however, which resembles the Greek key fret (an ornament consisting of small, straight bars intersecting one another in right angles) and is sometimes used on the later ceramic wares, appears on bronzes as early as the Shang and Chou dynasties, where it is called the cloud-and-thunder fret. The *t'ao-t'ieh*, a grotesque mask of uncertain origin, also appears on early bronzes and on later pottery and porcelain. Decorations based on Chinese literary sources are usually extremely difficult to trace to their origin.

The earliest Chinese pottery is of the Neolithic period and has been discovered in the provinces of Honan and

Earliest Chinese pottery

Kansu. Perhaps the best known of these wares is a series of large urns of red polished pottery with geometric decoration found in the Pan-shan cemetery and at Mach'ang, both in Kansu Province. These were made by hand, the latest specimens with perhaps some assistance from a slow wheel, and are at least as early as 2000 BC.

Shang (or Yin) and Chou dynasties (c. 1766–221 BC). The only known complete specimen of a fine white stoneware dating from *c.* 1400 BC (Freer Gallery of Art, Washington, D.C.) is decorated with chevrons (linked V-shapes) and a key-fret pattern, the shoulder motifs being reminiscent of those seen on contemporary bronze vessels. This ware is much better in quality than most other surviving pottery of the Shang period (*c.* 1766–1122 BC) or of the following Chou dynasty (*c.* 1122–221 BC). Much Chou pottery is decorated with rudimentary incised ornament, some of which resembles the impress of coarse textiles referred to as mat markings. The shapes used for these pieces were often inspired by bronze vessels.

The development of glazing in China may have started with the application of glass paste to some of the later Chou wares. Stoneware vessels of about the 3rd century BC have a glaze that is little more than a smear but one that has obviously been deliberately applied. This type persisted for several centuries.

Han dynasty (206 BC–AD 220) and the period of the Six Dynasties (AD 221–589). The first pottery to survive in appreciable quantities belongs to the Han dynasty; most of it has been excavated from graves. Perhaps the commonest form is the *hu*, a baluster-shaped vase copied from bronze vessels of the same name and sometimes decorated with relief ornament in friezes taken directly from a bronze original. The hill jar, which has a cover molded to represent the Taoist "Isles of the Blest," is another fairly frequent form, and many models of servants, domestic animals, buildings, wellheads, dovecots, and the like also have been discovered in graves. Some of this pottery is unglazed or decorated with cold (*i.e.*, unfired) pigments, but much of it is covered with a glaze that varies from copper green to yellowish brown; often the colours have become iridescent from long burial. The body is usually a dark red and approaches stoneware in hardness.

Han glaze is more glasslike than that of the Chou period and is of an excellent quality. It contains lead and was frequently coloured green with copper oxide.

Celadon glazes

Yüeh *yao* ("Yüeh ware") was first made at Yüeh-chou, Chekiang Province, during the Han dynasty, although all surviving specimens are later, most belonging to the Six Dynasties. They have a stoneware body and an olive or brownish-green glaze and belong to the family of celadons, a term that looms large in any discussion of early Chinese wares. It is applied to glazes ranging from the olive of Yüeh to the deep green of later varieties. These colours were the result of a wash of slip containing a high proportion of iron that was put over the body before glazing. The iron interacted with the glaze during firing and coloured it.

T'ang dynasty (AD 618–907). Chinese pottery reaches an important stage in its development during the T'ang dynasty.

Nearly everything that has survived has been excavated from tombs, many of them found accidentally by railway engineers and latterly by more systematic excavations. Excavations at Sāmarrā' on the Tigris, a luxurious residence built by the caliph al-Mu'taṣim (son of Hārūn ar-Rashīd) in AD 836 and abandoned in 873, have uncovered many fragments of T'ang wares of all kinds. Perhaps the most important finds from a historical viewpoint are the fragments of what is undoubtedly porcelain. An Islāmic record of travels in the Far East, written in 851, records "vessels of clay as transparent as glass." There can be little doubt, therefore, that translucent porcelain was made in the T'ang period, although it was not until the Yüan dynasty (1279–1368) that it began to resemble the type with which the West is most familiar.

Early porcelain

Perhaps the most important single development was the use of coloured glazes—as monochromes or splashed and dappled. The T'ang wares commonest in Western collections are those with either monochrome or dappled glazes covering a highly absorbent, buff, earthenware body. The dappled glazes were usually applied with a sponge, and they include blue, dark blue, green, yellow, orange, straw, and brown colours. These glazes normally exhibit a fine crackle and often fall short of the base in an uneven wavy line, the unglazed surface area varying from about one-third to two-thirds of the vessel.

Dappled glazes are also found on the magnificent series of tomb figures with which this period is particularly associated. Similar figures were made in unglazed earthenware and were sometimes decorated with cold pigment. Although the unglazed specimen or those covered only with the straw-coloured glaze are occasionally modelled superbly, many are crude and apparently made for the tombs of the less affluent and influential. Most of the glazed figures are much better in quality and occasionally reach a large size; figures of the Bactrian camel, for instance, are particularly impressive, some being nearly three feet high. The Bactrian pony, introduced into China about 138 BC, is to be found in many spirited poses. This

By courtesy of the Victorian and Albert Museum, London



Ceramic tomb figure decorated in characteristic coloured glazes, T'ang dynasty (AD 618–907). In the Victoria and Albert Museum, London. Height 71 cm.

fashion for tomb figures fell into disuse at the beginning of the Sung dynasty (AD 960–1279) but was revived for a short while during the Ming period (1368–1644), when T'ang influence is noticeable.

Marbled wares are seen occasionally. The effect was achieved either by combing slips of contrasting colours (*i.e.*, mingling the slips after they had been put on the pot, by means of a comb) or by mingling differently coloured clays. Another type of T'ang ware (probably from Honan) had a stoneware body with a dark-brown glaze streaked by pale blue. Most vessels stand on a flat base; although later T'ang wares sometimes were given a foot ring, for the most part this can be regarded as evidence in favour of a Sung dating.

Sung dynasty (AD 960–1279). The wares of the Sung dynasty are particularly noted for brilliant feldspathic glazes over a stoneware body and their emphasis on simplicity of form. Decoration is infrequent but may be incised, molded, impressed, or carved; a certain amount of painted decoration was done at Tz'u-chou in Chihli (now called Hopeh) Province (*see* below). The esteem accorded to the Sung wares accounts for the relatively large number that have survived. The principal varieties are Ju, *kuan*, Ko, Ting, Lung-ch'üan, Chün, Chien, Tz'u-chou, and *ying-ch'ing*.

Ju ware has a buff stoneware body and is covered with a

dense greenish-blue glaze that sometimes has a fine crackle. It was made in Honan at an Imperial factory that apparently had a life of about 20 years, starting in 1107.

Kuan ("official") is another Imperial ware that is also exceedingly scarce. It was probably first made in the north, the kilns being re-established at Hangchow in Chekiang Province about 1127, when the court fled southward to escape the Chin Tatar invaders. The body is of stoneware washed with brown slip. The glaze varies from pale green to a lavender blue, with a wide-meshed crackle emphasized by the application of brown pigment. Chinese references to "a brown mouth and an iron foot" can be identified with the colour of the rim and the foot ring.

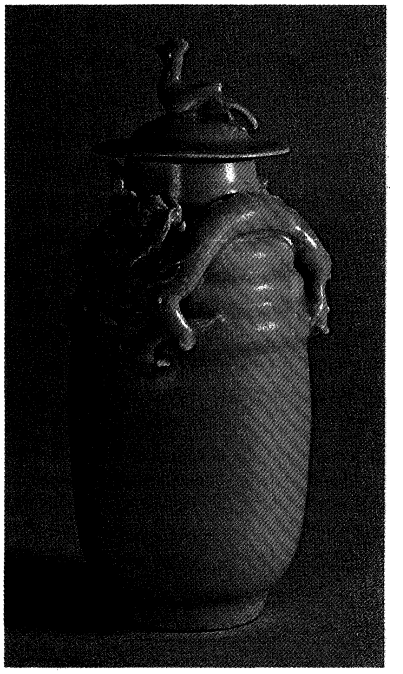
Ko ware is closely related to *kuan* ware. It has a dark stoneware body and a grayish-white glaze with a well-marked crackle, which was induced deliberately for its decorative effect.

Ting wares are white. Some exhibit an orange translucency, while the coarser varieties are opaque. The finest examples are called "white" (*pai*) Ting. On the exterior of bowls and similar vessels the glaze of white Ting is apt to collect in drops, called teardrops. Many articles, particularly bowls, were fired mouth downward, leaving an unglazed rim that was afterward bound with a band of copper or silver. (Bands appear occasionally on other Sung wares, notably *ying-ch'ing*, and were sometimes used to conceal damage rather than an unglazed rim.) Coarser varieties are known as "flour" (*fen*) Ting and "earthen" (*t'u*) Ting, and there are also a few examples of black Ting. As in the case of *kuan* ware, the kilns are said to have been removed southward in 1127, but it has so far proved impossible to differentiate between the northern and southern varieties. Other white wares made elsewhere during the period include those of Tz'u-chou and a variety covered with a white slip over a grayish body from Chü-lu Hsien (both in present Hopeh Province).

Lung-ch'üan

The celadons of Lung-ch'üan are, perhaps, the most common of the classic Sung wares. The town is in the province of Chekiang, near the capital of the southern Sung emperors at Hangchow. The kilns probably date back to the 10th century. The glaze, which is superb in quality, is a transparent green in colour. It is thick and viscous, usually with a well-marked crackle. (The glaze on early specimens is less transparent and more dense.)

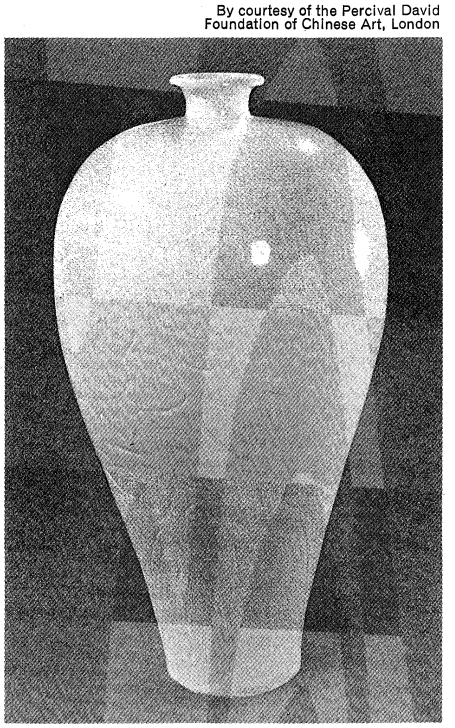
By courtesy of the Victoria and Albert Museum, London



Lung-ch'üan celadon ware wine jar and cover, with light bluish-green glaze, Sung dynasty, 12th century. In the Victoria and Albert Museum, London. Height 25.4 cm.

The body is gray to grayish white, best seen at the rim, where the glaze tends to be thin. By far the most frequent surviving examples of Lung-ch'üan celadon are large dishes, for which there was a thriving export trade, due in part to the superstition that a celadon dish would break or change colour if poisoned food were put into it. Bowls and large vases, both of which are scarce, were also made with this glaze. Decoration is usually incised, but molded decoration is also found. On some pots the molding was left unglazed, so that it burned to a dark reddish brown —an effective contrast to the colour of the glaze. The more finely potted wares are the scarcest and often the oldest. The heavier varieties were intended to withstand the rigours of transport to overseas markets, and probably most of them belong to the Yüan dynasty (1279–1368), when the export trade was considerably extended.

By courtesy of the Percival David Foundation of Chinese Art, London



Ting ware vase, white porcelain with an ivory-white glaze covering carved design of peony flowers and foliage, Sung dynasty (AD 960–1279). In the Percival David Foundation of Chinese Art, London. Height 36.8 cm.

Chün ware comes from the K'ai-feng district of Honan Province. The body is a grayish-white, hard-fired stoneware covered with a thick, dense, lavender-blue glaze often suffused with crimson purple. This is the first example of a reduced copper, or flambé, glaze. Conical bowls are especially numerous, and dishes are not unusual, but the finer specimens are usually flowerpots, sometimes said to have been made for Imperial use. Characteristic are barely perceptible channels or tracks caused by the parting of the viscous glaze; these are called earthworm tracks by the Chinese. The kilns probably continued to produce this ware until the 16th century, and it is difficult to separate some of the later productions from the earlier.

Chien ware is called after the original place of manufacture, Chien-an, in Fukien Province. Manufacture was later moved to nearby Chien-yang, probably during the Yüan period. The glaze is very dark brown, approaching black, over a dark stoneware body, and it usually stops short of the base in a thick treacly roll.

There are many variations in the colour of the glaze. Streaks in lighter brown are referred to by the Chinese as hare's fur. Silvery spots on the glaze are called oil spots. The most usual surviving form is the teabowl; these were much esteemed by the Japanese under the name of *temmoku* and were used in the tea ceremony (see below *Japan: Kamakura and Muromachi periods*).

The Tz'u-chou ware

The kilns of Tz'u-chou, formerly in Honan, are now in Hopeh Province. The earliest surviving examples are referable to the T'ang dynasty. In the Sung period, vases, wine jars, and pillows (which are more comfortable than they appear) were the most usual products. The body is usually a hard-fired, grayish-white stoneware that was first covered with a wash of white slip and then with a transparent glaze. For the first time painted decoration appears under the glaze, perhaps as a result of influence from the Middle East (see above *Islāmic pottery: 11th to 15th century in Mesopotamia and Persia*). Decoration is nearly always in brown or black; the motifs are usually floral and display a singular freedom of line that is very attractive. (The inclusion of human and animal figures suggests a Yüan or a Ming dating, at the least.) The slip covering was sometimes carved away, leaving a pattern in contrasting colour, a technique also used in conjunction with a dark brown glaze. A hare's-fur glaze, similar to that of Chien wares, was also employed. A blue glaze with painted decoration in black beneath it was obviously inspired by contemporary Persian pottery decorated in the same way. Another innovation, perhaps derived from the same source, is the use of colours applied over the glaze. These are limited to primitive reds and greens and yellows.

Ying-ch'ing ware

An important and not uncommon ware is *ying-ch'ing* ("shadowy blue"). It was manufactured both in the south (Kiangsi) and the north (Hopeh). Moreover, it was extensively exported and has been found as far west as the ruins of al Fusṭāṭ (old Cairo). The body is pale buff in colour, usually translucent, and thinly potted, breaking with a sugary fracture. Most genuine examples seem to belong to the Sung and Yüan periods, but it is probable that, in the north at any rate, manufacture started late in the T'ang dynasty and lasted well into the Ming period. Bowls of conical form are the commonest survival, and many are decorated with incised floral and foliate motifs. Lightly molded decoration occurs, as does combing of the clay. The *mei p'ing* vase is found with this glaze; it has a tall body with straight sides, high, rounded shoulders, and a short narrow neck and was intended to hold a single spray of prunus blossom. Stem cups, deep bowls, and ewers were also produced. Bowls sometimes have the rim bound with copper.

Yüan dynasty (1279–1368). The Yüan, or Mongol, dynasty is often regarded as being no more than transitional between Sung and Ming types. This is not entirely true. Undoubtedly, many Sung types were continued, just as the T'ang types were continued at the beginning of the Sung dynasty, but there are other wares that represent a new departure. The manufacturing centre of Ching-te-chen increased in importance and first manufactured the white translucent porcelain that was to have a revolutionary effect on Chinese wares. The use of painted decoration, begun during the Sung period at Tz'u-chou, also became much more widespread, and the two techniques were combined in a manner that later affected the course of porcelain manufacture throughout the world.

The *Ko ku yao lun* of 1387 refers to *shu fu* ware, a type of white porcelain. The base is unglazed. Decoration in relief, painted in slip or engraved, is to be seen on some surviving examples of porcelain. It appears to be a development of the earlier *Ting ch'ing*. Much more unusual is the appearance of a few specimens of Yüan date that are painted with reduced copper red under the glaze. As mentioned above, the potters of Chün-chou had achieved this colour, but only in the glaze.

The use of underglaze blue was introduced from the Middle East, where it had been employed at least as early as the 9th century, specimens thus decorated having been recovered at Sāmarrā' (see below *Islāmic pottery: 'Abbāsid*). The best known example of Yüan porcelain decorated in this manner, which is usually referred to as blue-and-white, is a pair of vases in the Percival David Foundation of Chinese Art in London. They bear a date equivalent to 1351. The peony scroll, carved or in applied relief, appears on some of these blue-and-white wares.

Blue-and-white ware

Ming dynasty (1368–1644). The Mongol emperor Shun Ti was defeated in a popular uprising, and the Hung-wu emperor, founder of the Ming dynasty, succeeded him in 1368. When the country had recovered from these internecine struggles, pottery art took a new lease of life, though under somewhat changed conditions. The Sung wares went out of favour, and the old factories sank into obscurity, while the fame and importance of the great porcelain town of Ching-te-chen, near the P'o-yang Hu in Kiangsi Province, overshadowed all the rest. The Imperial factory there was rebuilt and reorganized to keep the court supplied with the new porcelain. The staple product of Ching-te-chen was the fine white porcelain that made "china" a household word throughout the world; and as this ware lent itself peculiarly well to painted decoration, the vogue for painted porcelain rapidly replaced the old Sung taste for monochromes.

"China" ware

The reign of the Yung-lo emperor (1402–24) is remarkable for some extremely thin-walled pieces, referred to as "bodiless" (*t'o t'ai*) ware. Engraved examples are known, and Chinese commentaries refer to specimens decorated in red.

After this early period, Ming wares generally are fairly easily recognizable. Porcelain replaced stoneware as the usual medium, and polychrome decoration became widely employed. The largest single group of Ming porcelain is that painted in blue underglaze. Much of the pigment used was imported from Middle Eastern sources. Supplies of this so-called Mohammedan blue (*hui-hui ch'ing*), which came from the Kashān district of Persia, were not always obtainable and were interrupted on more than one occasion. The quality of the blue-painted wares, however, remained to a great extent dependent on its use until the end of the 16th century, when methods of refining native cobalt were devised.



Flask decorated with a dragon and wave scrolls in underglaze blue, Ming dynasty, 14th century. In the Victoria and Albert Museum, London. Height 36.8 cm.
By courtesy of the Victoria and Albert Museum, London

The wares lack much of the precision of the porcelain made during the following Ch'ing period (1644–1911), when a kind of factory system grew up that divided the work into a large number of repetitive operations. Little trouble was taken to smooth over imperfections of manufacture, and foot rings are often finished summarily. The glaze, too, frequently has minor defects, and articles, such as vases, are sometimes slightly distorted and carelessly finished. The shape of many examples can fairly be described as massive, in spite of the fact that most of them were made for export, and the difficulties of transporting them must have been considerable. None of these factors evinces a lack of skill, especially as the potters were quite capable of technical virtuosity when they wished to dis-

play it—some of the most thinly potted of all Chinese porcelain belongs to this period. It seems that the Ming potters disdained the attitude of mind that treated blemishes as important; occasional distortions, in fact, were regarded as lending interest to an object. The Chinese did not carry this aesthetic creed to the same lengths as the Japanese (see below *Japan*), but the difference seems to be largely one of degree. Ming wares can fairly be described as masculine, in contrast to the more feminine, more precisely finished wares of the later Ch'ing period.

Reign of the Hsüan-te emperor (1425–35). In this period the arts were particularly fostered and a high level of achievement attained. The blue painting is blackish in colour, with dark spots at intervals where thick blobs of pigment were deposited by the brush—the so-called heaped and piled effect. The motifs are floral and foliate, with the occasional use of fish and waterfowl. Sometimes vessels are bordered by a pattern of conventional rock amid waves—the Isles of Immortality—often referred to as the Rock of Ages pattern. The pattern appears frequently throughout the Ming period and later.

Heaped and piled effect

Contemporary Chinese commentaries refer to the use of underglaze copper red—often called sacrificial red for uncertain reasons. To a great extent sacrificial red was abandoned later in the dynasty in favour of overglaze iron red, although it was used again during the reign of the Ch'ing dynasty K'ang-hsi (1661–1722) and Yung-cheng (1723–35) emperors and appears in a rather primitive form from some provincial kilns. Both copper red and blue were used as monochromes and, occasionally, together; but since these pigments required a slightly different firing temperature, one or the other is usually deficient in quality.

The use of overglaze colours was rare, and the technique had by no means been fully mastered.

Reign of the Ch'eng-hua emperor (1464–87). Much overglaze decoration can be attributed with a reasonable measure of certainty to the reign of Ch'eng-hua, the finest examples being, perhaps, the chicken cups, so-called because they are decorated with chickens. Their decoration is outlined in underglaze blue and filled in with soft overglaze colours called "contending colours" (*tou ts'ai*). Ch'eng-hua overglaze colours were thin, subdued in colour, and pictorial in effect.

The practice of enamelling directly onto unglazed, or biscuit, porcelain instead of onto a glazed and fired body is sometimes thought to have begun in this reign, though that of the Chia-ching emperor (1521–66) is the more likely. Ming specimens are, in any case, extremely rare; most belong to the reign of the K'ang-hsi emperor (1661–1722) in the Ch'ing dynasty.

Reigns of the Hung-chih and Cheng-te emperors (1487–1521). The first use of a coloured overglaze ground can be attributed to the reign of the Hung-chih emperor (1487–1505), when a yellow of variable shade first appears. In the reign of the Cheng-te emperor (1505–21) the influence of the Muslim palace eunuchs who supervised the Imperial kilns is seen in such blue-and-white motifs as the Mohammedan scroll, which is composed of somewhat formal flowers joined by S-shaped stems, with scroll-like leaves at intervals along them. Mohammedan blue was again available. The earliest versions of this theme, which seems orginally to have come from a textile pattern, are the least stiffly drawn. The linear style of painting characteristic of earlier porcelain altered to one in which outlines were filled in with flat ungraduated washes.

Muslim influence

Reign of the Chia-ching emperor (1521–66). This reign is notable for a deterioration in the quality of the porcelain body, offset by the use of rich dark blue. Wares painted overglaze, too, were executed in good colours, with well-marked outlines. A characteristic colour, the opaque iron red (*fan hung*), sometimes called tomato red, was used as a monochrome with gilt traceries over it on bowls that sometimes had interior decoration in underglaze blue. Various wares have decoration in red and green, a palette that became more familiar later. A yellow glaze is found in conjunction with incised decoration (usually a dragon) in green. Very rarely was a green or blue monochrome used.

Reigns of the Lung-ch'ing and Wan-li emperors (1566–1620). The styles of Chia-ching were, to some extent, continued in the following reigns of the Lung-ch'ing (1566–72) and Wan li (1573–1620) emperors. A palette containing underglaze blue in conjunction with green, yellow, aubergine purple, and iron red (the precursor of the later Ch'ing *famille verte* palette) was known as "Wan-li five-colour" ware (Wan-li *wu ts'ai*). The red and green Chia-ching decoration was also used, and vast quantities of blue-and-white porcelain were produced for export. The body is quite unlike that used earlier in the dynasty, being thin, hard, crisp, and resonant. It is the commonest of all Ming wares in the West. During the reign of the Wan-li emperor, much pierced work (*ling lung*) was done. Pierced objects range from small brush pots to vases with coloured glazes sometimes termed *fa hua*.

The Ming dynasty ended in 1644. The wares of the last three emperors, for the most part, followed styles already established; perhaps an exception can be made for blue-and-white, which shows a number of new departures in both form and decoration. Many of the vases are without a foot-ring and stand on a flat, unglazed base. Forms based on European wares were obviously made for export.

Provincial and export wares. Most of the wares hitherto discussed were made in the Ching-te-chen area; it remains to consider the other wares of the period. The export of celadons went on, not only to the countries west of China but also to Japan, where they were much esteemed. Most celadons attributable to the Ming period have incised under the glaze floral and foliate decoration of a kind that also appears on blue painted wares.

The fine porcelain of Te-hua in Fukien Province was first made, perhaps, in the early part of the dynasty. Most of this porcelain was left undecorated, and received the name in Europe of blanc de chine. The glaze is exceptionally thick and lustrous, and early examples are often slightly ivory in tone. Overglaze painting is infrequent; virtually all early coloured specimens, figures or vessels, have been decorated in Europe, usually in the Netherlands. Figures especially were produced at the Te-hua kilns, with the Buddhist goddess Kuan-yin being a favourite subject.

Blanc de chine porcelain

The stoneware of I-hsing in Kiangsu Province was known in the West as boccaro ware and was copied and imitated at Meissen, at Staffordshire, and in the Netherlands by Ary de Milde and others. Its teapots were much valued in 17th century Europe, where tea was newly introduced. The wares of I-hsing are unglazed, the body varying from red to dark brown. The molding is extremely precise and was often sharpened by grinding on a lapidary's wheel. The body was sometimes polished in the same way.

Most of the Ming stoneware ridge tiles and roof finials were made at kilns near Peking. Many of them are decorated in green, yellow, turquoise, and aubergine-purple glazes, recalling the wares of the T'ang dynasty. A Ming date is exceedingly optimistic for most of them. To this group belong, it is thought, a few large figures that have sometimes been somewhat doubtfully awarded a T'ang date.

The provincial tile kilns also manufactured "three-coloured" (*san ts'ai*) wares, perhaps originally a product of the Tz'u-chou kilns. These were decorated with coloured glazes that were often kept from intermingling by threads of clay (cloisonné technique) or were used in conjunction with the pierced technique (*fa hua*). Others have engraved designs under the glazes. Most existing specimens are large vases, barrel-shaped garden seats, and the like. The best are extremely handsome and imposing, turquoise and dark-blue glazes being particularly effective.

Ch'ing dynasty (1644–1911). With the Ch'ing dynasty came the beginning of the immense vogue for porcelain in Europe that was to reach its height during the first half of the 18th century. Many varieties of Ch'ing ware are common in the West. Its wares differ, for the most part, from those of the Ming period in a fairly distinctive manner. Potters had their medium under almost complete control, and their products are much more precisely finished. Their finesse contrasts sharply with the struggles

Difference between Chinese and European pottery making

of potters in Europe, where porcelain manufacture did not emerge from the purely empirical stage until the 19th century. Letters written in 1712 and 1722 by a Jesuit missionary who spent some years at Ching-te-chen record that some Ch'ing pieces were handled by as many as 70 men, each contributing a small part to the total effect, and this is one of the reasons why many Ch'ing wares are found to lack the freshness and the spontaneity of Ming decoration.

The Imperial kilns of Ching-te-chen were fortunate in the support they received from the palace during the reigns of the K'ang-hsi (1661–1722), Yung-chen (1723–35), and Ch'ien-lung (1736–96) emperors. The K'ang-hsi emperor, in particular, was a patron of the arts on a considerable scale.

Underglaze blue and red. The blue-painted porcelain of the Ch'ing dynasty has been somewhat neglected in the 20th century. This is probably due to the ridiculously high value placed on it during the latter years of the 19th century, when it was called Nanking ware. Even the best, which belongs to the reign of the K'ang-hsi emperor, hardly bears comparison with the finer Ming wares, though its influence on European porcelain was far-reaching. Blue-and-white porcelain was exported to Europe in vast quantities, and many of the forms were especially made for export; the condiment ledge on plates and dishes, for instance, which first appeared in the reign of the Wan-li emperor (Ming dynasty), had been added for Western customers (the Chinese used the saucer dish). The blue-and-white of the K'ang-hsi period has an extremely white body, and the blue is exceptionally clear and pure. It is variable in shade, and the design is executed in graduated washes within lightly drawn outlines, a point of difference from Ming wares. Many of the designs of the Ming period were in use, and, of the later patterns, those illustrating literary and historical themes are probably of the highest quality.

Ginger jars decorated with prunus blossom reserved in white on an irregular blue ground, intended to represent the cracked ice of spring and sometimes described as pulsating, were once valued highly; in the mid-20th century a more realistic attitude was taken toward them.

Underglaze copper red was also used during the 18th century. The stem cups of the Yung-cheng period with three fruit or three fish in silhouette, which imitate those of Hsüan-te, are much better known than the wares they copied. Copper red also appears in conjunction with underglaze blue, and a greenish-toned glaze is common with pieces thus decorated.

Underglaze blue was sometimes used as a monochrome ground colour. It was blown on the surface in powder form before glazing; a bamboo tube, closed with gauze at one end, was employed for the purpose. It is thus called powder blue, or, in Chinese, *ch'ui ch'ing* ("blown blue"), and is distinct from the sponged blue grounds of the Ming dynasty. It was subsequently used at several of the porcelain factories in Europe. Clair du lune (*yüeh pai*, "moon white"), a cobalt glaze of the palest blue shade, was also used.

Use of powder blue

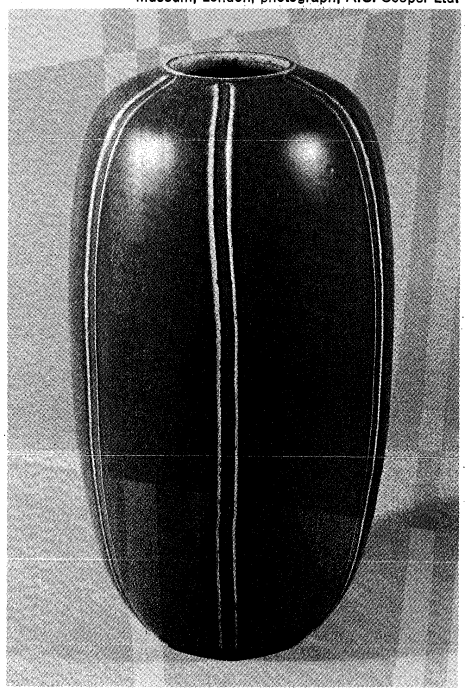
Coloured glazes. Copper red, called oxblood (*sang-de-boeuf*) by the French, appears in monochrome form as Lang *yao*. This glaze was also known to the Chinese as "blown red" (*ch'ui hung*). It was certainly used as a monochrome in early Ming times and possibly even earlier, and is the direct ancestor of the showy flambé glazes (*yao pien*) of the Ch'ien-lung period that are often vividly streaked with unreduced copper blue.

Another variation, no doubt at first accidental, is the glaze known in the West as "peach bloom," a pinkish red mottled with russet spots and tinged with green. The Chinese have various names for it, but perhaps the commonest is "bean red" (chiang-tou *hung*). It is used on a white body. Most objects glazed in this way are small items for the writer's table.

Monochromes of all kinds are a distinct and important section of Ch'ing wares, and many reproductions of Sung monochromes were made. The use of iron as a pigment can be seen in a revival of the celadon glaze. The Ching-te-chen celadons have, generally, a pale-green glaze over white porcelain, the foot-ring being given a wash of brown to simulate the old ware. Meanwhile, celadons of the Lung-ch'üan type were still being made. In addition to the celadon glaze, iron was used to produce colours varying between café au lait and pale yellow and also "dead-leaf" brown.

Sometimes panels were reserved in white and painted in overglaze colours. Specimens thus glazed appear in the old Dutch catalogues as Batavian ware, because the wares were imported via the Dutch centre of trade and trans-shipment at Batavia (modern Djakarta), in Java. They are also related to "mirror black" (*wu chin*), a lustrous colour obtained by the addition of manganese, and sometimes decorated with gilding or even, as in at least one extant specimen, with both gilding and silvering. Imperial yellow, a lead glaze often used over engraved dragons and similar designs, was again employed during the 19th century.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper Ltd.



Vase with *flambé* glaze *(yao pien)* of reduced copper, Ch'ing dynasty, reign of the Ch'ien-lung emperor, 1736–96. In the Victoria and Albert Museum, London. Height 32.4 cm.

Brilliant turquoise glazes derived from copper have been produced up to the mid-20th century, although later examples seldom have the quality of the earlier ones. The glaze is usually covered with a network of fine crackle, and in some examples there is engraved decoration under the glaze. Related glazes are the copper greens—for example, leaf green and cucumber green, the latter being speckled with a darker colour. Apple green is an overglaze colour used as a ground and applied over a crackled gray glaze. Most greens are relatively late.

Purple, or aubergine, glazes derived from manganese are seen occasionally. Brinjal bowls, decorated with engraved flowers, have an aubergine ground in conjunction with dappled green and yellow glazes. (Brinjal, in fact, means aubergine, or eggplant, which is a favourite food in parts of the East.) Bowls with engraved dragons and a combination of only two of these colours are somewhat better in quality.

Overglaze colours. Overglaze colours were sometimes used as monochromes; for example, iron red or, as it is sometimes called because it varies a little in shade, coral red. The surface is usually glossy but occasionally mat. The rose colour, discussed below, was used both as a monochrome and as a ground colour.

The wares enamelled on biscuit are a much sought after group. They are a development of the Ming *san ts'ai*

wares, which were still being made during the Ch'ing period. The effect of painting directly on biscuit was to produce a soft and distinctive colouring that is extremely attractive. The outlines were first painted directly on the unglazed surface in brownish black; some of the colours were then painted within these outlines and others were washed over them; however, red or blue overglaze colours, when they appear, are usually provided with a patch of glaze underneath them. The practice seems hardly to have survived the K'ang-hsi period, except for deliberately made later copies.

Famille verte overglaze

During the reign of the K'ang-hsi emperor the wares decorated overglaze were painted in the *famille verte* palette, usually over a white glaze. The name *famille verte* ("green family") is derived from the dominant distinctive green employed, but the wares are in fact a development of the Wan-li five-colour ware, the most important difference being the replacement of the earlier underglaze blue by an overglaze blue. On most genuine examples it is possible to see a distinct halo around the overglaze blue, but its absence does not necessarily condemn the piece as not genuine. The *famille noire* has the *verte* palette in conjunction with a black ground; the *famille jaune* uses the same colours, but is used in conjunction with a yellow ground. In each case the white porcelain disappears under the colours.



Trumpet-shaped vase with floral decoration on background of green enamel, *famille verte*, Ch'ing dynasty, reign of the K'ang-hsi emperor, 1661–1722. In the Victoria and Albert Museum, London. Height 61 cm.

By courtesy of the Victoria and Albert Museum, London

Toward the close of the reign of the K'ang-hsi emperor (*c.* 1685) an opaque overglaze rose colour appears. This and its related colours were called "foreign colours" (*yang ts'ai*). It soon formed the characteristic colour of a group of wares, referred to as the *famille rose*, which was particularly developed during the reign of the Yung-cheng emperor. It more or less replaced the *verte* palette. The translucent overglaze colours of the earlier period tended to become opaque, and painting has a more feminine quality to its style.

During the 18th century the white wares of Ching-te-chen were made mostly for the home market, though a few were exported. They included examples of the bodiless ware and the *an hua* (literally "secret language"). The latter, copied from a traditional Yung-lo (1402–24) type, has designs lightly incised or painted with white slip. The body is white, and the whole is covered with clear glaze. The decoration can only be seen plainly if light is allowed to shine through it. Pierced work was revived in certain rare pieces inspired by jade; the use of piercing that was filled with glaze was derived from Persian Gombroon ware.

European influence and the export trade. Before mid-18th century, some European wares had found their way to China, as witness certain copies of early Meissen porcelain. The taste of the European trader, though hardly representative of the more cultured section of Western civilization, also began to have considerable influence at this time.

Much decoration was done in studios in and around the port of Canton, white porcelain being sent from Ching-te-chen for the purpose. Enormous quantities of *famille rose* porcelain were painted there, including most of the "ruby-backed" dishes, which are completely covered on the reverse, except for the interior of the foot ring, with a ground of overglaze opaque rose. They often have an elaborate arrangement of minutely delineated border patterns around the central subject (usually pretty women), demonstrating the new, and later widespread, idea that the beauty of an object is directly proportional to the amount of decoration on it. This theory was to be one of the causes of the degeneration of later Chinese and Japanese wares; it was, however, by no means confined to the Orient and can be seen in most 19th-century European porcelain.

The Yung-cheng painters were the first to carry foliate decoration over on to the back of the dish, usually as a prolongation of the stem. This was repeated later during the reign of the Tao-kuang emperor (1821–50). The European tendency to draw flowers in a naturalistic manner also appears in China from the Yung-chen period onward, although the practice was not carried to the same lengths.

An attempt to imitate the European method of overglaze painting, in which colours were applied in flat washes that partly sank into soft porcelain glazes, can be seen in the "ancient moon pavilion" (*ku yüeh hsüan*) wares. These will sometimes have a European subject, for example, a Watteau shepherdess, but Chinese subjects were also used.

Oriental Lowestoft

Of the wares more directly due to European intervention perhaps the best known is Chinese export porcelain, still sometimes known as Oriental Lowestoft. The name is due to an error on the part of William Chaffers (the author of a book on pottery marks), who persisted in attributing these wares to the small English factory at Lowestoft. If this porcelain is important at all, it is as a curiosity; the artistic value is nearly always negligible. The styles are usually based on those of European pottery or metalwork or on a combination of Western and Oriental motifs in an unpleasing jumble. The designs were provided by Western traders, and coats of arms are comparatively common.

Other wares connected with the export trade are those decorated with the Mandarin patterns; these came from Cantonese studios and were introduced toward the end of the 18th century. They have figure subjects in panels that are surrounded with coloured grounds and an excess of floral and other ornament in unprepossessing combinations of colours.

Much white porcelain was sent to Canton to be decorated, but much, too, was shipped to Europe for the same purpose. Many examples were painted by German studio

painters, by Dutch enamellers, and by English "outside decorators."

Europe, of course, was not the only export market open to the Chinese. Much blue-and-white was exported to the traditional markets in Persia and the Near East (Arabic inscriptions can be seen on some specimens) and elsewhere—India, Thailand, Burma, and Tibet.

Other Chinese kilns

The wares discussed so far have been principally those of Ching-te-chen. Those of Te-hua in Fukien Province, however, are also important. Figures of the Buddhist goddess Kuan-yin in particular were exported in enormous quantities, and the *an hua* and pierced decorations often came from Te-hua. Vessels, such as libation cups with applied prunus sprigs, were copied by European factories in the 18th century, notably by those at Meissen, Chelsea and Bow, and Saint-Cloud. The body is usually white, sometimes with an ivory tone, and the glaze is thick, rich, and lustrous. European forms are to be seen occasionally, and most coloured examples have been decorated in the West. The kilns of I-hsing also continued making the traditional wares.

19th and 20th centuries. The 19th century has little to offer that is new or of good quality. Snuff bottles painted with miniature designs were first made toward the end of the 18th century, but most belong to the reign of the Chia-ch'ing (1796–1820) and Tao-kuang (1821–51) emperors. Bowls with circular medallions painted in overglaze colours with yellow or rose grounds are, perhaps, among the finer wares. Also of good quality are bowls covered with an opaque ground, rose or yellow, with designs engraved into it. These were first made in the 18th century and extend to the reign of the Tao-kuang emperor.

Most of the wares of Tao-kuang are poor in quality, although some examples in the style of Yung-cheng are better. The glaze has a musliny texture similar to that seen on some early Ming wares and on Japanese porcelain from Arita. Translucent overglaze colours over underglaze blue are a Yung-cheng type that had a revival at this time. In addition, the *rose-verte* palette was commonly used.

In 1853 the Taiping Rebellion led to the destruction of the kilns at Ching-te-chen, which were not rebuilt until 1864. The reign of the T'ung-chih emperor (1862–75) is principally notable for poor copies of earlier monochromes, including the peach-bloom glaze. Nearly all wares from this time onward are slick copies of older work.

KOREA

Because Korea lies to the north of China and close to the islands of Japan, it has usually formed a cultural link between the two countries. During the Japanese invasion of 1592, for instance, many Korean potters were taken to Japan, where they were set to work making tea ceremony wares, which had hitherto been imported, and they later helped to found the porcelain industry.

Silla dynasty (57 BC–AD 935). It is difficult to distinguish some Korean wares from those made in the northern provinces of China during the contemporary Han to T'ang period. The wares of the Silla period include some reminiscent of those of the Chou dynasty. Specimens of stoneware obviously based on metalwork are distantly related to some of the Han bronzes. Patterns on these wares are geometric and incised into the clay before firing.

An olive-green glaze was introduced later in the Silla dynasty, probably about the 9th century. Roof tiles and finials have a brown or a green glaze and may be contemporary with the Han dynasty.

Koryŏ dynasty (918–1392). The wares of the Koryŏ dynasty (roughly corresponding to the Chinese Sung and Yüan dynasties) exhibit a much greater diversity and fall into rather more clearly defined groups. The attribution of certain black-glazed *temmoku* types (see above *China: Sung dynasty*) is controversial, but it seems that at least some were made in Korea. Many celadons, too, have typical Korean lobed forms, based on the melon or the gourd. These are also to be seen in porcelain, much of which has a bluish-white glaze. Some lobed boxes, usually circular, are decorated with impressed designs and are probably always Korean.

One of the difficulties in the study of Korean pottery is that practically everything has been recovered from tombs; few actual kiln sites have been discovered. Nevertheless, one such excavation at Yuch'ŏn-ni has disclosed shards of both the celadon glaze and of white porcelain from which it seems evident that white porcelain resembling both the *ying-ch'ing* and Ting types was made (see above *China: Sung dynasty*). The earliest vessels were probably fairly close copies of Chinese styles, while the distinctive Korean style followed rather later. A crazing of the glaze and a certain amount of flaking are characteristic. A mere handful of specimens, some fragmentary, of inlaid white porcelain have survived. They are best represented by a vase in the Duksoo Palace Museum of Fine Arts in Seoul that has panels of black-and-white inlay beneath a celadon glaze. Decoration on much Korean porcelain of the period is either incised (foliage being a frequent motif), combed, or molded in shallow relief.

White porcelain

Korean celadons have a stoneware body covered with a glaze varying from bluish green to a putty colour; some are obviously analogous to the celadons of Yüeh-chou (see above *China: Han dynasty*). Characteristic of Korean pots are the stilt or spur marks to be seen on the otherwise glazed base; these are the points on which the pots rested in the kiln. Many of the forms are lobed. Perhaps the most important divergence from the usual Chinese celadon is the presence of inlaid decoration beneath the glaze of many specimens, later examples of which are often referred to as *mishima*.

Mishima decoration

The designs were first incised into the clay, and the incisions were then filled with black and white slip. The inlaid patterns are diverse, but most of the subjects are floral; birds are to be seen occasionally. Isolated flowers with symmetrically radiating petals are also found, principally on boxes.

By courtesy of the Victoria and Albert Museum, London; photograph, Wilfrid Walter



Korean bottle with a celadon glaze and *mishima* (inlaid decoration), Koryŏ dynasty, 13th century. In the Victoria and Albert Museum, London. Height 34.6 cm.

Yi dynasty (1392–1910). While most Korean wares of this period are distinctly rougher than those of China in the Ming and Ch'ing periods, the decoration is often magnificent in quality. Most can be clearly distinguished

from Chinese wares by their forms, which show distinct differences in almost every case. Lobed forms suggested by the melon are very characteristic, and the pear-shaped bottle differs in its proportions from that of the Chinese. The large rugged jars with high shoulders are not so precisely potted as similar jars from China, often showing a marked degree of asymmetricality. Twisted rope handles are also peculiar to Korea. Many of the ewers are obvious adaptations from metalwork.

Painting in brownish black beneath a celadon glaze, which had begun in the Koryŏ dynasty, continued in the Yi dynasty. Inlaid decoration was also executed during the early part of this period, the pattern often being engraved by stamps rather than incised freehand. Sgraffito decoration, in which patterns were incised through a grayish-white slip, is also seen occasionally.

Some excellent painted designs in an underglaze blue of variable colour but usually distinctly grayish in tone were executed on a rough procelain body that is almost stoneware. The designs are particularly notable for great economy of brushwork and superb drawing. Their affinities are much more with Japanese pottery than with contemporary Chinese wares. A typical Japanese technique, "brush" (*hakeme*), or brushed slip, is used in conjunction with painted decoration in the early part of the dynasty, but later it is used alone. Korean influence on Japanese pottery was probably at its strongest during the ascendancy of the Japanese warrior Hideyoshi (1536–98), who invaded Korea. It is unlikely that much important work was done in Korea itself after this invasion.

JAPAN

Since Japan is a well-wooded country, wood has always been used for domestic utensils of all kinds, either in a natural state or lacquered. Until recent times, therefore, pottery and porcelain were not employed extensively for general domestic use but were reserved for such special purposes as the tea ceremony (see below). In pottery the Japanese especially admire accidental effects that resemble natural forms. Objects that appear misshapen and glazes that exhibit what would normally be regarded as serious imperfections in the West are admired by the Japanese connoisseur. The Japanese potter liked his work to reveal the impress of the hand that had made it. Marks, such as the ridges left by the fingers in a newly-thrown vessel, were frequently accentuated instead of being obliterated, and marks made by tools were often left untouched.

Accidental effects

Hand modelling was practiced long after the wheel was known, and asymmetries and irregularities of form were purposely sought. Similar accidental effects were encouraged in glazing: coloured glazes were allowed to run in streaks and were irregularly applied. They were often thick, with many bubbles, and with a semifluid or treacly appearance. Crackled glazes and those deeply fissured (the latter called dragon skin or lizard skin) were deliberately induced. Painted decoration, frequently blue, brown, or iron red, is often summary and almost calligraphic in its simplicity. The aim was to give an overall effect that resembled such natural objects as stones, in being largely uncontrived.

From the 15th century onward, the art of the potter was also affected by the elaborate tea ceremony (the *cha-no-yu*). In its original form it was probably introduced from China by Zen priests, but at the court of the shogun (military governor) Yoshimasa (1435–90), in Kyōto, it developed into a fixed ceremonial pattern. Possibly the ceremony was first exploited as a means of settling feudal disputes. It is held in a small room or pavilion, usually surrounded by a carefully designed garden. When the guests are summoned they enter a sparsely furnished room through a very low doorway. The fact that guests must crawl into the room is thought to have served the purpose of preventing them from concealing a sword under their robes.

In a recess called the tokonoma, a picture mounted on brocade or silk is hung, and the guests bow to this in appreciation. The tea master puts a little powdered tea in a bowl and pours on it water that has been heated over a charcoal brazier. The tea is whipped to a froth with a bamboo whisk and then passed from hand to hand. The various utensils (the teabowl, tea caddy, water container, boxes, plates, and iron tea kettle) have been carefully selected by the tea master and are often of great age. The tea drinking is followed by a discussion and appreciation of the qualities of the utensils. The bowls are valued for their heat-retaining properties and the way in which they fit the hand as well as for their appearance. Sometimes a newly acquired work of art is produced by the host for the delectation of his guests. Since the tea masters were the aesthetic arbiters, their influence on Japanese pottery was profound.

Early history. The early history of Japan is considerably more obscure than that of China. The first Japanese pottery belongs to the Jōmon period (*c.* 2500–250 BC). It has a black body, and the decoration is usually an impressed representation of coiled rope or matting (*jōmon* means "coiled"). *Jōmon-shiki* ("pottery") is widely distributed throughout the islands, but complete specimens are very rare. It was followed by Yayoi pottery, specimens of which have been excavated throughout Japan. The body is somewhat finer in quality than Jōmon pottery and is usually red or gray. Decoration is simple, and forms will sometimes show the influence of Korean pottery of the period. It ceased to be produced about the 6th century AD.

Jōmon pottery

Meanwhile, from about the 3rd to the 6th century AD, large tombs were constructed in the form of oval or circular tumuli from whose bases have been recovered the *haniwa* ("clay circle") figures of warriors, women, horses, and so forth. They are hollow, and, though vigorously modelled, they are more primitive than analogous tomb figures from China.

In the Asuka or Sueki period (AD 552–710) that followed, wares are much more sophisticated. Unlike the preceding types, they were made with a wheel, and firing took place in a rudimentary kiln at a much higher temperature than previously. Widespread manufacture continued through the Nara period (710–794) and the early part of the following Heian, or Fujiwara, period (794–1185). Some examples have a smear glaze, no doubt at first caused accidentally by wood ashes coming into contact with the surface. Three colours of glaze—green, yellowish brown, and white—were used either alone or in combination and resemble those of T'ang earthenware. Pottery of this kind has been found around Ōsaka and Kyōto. The principal pottery productions of the period were vases, dishes, bowls, and bottles of various descriptions.

The influence of Korea and of T'ang China is noticeable. Toward the end of the Heian period contacts with China were severed, and there was a corresponding decline in the art of pottery; even the traditional Sueki ware disappeared.

Kamakura and Muromachi periods (1192–1573). A revival in the Kamakura period (1192–1333) followed the visit of the potter Katō Shirōzaemon (Tōshirō) to China in 1227, where he learned the secrets of pottery making. He established himself at Seto, Owari (now Aichi Prefecture), which speedily became a large centre of manufacture. There were soon about 200 kilns in the vicinity making a variety of wares, some of which were glazed in black in imitation of the *temmoku* wares of China (see above *China: Sung dynasty*). The early wares were mainly for ritual purposes, but by the beginning of the Muromachi, or Ashikaga, period (1338–1573) teabowls, plates, jars, and saucers of domestic utility were also being made. Wares of the Kamakura period are decorated with incised designs or with impressed or applied ornament. The Muromachi wares are much plainer as the result of the growing influence of the tea ceremony, especially the *wabi* school of the cult, which concentrated on rustic simplicity. The wares of both of these periods have a feldspathic glaze, but the Muromachi glaze is more even in quality than the Kamakura, which has a tendency to run in rivulets. A transitional type has a soft-yellowish glaze or a dark-brown glaze sometimes called Seto *temmoku*.

A large number of kilns were in existence, the more important known as the "six pottery centres of ancient Japan." These were Seto; Tokoname (also in Aichi Prefecture), which may have exceeded Seto in the size of its production; Bizen (Okayama Prefecture), which has produced an excellent unglazed stoneware from the Heian period to the 20th century; Tamba (Kyōto Prefecture); Shigaraki (Shiga Prefecture); and Echizen (Fukui Prefecture). The wares of Seto, especially those made for Buddhist ceremonies, were regarded as the finest pottery of this period.

Seto–Mino wares

Momoyama period (1574–1600). Production had been interrupted during the civil wars of the 15th and 16th centuries. Toward the end of the 16th century the Seto kilns were removed for a time to the Gifu Prefecture of Mino Province, where they received the protection of the feudal baron (daimyo) of Toki. The Mino pottery was founded by Katō Yosabei, whose sons started other potteries in the vicinity, notably that under the aegis of the tea master Furuta Oribe Masashige. New kilns were also built elsewhere, and pottery, while retaining its importance in the tea ceremony, became much more widely used for ordinary purposes. The inspiration for most of its shapes and designs came from the Mino region. The later wares of these kilns are much less austere than those attributed to the Muromachi period, since the cult of the tea ceremony, now widespread, had lost something of its earlier simplicity. Characteristic tea ceremony wares of the early years of the 17th century are Shino, which has a thick, crackled glaze and is sometimes summarily painted in blue or brown; yellow Seto (*ki-Seto*), whose crackled yellow glaze covers a stoneware body; and, at Narumi, in the adjoining Owari region, a ware of the kind associated with Oribe (which had become a generic term for pottery influenced by the tea master of that name), which is glazed in white, straw colour, yellowish green, and pinkish red, with sometimes the addition of slight painting in brown.

Toward the end of the 16th century the tea ceremony was reformed by Sen Rikyū (1521–91), the tea master to the military dictator of Japan, Toyotomi Hideyoshi. Sen Rikyū was principally responsible for the replacement of the hitherto much admired *temmoku* bowls from China by others patterned after unsophisticated Korean wares; his influence has persisted to the 20th century. In the 1590s Hideyoshi twice invaded Korea, and as a result of these wars many Korean potters were taken to Japan, where their influence was considerable.

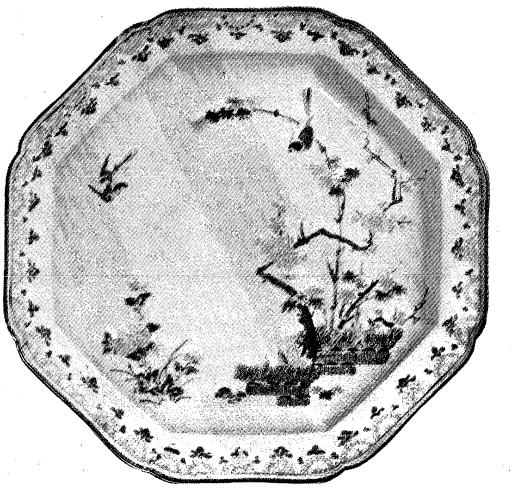
By courtesy of the Museum of Fine Arts, Boston



Raku ware water jar and cover, Tokyo, Edo period, 18th century. In the Museum of Fine Arts, Boston. Height 20.3 cm.

Raku ware

A tilemaker named Ameya, who is said to have been a Korean, introduced a new type of ware that was covered with a lead glaze and fired at a comparatively low temperature. His son Tanaka Chōjirō and his family extended this technique to the tea ceremony bowl, and, about 1588 their wares were brought to the notice of Hideyoshi, who awarded them a gold seal engraved with the word *raku* ("felicity"). The *raku*, made in Kyōto, are among the most famous of all Japanese wares. The shape of the vessels is extremely simple: a wide straight-sided bowl set on a narrow base. At first the glaze was dark brown, but a light orange red was developed later, to be followed in the 17th century by a straw colour. Still later, green and cream and other colours were introduced. Teabowls attributed to the first Chōjirō are greatly valued in Japan.



Kakiemon ware decorated with iron red and coloured enamels, Hizen Province, Japan, Edo period, 17th century. In the Metropolitan Museum of Art, New York. Diameter 24.4 cm.
By courtesy of the Metropolitan Museum of Art, New York, Rogers Fund, 1918

Karatsu ware

The kilns of Karatsu, a district in the north of Hizen Province, may have been established by Korean potters, since the influence of Korea is perceptible in some of them. The term Karatsu ware encompasses a great variety of shapes and styles: "undecorated" (*muji*), "painted" (*e*), "speckled" (*madara*), in the Korean style (Chosen), which has a thick opaque glaze, and in the style of Seto (Seto), which has a white glaze. The earliest Karatsu ware belongs probably to the end of the 16th century, although it is sometimes awarded a still earlier date. Most surviving examples belong to the 17th century. The most valued pieces are those made for the tea ceremony.

Edo Period (1603–1867). According to tradition, the first Japanese porcelain was made at the beginning of the 16th century after Shonzui Goradoyu-go brought back the secret of its manufacture from the Chinese kilns at Ching-te-chen. Another account claims that a Korean potter called Ri Sampei, who was brought to Japan by Hideyoshi, discovered porcelain clay in the Izumi Mountain, near Arita (Saga Prefecture); this version is feasible, since no porcelain made before the end of the 16th century has been identified.

The first Arita manufacture was decorated in blue underglaze, simple and excellent in quality. Specimens soon found their way to Europe in Dutch ships, and the Dutch were awarded a trading monopoly in 1641. Some of these early Japanese export wares are based on contemporary European metalwork and faience.

The family of Sakaida is especially connected with the Arita kilns. The first recorded member, born about 1596, worked in underglaze blue until the family learned the secret of using overglaze colours. According to tradition, it was told to them by a Chinese met by chance in the port of Nagasaki. This overglaze technique was perfected soon after the middle of the 17th century. It was continued by the family, and, since many of them were called Kakiemon, the style has become known by that name. The palette is easily recognized—iron red, bluish green, light blue, yellow, and sometimes a little gilding; many examples have a chocolate-brown rim. Octagonal and

Kakiemon porcelain

square shapes are especially frequent. Themes of decoration are markedly asymmetrical, with much of the white porcelain surface left untouched. This technique and style spread rapidly to other provinces, and its influence on porcelain that was manufactured in Europe during the first half of the 18th century was at least as great as that of Chinese porcelain. A later Kakiemon development in which "brocade" (*nishikide*) patterns in compartments were used (at the suggestion of Dutch traders) prove to be less pleasing. These later coloured wares from Arita became known as Imari, after the port from which they were shipped.

Like 18th-century Chinese white porcelain, Japanese white wares were shipped to Europe, where they were subsequently decorated by Dutch and other European enamellers.

Kutani porcelain

Of considerable importance but more rarely seen in Europe is the porcelain called Kutani. The kilns at Kutani-mura (Ishikawa Prefecture) were founded about 1652 and operated for approximately 60 years. Greatly valued, old-Kutani or *ko*-Kutani, porcelain is among the finest of the Japanese wares. The body is heavy, approaching stonewares, and the designs are executed boldly and in rich colours.

The Mikawachi kilns under the protection of the prince of Hirado made porcelain principally for his use. The delicate, very white body is usually decorated in miniature style with underglaze blue. Kyōto imitated Sung celadons and the Ming green and red wares. Seto made no porcelain until about 1807; the first production was decorated in underglaze blue (*sometsuke*). Overglaze colours date from about 1835.

The manufacture of earthenware was continued during the 17th and 18th centuries, and much of it is notable for its decoration. Toward the end of the 17th century Ninsei (Nonomura Seisuke) began work at Kyōto and was responsible for much finely enamelled decoration on a cream earthenware body covered with a finely crackled glaze. Also at Kyōto, the work of Kenzan, who used rich and subtly coloured slips often as a background for plant motives, and of the Dōhachi family, famous for their overglaze decoration, are among the wares which are much sought in Japan.

19th and 20th centuries. Japanese productions during the 19th century, in common with those in most other parts of the world, greatly deteriorated in taste. A typical product of the period is the so-called Satsuma pottery, most of which was not made at Satsuma, but at Kyōto, and was then sent to Tokyo to be decorated especially for export. The designs are overcrowded and debased, and its popularity undoubtedly retarded an appreciation of work in the true Japanese taste among Western students and collectors.

Like Western pottery manufacture in the mid-20th century, Japan's is largely industrialized, and most products are derivative, but the Japanese tradition of pottery making in small and private kilns continues.

THAILAND AND ANNAM

Pottery was made in the old Siamese capitals of Sukhothai and Sawankhalok. It is also thought that potteries persisted at Ayutthaya until the 18th century. Little is known of the early history of the region, and definite information on its pottery is almost nonexistent. Dating of the pottery from these regions for the most part has been by analogy with related Chinese wares, which greatly influenced Siamese work.

Kilns have been excavated on the site of old Sawankhalok, about 200 miles (320 kilometres) north of Bangkok. The principal type of ware is a grayish-white stoneware covered with a translucent celadon glaze, usually grayish green in colour. The glaze is commonly crackled; this appears to be fortuitous, since little trouble was taken to achieve a precise finish. A particularly common decoration consists of roughly scored vertical flutes, with incised circles at the shoulder to accentuate the form. Decoration of a more definite kind is always incised under the glaze and is usually floral. Flowers are stylized, sometimes with combed lines on the petals. Covered bowls, dishes, ewers, and bottles with two small loop handles at the neck are the most common forms.

Another type of ware, with similar forms, has a rather treacly-brown glaze. Some well-modelled animals are found with this glaze. There are also tiles and bricks with crudely modelled figures in relief on them. They are analogous in form and technique to Chinese pottery of the Sung dynasty and are generally regarded as being contemporary with the Sung or Yüan period. Some small covered jars of a gray porcelaneous ware, summarily decorated with stylized floral and foliate patterns, appear to have been made at Sawankhalok (the date is probably equivalent to that of the early Ming period). These Martabani wares were widely exported throughout the East during this period.

Annam. Little is known of wares made in Annam, but some brownish celadons are regarded as likely to have been made there, as well as some small covered jars painted in a poor underglaze blue.

IV. American Indian pottery

The American Indians are of Asiatic descent; their route to the New World was from Siberia into Alaska across the Bering Straits. The usually quoted period of their migration is between 40,000 and 10,000 years ago. Since they were nomadic peoples, it is unlikely that they brought the knowledge of pottery making with them. When pottery making did begin, it was fundamentally unlike any known work from the Old World, and the few remote resemblances to Oriental motifs are almost certainly fortuitous. The wheel remained unknown until the arrival of Europeans, although there is reason to think that a turntable, or slow wheel, may have been used occasionally. Most of the pottery was made by coiling, some by molding—both are techniques that could have arisen spontaneously. It is likely that most of the work was done by women rather than by the men. This is nearly always the case with primitive potters when the wheel is not used, and Pueblo Indian women still do this kind of work.

Slips were used to cover the body, and coloured slips provided the material for much of the painted freehand decoration. Glazes are rare, although examples can be found among the Pueblo Indians of New Mexico from about AD 1300 onward, on a few vessels from the Chimú area in the Andes, and occasionally in Central America. The effect of a reducing atmosphere was understood, so that gray and black pots are found as well as the red and brown ones fired in an oxidizing flame. Undecorated surfaces were often highly polished.

NORTH AMERICA

The most important North American pottery was made in the southwest—an area including Arizona, New Mexico, and also parts of Utah and Colorado. The people who inhabited the plateau land from about 100 BC are often referred to as the Anasazi, a Navaho Indian word meaning ancient people. They are the ancestors of the Pueblo, who began to emerge about AD 700. The Anasazi were nomadic hunters; although they did not at first make pottery, they did make excellent baskets. Fixed dwellings appear about AD 50, and this probably marks the beginning of pottery manufacture. The earliest pots appear to have been baskets that were smeared with clay and then dried in the sun.

Next came basket-shaped wares coiled in a gray body, used principally for cooking. They were followed by more decorative bowls and pots, with striking black and white geometric designs that seem to have been executed about AD 700. Slightly later there is another type of ware that has black decoration on a red slip. After the 12th century the earlier types began to disappear and were replaced by polychrome wares decorated with stylized birds, feathers, animals, and human figures amid the geometric patterns. The principal colours are yellow and red. A small quantity of glazed ware was made in the Zuñi area of New Mexico.

The Hohokam tribes (a Pima word meaning "those who have gone"), who lived in the desert of southern Arizona

Pottery figures for religious purposes

and were approximately contemporary with the Anasazi, made pottery figures for religious purposes, usually of crudely modelled naked women. Some of this pottery is a gray ware, but most of it is buff, with decoration in iron red that has a quality lacking the stiffness of the Pueblo designs.

The Mogollon culture of New Mexico produced, during the Mimbres period of the 11th and 12th centuries, a ware remarkable for its lively black and white decoration depicting human, animal, and insect forms in a much less stylized manner than the paintings on most other wares from the southwest.

There is little pottery of importance from other parts of the United States. Primitive pots have been found on the Atlantic coast, in Georgia and Florida, on the Gulf Coast, and elsewhere, some of which are based on basketwork. Geometric decoration, usually incised, is the rule. Eskimo pottery, which is generally rather crude, bears some resemblance to early Asiatic types.

CENTRAL AMERICA

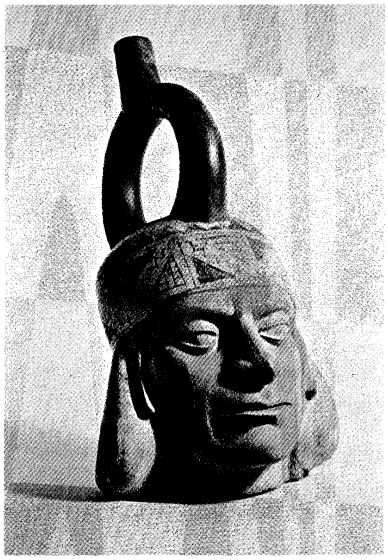
The pottery of Mexico and the rest of Central America is of considerable interest, but the wares are so diverse that it is impossible to summarize them adequately. They probably date from the 2nd millennium BC onward and were made by the Mayas, the Zapotecs, the Toltecs, and the Aztecs. Generally speaking, geometric patterns are common, and slips in black, brown, white, or red were frequently used. A curiosity of Central America (possibly adopted from South America) is a technique that resembles to some degree the batik method of dyeing textiles.

Maya, Zapotec, Toltec, and Aztec pottery

The surface of the pot was coated either with wax or gum. This was then scraped away in part to form a predetermined pattern, and the whole surface of the pot was covered with pigment. In firing, the gum burned away, leaving only the scraped parts in colour. Ornament carved in low relief after firing is to be seen occasionally and has few parallels outside the Americas. An unusual technique from the Mexican highlands consisted of covering the whole surface of the pot with a kind of thick slip, most of which was then scraped away, leaving only thin partitions. These compartments were filled with slips of a contrasting colour. The commonest shapes are bowls and wide-mouthed vases; many of these were made with legs, usually three, so that they could be set down on uneven ground. Figurines, some of which are painted, have also been found.

Between about 600 BC and AD 1000, the Mayas were making an excellent polychrome pottery in which designs in red and black were painted on a cream or orange slip.

By courtesy of the trustees of the British Museum



Stirrup spout vessel in the form of a portrait head, Mochica culture, Peru (100 BC–AD 900). In the British Museum. Height 21.6 cm.

Between roughly the 4th and the 10th centuries AD, the Zapotecs, whose chief ceremonial site was Monte Albán in Oaxaca State, made striking urns in the forms of their gods. An orange-coloured pottery decorated in a great diversity of styles is associated with the Toltecs, as is a dark-coloured pottery with a glossy appearance and incised ornament (plumbate ware). Both of these types were widely distributed throughout Central America from the 11th to 14th centuries. Little is known of the Aztecs until about 1325, the date of the foundation of Tenochtitlán (Mexico City). Much of their later pottery utilizes an orange-burning clay that was painted with black curvilinear geometric motifs, in contrast to their earlier rectilinear style. During the period of Montezuma I in the 15th century, designs became more naturalistic, and birds, fish, and plant forms were freely utilized. European motifs first appear after the conquest, and such techniques as tin glazing were used from the 17th century onward.

SOUTH AMERICA

Most South American pottery was made at centres in the Andes and on the west coast, particularly in Bolivia and Peru. Pottery of lesser importance comes from Ecuador, Colombia, northwest Argentina, and northern Chile. In some places a very high degree of skill was attained, especially in the central Andes, where the earliest wares seem to date from the end of the 2nd millennium BC. Much of the pottery was made in molds. The stirrup-shaped spout on many jars is a characteristic feature. The batik type of decoration already mentioned was also used. Vessels were modelled in the shape of animal or human figures, which were also used as motifs for painted decoration. The puma god worshipped by the early peoples appears in many forms. Depictions of erotic themes also appear in South American pottery, particularly in the Mochica and Chimú cultures.

Mochica culture

The work of the Mochica culture, which flourished around the northern coast of Peru, is at its best about the 7th century AD. Jars in the form of human heads, some of which may be portraits, are remarkable both for the naturalism of the treatment and the skill of the potter. These have the stirrup spout. Painted decoration is often stylized although with a considerable degree of realism, and the subjects are nearly always ceremonial or religious nature.

The pottery of the Nazcas, who lived on the southern coast of Peru at much the same period, is noted for its painting. A varied palette included several shades of red and blue, yellow, orange, green, brown, black, gray, and white. The stirrup spout here becomes two spouts joined by a flat bridge. The earliest painting is on a red ground, white grounds becoming more common later. Geometric patterns are to be seen in conjunction with stylized birds, human heads, and the like. The naturalistic portrait jars of the Mochica do not appear, but there are some vessels in the form of figures modelled in a much more conventional style and similarly painted. Puma's heads occur in relief, with the body of the animal completed in brushwork. The centipede god is a motif that does not appear elsewhere.

The people of Tiahuanaco, who lived in the region around Lake Titicaca, were influenced by the Nazca wares, though painted decoration, often carried out on a red slip ground, is more limited in colour than the Nazca. The puma head was used as a motif. In general, decoration is extremely stylized with a very strong geometric flavour.

Chimú culture

The Chimú culture succeeded the Mochica in the northern area and lasted until the arrival of the Incas. The most familiar ware is in a body that varies from gray to black, although a red polished ware, sometimes painted in white slip, was also made. The influence of the Mochica tradition can be seen in the retention of the stirrup spout on some jars; others have the double spout connected by a flat bridge. The modelled wares of the Mochica culture were also revived but are of a generally inferior quality.

The Incas originally settled in Cuzco, the old capital of

Peru, at the end of the 11th century. During the 15th century they established themselves firmly over a wide area, including the territory of the Chimús. They were principally soldiers and administrators with small inclination toward luxury; and their pottery, of excellent quality particularly in the 15th century, is designed in good taste without an excess of decoration. Most Inca pottery is red polished ware. It is usually painted with geometric designs in red, white, and black, although relief decoration is also to be seen on black ware, especially from the Chimú region.

The commonest surviving form has been called an aryballos, although its resemblance to the Greek form is remote and fortuitous. It has a conical base, and the neck finishes in a flaring mouth. Two loop handles are set low on the body. The assumption that this vessel was made for carrying water on the back seems a little doubtful in view of its shape and the disposition of the handles. Little fine pottery was made after the arrival of the Spaniards in the 16th century.

BIBLIOGRAPHY. Books on pottery are fairly numerous, but many have been written in recent years for the popular market and are not always reliable or duplicate information. The following titles may be regarded as standard or major works.

General works: Two good introductory surveys are GEORGE SAVAGE, *Pottery Through the Ages* (1963) and *Porcelain Through the Ages*, 2nd ed. (1963). ROBERT J. CHARLESTON (ed.), *World Ceramics* (1968), is a lavishly illustrated history of ceramics written by many noted specialists in Europe and the United States. Highly recommended is W.B. HONEY, *European Ceramic Art, from the End of the Middle Ages to about 1815*, 2nd ed., 2 vol. (1963), which has a comprehensive list of marks and an excellent bibliography. Also useful is EMIL HANNOVER, *Keramisk Haandbag*, 2 vol. (1919–24; Eng. trans., *Pottery and Porcelain*, 3 vol., 1925); and WARREN COX, *The Book of Pottery and Porcelain* (1944). GEORGE SAVAGE, *The Dictionary of Antiques* (1970), includes information about continental wares on a less comprehensive scale, but its extensive bibliography has been brought up to date. It also differentiates between works in print and those available only in libraries or secondhand. For the technical side of pottery, see BERNARD LEACH, *A Potter's Book* (1940); W.B. HONEY, *The Art of the Potter* (1946); and PAUL RADO, *An Introduction to the Technology of Pottery* (1969). Factory marks are discussed and listed in W.B. HONEY and JOHN P. CUSHION, *Handbook of Pottery and Porcelain Marks*, 3rd rev. ed. (1965); Cushion has also published separate handbooks on *English Ceramic Marks and Those of Wales, Scotland, and Ireland* (1959), on *German Ceramic Marks and Those of Other Central European Countries* (1961), and on *French and Italian Ceramic Marks* (1965). GEOFFREY GODDEN, *Encyclopaedia of British Pottery and Porcelain Marks* (1964), is another recommended source. WILLIAM CHAFFERS, *Marks and Monograms on Pottery and Porcelain*, 14th ed. (1931), formerly the standard work, has now been superseded and is sometimes inaccurate. More specialized is ARTHUR BEHSE, *Deutsche Fayencemarken–Brevier* (1955), which should be consulted for German wares. Ceramic terms are defined by GEORGE SAVAGE and H. NEWMAN in the *Illustrated Dictionary of Ceramic Terms* (1973).

Ancient Near East and Egypt: Books dealing specifically with the pottery of this area and period are scarce. See WALTER ANDRAE (ed.), *Coloured Ceramics from Assur* (1925); and HENRY WALLIS, *Egyptian Ceramic Art* (1898). More information will be found in books dealing with the arts of the ancient Near East and Egypt.

Ancient Aegean, Greece, and Italy: Works on ancient pottery from these regions are fairly numerous. A comprehensive selection is listed below. HENRY B. WALTERS, *A History of Ancient Pottery*, 2 vol. (1905, reprinted 1971); CHARLES T. SELTMAN, *Attic Vase-Painting* (1933); VINCENT D'ARBA DESBOROUGH, *Protogeometric Pottery* (1952); G.M.A. RICHTER, *A Handbook of Greek Art*, 6th ed. (1969); *Ancient Italy* (1955); *Attic Red-Figured Vases: A Survey*, 2nd ed. (1958); and *The Craft of Athenian Pottery* (1923); ARTHUR LANE, *Greek Pottery*, 3rd ed. (1971); ROBERT M. COOK, *Greek Painted Pottery* (1960); T.B.L. WEBSTER, *Greek Terracottas* (1951); ROBERT J. CHARLESTON, *Roman Pottery* (1955); JOHN D. BEAZLEY, *Etruscan Vase Painting* (1947).

Islāmic pottery: There are few works easily available. The best are ARTHUR LANE, *Early Islamic Pottery*, rev. ed. (1958) and *Later Islamic Pottery*, 2nd ed. (1972); see also MAURICE S. DIMAND, *A Handbook of Muhammadan Art*, 3rd ed. rev. (1958); and ROBERT L. HOBSON, *A Guide to Islamic Pottery of the Near East* (1932).

Western pottery: Where they exist, the titles of English works on continental pottery are given below. Those listed in other languages are illustrated works that will supplement the deficiency of works in English. In the case of porcelain, English works that discuss the wares of individual factories are given only in special cases since they are extremely numerous and most general works on the subject include a bibliography. (*Spain*): ALICE W. FROTHINGHAM, *Catalogue of Hispano-Moresque Pottery in the Collection of the Hispanic Society of America* (1936); *Talavera Pottery, with a Catalogue of the Collection of the Hispanic Society of America* (1944); *Lustreware of Spain* (1951); ALBERT VAN DER PUT, *Hispano-Moresque Ware of the XVth Century* (1904). (*Italy*): JOSEPH CHOMPRET, *Répertoire de la Mojolique Italienne*, 2 vol. (1949); BERNARD RACKHAM, *Guide to Italian Maiolica* (1933); *Catalogue of Italian Maiolica*, 2 vol. (1940); and *Italian Maiolica* (1952)—Rackham's books are the best source of information in English; ARTHUR LANE, *Italian Porcelain* (1954); GIUSEPPE MORAZZONI, *Le Porcellane Italiane* (1935). (*France and Belgium*): CHARLES DAMIRON, *La Faïence artistique de Moustiers* (1919) and *La Faïence de Lyon* (1926); JEANNE GIACOMOTTI, *Faïences françaises* (1963; Eng. trans., *French Faïence*, 1963); HANS HAUG, *Les Faïences et porcelaines de Strasbourg* (1922); ARTHUR LANE, *French Faïence*, 2nd ed. (1970), the best introduction to the subject; FRANCOIS PONCETTON and GEORGE SALLES, *Les Poteries Françaises* (1928); PAUL ALFASSA and JACQUES GUERIN, *Porcelaine française du XVII^e au milieu du XIX^e siècle* (1932); EMILE BOURGEOIS, *Le Biscuit de Sèvres au XVIII^e siècle*, 2 vol. (1909); W.B. HONEY, *French Porcelain of the 18th Century* (1950); EUGENE J. SOIL DE MORAIME, *La Manufacture impériale et royale de porcelain de Tournay*, 3rd ed. of his *Céramique tournaisienne* by LUCIEN DELPLACE DE FORMANOIR (1937); GEORGE SAVAGE, *Seventeenth and Eighteenth Century French Porcelain* (1960); PIERRE VERLET, SERGE GRANDJEAN, and MARCELLE BRUNET, *Sèvres* (1953). (*Germany and Austria*): KARL KOETSCHAU, *Rheinisches Steinzeug* (1924); HANS MEYER, *Böhmisches Porzellan und Steingut* (1927); GUSTAV E. PAZAUREK, *Deutsche Fayence und Porzellan-Hausmaler*, 2 vol. (1925, reprinted 1970) and *Steingut: Formgebung und Geschichte* (1921); O. RIESEBIETER, *Die deutschen Fayencen des 17. und 18. Jahrhunderts* (1921), the most comprehensive general work on the subject; EDMUND W. BRAUN and JOSEPH FOLNESICS, *Geschichte der K.K. Wiener Porzellan-Manufaktur* (1907); HANS CHRIST, *Ludwigsburger Porzellanfiguren* (1921); JOHN F. HAYWARD, *Viennese Porcelain of the Du Paquier Period* (1952); FRIEDRICH H. HOFMANN, *Frankenthaler Porzellan*, 2 vol. (1911); *Geschichte der Bayerischen Porzellan-Manufaktur Nymphenburg*, 3 vol. (1921–23); and *Das Porzellan der Europäischen Manufakturen im 18. Jahrhundert* (1932); W.B. HONEY, *Dresden China*, 2nd ed. (1954) and *Germain Porcelain* (1948); GEORG LENZ, *Berliner Porzellan: Die Manufaktur Friedrichs des Grossen, 1763–1786*, 2 vol. (1913); HUGO MORLEY-FLETCHER, *Antique Porcelain in Color: Meissen* (1971); E. POCHE, *Bohemian Porcelain* (n.d.); KARL ROEDER and MICHE OPPENHEIM, *Das Höchster Porzellan* (1925); MAX SAUERLANDT, *Deutsche Porzellanfiguren des XVIII. Jahrhunderts* (1923); GEORGE SAVAGE, *18th–Century German Porcelain*, 2nd ed. (1967); CHRISTIAN SCHERER, *Das Fürstenberger Porzellan* (1909). (*Switzerland*): SIEGFRIED DUCRET, *Zürcher Porzellan des 18. Jahrhunderts* (1944). (*The Netherlands*): CAROLINE H. DE JONGE, *Delft Aardewerk* (1965; Eng. trans., DELFT-CERAMICS, 1969) and *Nederlandse tegels* (1971; Eng. trans., *Dutch Tiles*, 1971); FERRAND W. HUDIG, *Delfter Fayence* (1929); JEAN JUSTICE, *Dictionnaire des marques et monogrammes de la faïence de Delft* (1920; Eng. trans., *Dictionary of Marks and Monograms of Delft Pottery*, 1930); ELISABETH NEURDENBURG, *Old Dutch Pottery and Tiles* (Eng. trans., 1923); BERNARD RACKHAM, *Early Netherlands Maiolica* (1926), a discussion of the earliest pre-Delft wares. (*Scandinavia*): RICHARD MARSSON, *Die Stralsunder Fayencefabrik, 1757–1790* (1928); ARTHUR HAYDEN, *Royal Copenhagen Porcelain* (1911); ERIK WETTERGREN, *The Modern Decorative Arts of Sweden* (Eng. trans. 1926). (*Russia*): GEORGY LUKOMSKY, *Russisches Porzellan, 1774–1923* (1924); MARVIN C. ROSS, *Russian Porcelains* (1968). (*England*): HARRY BARNARD, *Chats on Wedgwood Ware* (1924); GEOFFREY BEMROSE, *Nineteenth Century English Pottery and Porcelain* (1952); FREDERICK H. GARNER, *English Delftware* (1948); JOHN E. and EDITH HODGKIN, *Examples of Early English Pottery, Named, Dated and Inscribed* (1891); ELIZA METEYARD, *The Life of Josiah Wedgwood*, 2 vol. (1865); ERNEST MORTON NANCE, *The Pottery and Porcelain of Swansea and Nantgarw* (1942); W.J. POUNTNEY, *Old Bristol Potteries* (1920); E. STANLEY PRICE, *John Sadler, a Liverpool Pottery Printer* (1949); BERNARD RACKHAM, *Mediaeval En-*

glish Pottery (1947) and *Early Staffordshire Pottery* (1951); BERNARD RACKHAM and HERBERT READ, *English Pottery: Its Development from Early Times to the End of the Eighteenth Century* (1924); GEORGE W. RHEAD, *The Earthenware Collector* (1920); JOSIAH WEDGWOOD, *Selected Letters*, ed. by ANN FINER and GEORGE SAVAGE (1965); DONALD TOWNER, *Handbook of Leeds Pottery* (1951); FREDERICK WILLIAMSON, *The Derby Pot Manufactory Known as Cockpit Hill* (1931); HUGH WAKEFIELD, *Victorian Pottery* (1962); PATTERSON D. GORDON PUGH, *Staffordshire Portrait Figures and Allied Subjects of the Victorian Era* (1971); MURIEL ROSE, *Artist-Potters in England* (1955); JOSEPH L. DIXON, *English Porcelain of the Eighteenth Century* (1952); ENGLISH CERAMIC CIRCLE, *English Pottery and Porcelain: Commemorative Catalogue of an Exhibition Held at the Victoria and Albert Museum* (1949); STANLEY W. FISHER, *The Decoration of English Porcelain* (1954) and *English Blue and White Porcelain of the 18th Century* (1947); GEOFFREY GODDEN, *An Illustrated Encyclopaedia of British Pottery and Porcelain* (1966); W.B. HONEY, *Old English Porcelain* (1948); WILLIAM KING, *English Porcelain Figures of the Eighteenth Century* (1925); GEORGE SAVAGE, *18th-Century English Porcelain*, new ed. (1964); CYRIL COOK, *The Life and Work of Robert Hancock* (1948); *William Duesbury's London Account Book, 1751–1753*, with an introduction by MRS. DONALD MCALISTER (1931); WILLIAM H. TAPP, *Jefferyes Hamett O'Neal, 1734–1801* (1938). (*American Indian pottery*): G.H.S. BUSHNELL and ADRIAN DIGBY, *Ancient American Pottery* (1955). (*United States*): JOHN RAMSAY, *American Potters and Pottery* (1939); EDWIN A. BARBER, *Tulip Ware of the Pennsylvania-German Potters* (1903); WARREN COX, *The Book of Pottery and Porcelain*, vol. 2 (1944).

Non-Western pottery: (*China*): The best general work is still W.B. HONEY, *The Ceramic Art of China, and Other Countries of the Far East* (1945); but ROBERT L. HOBSON, *Chinese Pottery and Porcelain*, 2 vol. (1915), is often consulted. Hobson's *Wares of the Ming Dynasty* (1923) and *Later Ceramic Wares of China* (1925), are scholarly works, well illustrated for their period. *Chinese Art*, 4 vol. (1960–65), by a number of internationally known experts, covers the whole field of Chinese art, including pottery and porcelain, and is of considerable value to the student. The earliest Chinese wares are discussed in English by CHIN-TING WU in *Prehistoric Pottery in China* (1938). Wares till the Sung dynasty are discussed by BASIL GRAY in *Early Chinese Pottery and Porcelain* (1952); and by ARTHUR L. HETERINGTON in *Early Ceramic Wares of China* (1922) and *Chinese Ceramic Glazes*, 2nd rev. ed. (1948). For wares of the Ming dynasty and subsequently, see ROGER SOAME JENYNS, *Ming Pottery and Porcelain* (1953) and *Later Chinese Porcelain*, 2nd ed. (1959). (*Korea*): G. ST. G.M. GOMPERTZ, *Korean Celadon, and Other Wares of the Koryŏ Period* (1963); and CHEWON KIM, *The Ceramic Art of Korea* (1961); W.B. HONEY, *Corean Pottery* (1947). (*Japan*): Japanese porcelain is better documented than the pottery of that country. ROGER SOAME JENYNS, *Japanese Porcelain* (1965), is the definitive work in English on the subject. See also TAOANARI MITSUOKA, *Ceramic Art of Japan* (1949); and ROGER SOAME JENYNS, *Japanese Pottery* (1971). (*Chinese and Japanese export porcelain*): This includes works on the erroneously termed Oriental Lowestoft. MICHEL BEURDELEY, *Porcelaine de la compagnie des Indes* (1962; Eng. trans., *Porcelain of the East India Companies;* U.S. title, *Chinese Trade Porcelain;* 1962); SIR HARRY M. GARNER, *Oriental Blue and White*, 3rd ed. (1970), discusses early blue-and-white export wares; JOHN A. LLOYD HYDE, *Oriental Lowestoft* (1954); JEAN MCCLURE MUDGE, *Chinese Export Porcelain for American Trade, 1785–1835* (1962); JOHN G. PHILLIPS, *China Trade Porcelain* (1956); WALTER A. STAEHELIN, *The Book of Porcelain* (1966), a book of Chinese illustrations relating to the export trade in the 18th century; SIR A. TUDOR CRAIG, *Armorial Porcelain of the Eighteenth Century* (1925); T. VOLKER, *Porcelain and the Dutch East India Company* (1954), an examination of the records in relation to the import of Chinese and Japanese porcelain in the 17th century.

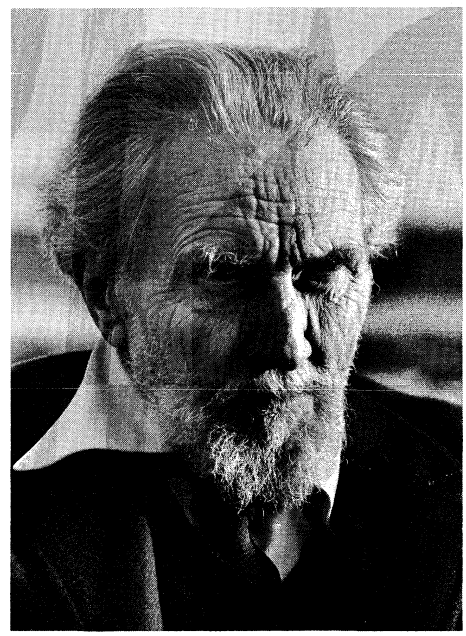
(Ge.S)

Pound, Ezra

One of the most important poets and critics of his time, as well as a supremely discerning and energetic entrepreneur of the arts, Ezra Pound did more than any other single figure to advance a "modern" movement in English and American literature by promoting—and also by occasionally helping to shape—the work of such widely different poets and writers as W.B. Yeats, James Joyce, Ernest Hemingway, Robert Frost, D.H. Lawrence, and T.S. Eliot.

Ezra Loomis Pound was born on October 30, 1885, in the small mining town of Hailey, Idaho, the only child of a Federal Land Office official, Homer Loomis Pound of Wisconsin and Isabel Weston of New York City. About 1887 the family moved to the eastern states, and in June 1889, following Homer Pound's appointment to the U.S. Mint in Philadelphia, they settled in nearby Wyncote, where Pound lived a normal middle class childhood. After two years at Cheltenham Military Academy, which he left without graduating, he attended a local high school. From there he went for two years (1901–03) to the University of Pennsylvania, where he met his lifelong friend, the poet William Carlos Williams. He took his Ph.B. (bachelor of philosophy) degree at Hamilton College, New York, in 1905 and returned to the University of Pennsylvania for graduate work. He received his M.A. in June 1906 but withdrew from the university after working one more year toward his doctorate. He left with a knowledge of Latin, Greek, French, Italian, German, Spanish, Provençal, and Anglo-Saxon, as well as of English literature and grammar.

Life Magazine—David Lees, © 1964, Time Inc.



Ezra Pound.

In the autumn of 1907, Pound became professor of Romance languages at Wabash Presbyterian College, Crawfordsville, Indiana. Although his general behaviour fairly reflected his Presbyterian upbringing, he was already writing poetry and was affecting a bohemian manner.

His career came quickly to an end, and in February 1908, with light luggage and the manuscript of a book of poems that had been rejected by at least one American publisher, he set sail for Europe.

He had been to Europe twice before, once with his great aunt in 1898 and once again in the summer of 1906, when he had gathered the material for his first three published articles: "Raphaelite Latin," concerning the Latin poets of the Renaissance, and "Interesting French Publications," concerning the troubadours (both published in the *Book News Monthly*, Philadelphia, September 1906); and "Burgos, a Dream City of Old Castile" (October issue).

Now, with little money, he sailed to Gibraltar and southern Spain, then on to Venice, where in June 1908 he published, at his own expense, his first book of poems, *A lume spento*. About September 1908 he went to London, where he was befriended by the writer and editor Ford Madox Ford (who published him in his *English Review*), entered William Butler Yeats's circle, and joined the "school of images," a modern group presided over by the philosopher T.E. Hulme. In England, success came quick-

Success in London

ly to Pound. A book of poems, *Personae*, was published in April 1909; a second book, *Exultations*, followed in October; and a third book *The Spirit of Romance*, based on lectures delivered in London (1909–10), was published in 1910.

After a trip home—a last desperate and unsuccessful attempt to make a literary life for himself in Philadelphia or New York—he returned to Europe in February 1911, visiting Italy, Germany, and France. Toward the end of 1911 he met an English journalist, Alfred R. Orage, editor of the Socialist weekly *New Age*, who opened its pages to him and provided him with a small but regular income during the next nine years. In 1912 Pound became London correspondent for the small magazine *Poetry* (Chicago); he did much to enhance the magazine's importance and was soon a dominant figure in Anglo-American verse. He was among the first to recognize and review the poetry of Robert Frost and D.H. Lawrence and to praise the sculpture of the modernists Jacob Epstein and Henri Gaudier-Brzeska. As leader of the Imagist movement of 1912–14, successor of the "school of images," he drew up the first Imagist manifesto with its emphasis on direct and sparse language and precise images in poetry and edited the first Imagist anthology, *Des Imagistes* (1914). Though his friend Yeats had already become famous, Pound succeeded in persuading him to adopt a new leaner style. In 1914, the year of his marriage to Dorothy Shakespear, daughter of Yeats's friend, Olivia Shakespear, he began a collaboration with the then unknown James Joyce. As unofficial editor of *The Egoist* (London) and later as London editor of *The Little Review* (New York), he saw to the publication of Joyce's *Portrait of the Artist as a Young Man* and *Ulysses*, thus spreading Joyce's name and securing financial assistance for him. In the same year he gave T.S. Eliot a similar start in his career as poet and critic.

Editor of *Des Imagistes*

In the years following he continued to publish his own poetry (*Ripostes*, 1912; *Lustra*, 1916) and prose criticism (*Pavannes and Divisions*, 1918). From the literary remains of the great Orientalist Ernest Fenollosa, which had been presented to Pound in 1913, he published highly acclaimed English versions of early Chinese poetry, *Cathay* (1915), and two volumes of Japanese Nō plays (1916–17).

Unsettled by the slaughter of World War I and the spirit of hopelessness he felt was pervading England after its conclusion, Pound decided to move to Paris, publishing before he left two of his most important poetical works, "Homage to Sextus Propertius," in the book *Quia Pauper Amavi* (1919), and *Hugh Selwyn Mauberley* (1920). "Propertius" is a comment on the British Empire in 1917, by way of Propertius and the Roman Empire; *Mauberley*, a finely chiselled "portrait" of one aspect of British literary culture in 1919, is one of the best and most praised poems of the 20th century.

Meeting with Hemingway

In Paris (1921–24) Pound met and helped the young American novelist Ernest Hemingway; wrote an opera, *Le Testament*, based on poems of François Villon (performed in Paris in 1926 and by the BBC, London, in 1931 and 1962); assisted T.S. Eliot with the editing of his long poem *The Waste Land*; and acted as correspondent for the New York literary journal *The Dial*. In 1924 he tired of Paris and moved to Rapallo, Italy, which was to be his home for the next 20 years. In 1925 he had a daughter, Maria, by the expatriate American violinist Olga Rudge, and in 1926 his wife, Dorothy, gave birth to a son, Omar. The daughter was brought up by a peasant woman in the Italian Tirol, the son by relatives in England. In 1927–28 Pound edited his own magazine, *Exile*, and in 1930 brought together, under the title *A Draft of XXX Cantos*, various segments of his ambitious long poem "The Cantos," which he had begun in 1915.

The 1930s saw the publication of further volumes of "The Cantos" (*Eleven New Cantos*, 1934; *The Fifth Decad of Cantos*, 1937; *Cantos LII-LXXI*, 1940) and a collection of some of his best prose (*Make It New*, 1934). A growing interest in music caused him to arrange a long series of concerts in Rapallo during the 1930s; and, with the assistance of Olga Rudge, he played a large part in the rediscovery of the 17th-century Italian composer Antonio Vivaldi. The results of his continuing investigation in the areas of culture and history were published in his brilliant but fragmentary prose work *Guide to Kulchur* (1938).

Following the worldwide depression of the 1930s, he turned more and more to history, especially economic history, a subject in which he had been interested since his meeting in London in 1918 with C.H. Douglas, the founder of Social Credit, an economic theory stating that maldistribution of wealth due to insufficient purchasing power is the cause of economic depressions. Pound had come to believe that a misunderstanding of money and banking by governments and the public, as well as the manipulation of money by international bankers, had led the world into a long series of wars. He became obsessed with monetary reform (*ABC of Economics*, 1933; *Social Credit*, 1935; *What Is Money For?*, 1939), involved himself in politics, and declared his admiration for the Italian dictator Benito Mussolini (*Jefferson and/or Mussolini*, 1935). The obsession affected his "Cantos," which even earlier had shown evidence of becoming an uncontrolled series of personal and historical episodes.

As war in Europe drew near, Pound returned home (1939) in the hope that he could help keep the peace between Italy and the United States. He went back to Italy a disappointed man, and between 1941 and 1943, after Italy and the U.S. were at war, he made several hundred broadcasts over Rome Radio on subjects ranging from James Joyce to the control of money and the U.S. government by Jewish bankers and often openly condemned the U.S. war effort. He was arrested by U.S. forces in 1945 and spent six months in a prison camp for army criminals near Pisa. Despite harsh conditions there, he translated Confucius into English (*The Great Digest & Unwobbling Pivot*, 1951) and wrote *The Pisan Cantos* (1948), the best and most moving section of his long poem. Returned to the U.S. to face trial for treason, he was pronounced "insane and mentally unfit for trial" by a panel of doctors and spent 12 years (1946–58) in St. Elizabeth's Hospital for the criminally insane in Washington, D.C. During this time he continued to write and publish "The Cantos" (*Section: Rock-Drill*, 1955; *Thrones*, 1959), translated ancient Chinese poetry (*The Classic Anthology*, 1954) and Sophocles' *Trachiniai* (*Women of Trachis*, 1956), received visitors regularly, and kept up a voluminous and worldwide correspondence. Controversy surrounding him burst out anew when, in 1949, he was awarded the important Bollingen Prize for his *Pisan Cantos*. When on April 18, 1958, he was declared unfit to stand trial and the charges against him were dropped, he was released from St. Elizabeth's. He returned to Italy, dividing the year between Rapallo and Venice. For a while he wrote and worked but soon lapsed into silence; he died in Venice on November 1, 1972.

Anti-American broadcasts

Out of his 60 years of publishing activity came 70 books of his own, contributions to some 70 others, and more than 1,500 articles.

MAJOR WORKS

POEMS: *A lume spento* (1908); *Personae* (1909); *Exultations* (1909); *Provença* (1910); *Canzoni* (1911); *Ripostes* (1912); *Lustra* (1916); *Quia Pauper Amavi* (1919); *Hugh Selwyn Mauberley* (1920); *Umbra* (1920), collected early poems; *Poems 1918–1921* (1921); *A Draft of XVI Cantos* (1925); *Personae (The Collected Poems)* (1926); *A Draft of the Cantos 17–27* (1928); *A Draft of XXX Cantos* (1930); *Eleven New Cantos* (1934); *The Fifth Decad of Cantos* (1937); *Cantos LII-LXXI* (1940); *The Pisan Cantos* (1948); *The Cantos* (collected edition, 1948); *Section: Rock-Drill*, cantos 85–95 (1955); *Thrones, 96–109 de los cantares* (1959); *Redondillas, or Something of That Sort* (1967), unpublished long poem of 1911; *The Cantos* (1970).

PROSE: *The Spirit of Romance* (1910); *Gaudier-Brzeska* (1916); *Pavannes and Divisions* (1918); *Instigations* (1920); *Indiscretions* (1923); *Antheil and the Treatise on Harmony* (1924); *Imaginary Letters* (1930); *How to Read* (1931); *ABC of Economics* (1933); *ABC of Reading* (1934); *Make It New* (1934); *Social Credit: An Impact* (1935); *Jefferson and/or Mussolini* (1935); *Polite Essays* (1937); *Guide to Kulchur* (1938); *What Is Money For?* (1939); *"If*

This be Treason . . ." (1948); *Literary Essays* (1954); *Pavannes and Divagations* (1958); *Impact* (1960), economic, political, and cultural essays.

TRANSLATIONS: *The Sonnets and Ballate of Guido Cavalcanti* (1912); *Cathay* (1915); *Certain Noble Plays of Japan* (1916); *"Noh" or Accomplishment* (1917, dated 1916); *Dialogues of Fontenelle* (1917); *The Natural Philosophy of Love*, from Remy de Gourmont's *Physique de l'amour* (1922); *Ta Hio: The Great Learning* (1928); *Guido Cavalcanti Rime* (1932); *Confucius: Digest of the Analects* (1937); *The Classical Anthology Defined by Confucius* (1954); *Sophokles Women of Trachis* (1956); *Moscardino*, first volume of Enrico Pea's four-volume novel, *Il romanzo di Moscardino* (1956); *Love Poems of Ancient Egypt*, with Noel Stock (1962).

ITALIAN PROSE: (translated into English by John Drummond)—*Carta da Visita* (1942; *A Visiting Card*, 1952); *L'America, Roosevelt e le cause della guerra presente* (1944; *America, Roosevelt and the Causes of the Present War*, 1951), *Oro e lavoro* (1944; *Gold and Work*, 1952); also, trans. by Carmine Amore, "Introduzione alla natura economica degli S.U.A." (1944; Money Pamphlets by £, No. 1, "An Introduction to the Economic Nature of the United States," 1950).

BIBLIOGRAPHY. A complete listing of Pound's works may be found in DONALD GALLUP, *A Bibliography of Ezra Pound* (1963, revised reissue 1969), thorough and accurate; and MARIE HENAULT, *Checklist of Ezra Pound* (1970). Most of the writing on which his fame now rests may be found in *Personae (The Collected Poems)* (1926, new ed. 1949), a selection of poems Pound wished to keep in print in 1926, plus a few earlier and later poems added 1949; *The Cantos* (1970), cantos 1–117, a collection of all the segments published to date; *The Spirit of Romance* (1910, new ed. 1953; paperback with new prefatory note by Pound, 1968); *Literary Essays* (1954), the bulk of his best criticism, ed. with introduction by T.S. ELIOT; *Guide to Kulchur* (1938, new ed. 1952); *The Letters of Ezra Pound, 1907–1941*, ed. by D.D. PAIGE (1950), an excellent introduction to Pound's literary life and inimitable epistolary style.

Biographical studies: NOEL STOCK, *The Life of Ezra Pound* (1970); CHARLES NORMAN, *The Case of Ezra Pound* (1968); *Pound/Joyce: The Letters of Ezra Pound to James Joyce*, ed. with commentary by FORREST READ (1967); JULIEN CORNELL, *The Trial of Ezra Pound* (1966); HARRY MEACHAM, *The Caged Panther: Ezra Pound at Saint Elizabeths* (1967); MARY DE RACHEWILTZ, *Discretions* (1971).

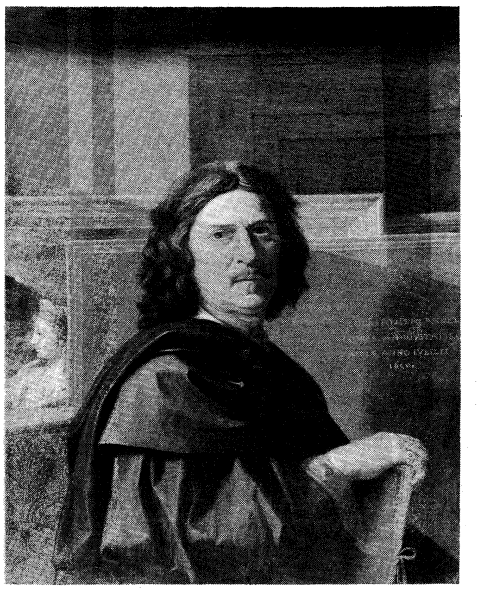
Criticism and interpretation: DONALD DAVIE, *Ezra Pound: Poet as Sculptor* (1964); EARLE DAVIS, *Vision Fugitive: Ezra Pound and Economics* (1968); N. CHRISTOPH DE NAGY, *The Poetry of Ezra Pound: The Pre-Imagist Stage*, 2nd rev. ed. (1968); NOEL STOCK, *Poet in Exile: Ezra Pound* (1964); HUGH KENNER, *The Poetry of Ezra Pound* (1951) and *The Pound Era* (1972); EVA HESSE (ed.), *New Approaches to Ezra Pound: A Co-ordinated Investigation of Pound's Poetry and Ideas* (1969); J.P. SULLIVAN (ed.), *Ezra Pound: A Critical Anthology* (1970). *On the Cantos:* NOEL STOCK, *Reading the Cantos* (1967); J.H. EDWARDS and W. VASSE, *Annotated Index to the Cantos of Ezra Pound* (1957).

(N.St.)

Poussin, Nicolas

Nicolas Poussin has been regarded by artists and connoisseurs in France as the greatest French painter of the 17th century, and perhaps in the history of French art, though he in fact lived and worked in Rome—except for the year and a half he spent at the Royal Court of France (1640–42)—from his 30th year to his death. His celebrity began during his lifetime—even the great sculptor Gian Lorenzo Bernini, who was the leading Baroque artist in Rome, was one of his fervent admirers—and his fame and influence were kept alive by such artists as Jacques-Louis David during the 18th century and in the 19th century by Eugène Delacroix, Jean-Auguste-Dominique Ingres, and Paul Cézanne, who proclaimed his own art an attempt to "do Poussin over again after nature."

Childhood and early travels. Poussin was born in 1594 in the hamlet of Villers near the small town of Les Andelys, on the River Seine, north of Paris. He was the son of very poor parents who earned a scanty living as small farmers, although his father once had been a soldier. Poussin spoke of the good education he received at Les Andelys, yet it seems that he was not even instructed in a trade or craft until his 16th year. He apparently did not show any interest in the arts until the painter Quentin



Poussin, self-portrait, oil painting, 1650. In the Louvre, Paris.
Giraudon

Varin (1570–1634) visited Les Andelys in 1612 to produce several paintings for the church of Le Grand Andely. Poussin's interest in the arts was awakened, and he decided to become a painter. As this was impossible in Les Andelys, he left his home, going first to Rouen and then to Paris to find a suitable teacher. His poverty and ignorance made this search very difficult. He found no master who appeared satisfactory and studied at different times under several minor painters who were either Mannerists in the Dutch tradition or portraitists. In another respect Poussin was more successful. A young nobleman from the French province of Poitou asked Poussin to come to his château and paint several pictures for him. But the nobleman's mother tried to use Poussin as an ordinary servant, so the young artist secretly left the estate to return to Paris. Travelling on foot and without any means, Poussin endured, according to his own statement, the greatest hardship and fatigue. He did not reach Paris but had to return to his paternal home, where he arrived not only very ill but much humiliated.

Early influences and training

Recovering after a year, Poussin again set out for Paris, not only to continue his studies but also to pursue another aim. While previously resident in the French capital, he had met a chamberlain of the Queen Mother, Marie de Médicis, an amateur of high rank who introduced him to the art of the Italian High Renaissance through reproductions of Raphael's paintings. These engravings, according to his biographer Passeri, inspired him with "that light he had ever longed for." He yearned, therefore, to go to Rome, which was then the centre of the European art world. Only after three attempts was he successful in reaching the Papal City. His first attempt to reach Rome met with disaster in Florence, the second at Lyon where he was stopped by police because of debts he had left unpaid in Paris. His third and final attempt to get to Rome was aided by cavalier Giambattista Marino, the Italian court poet to Marie de Médicis.

Journey to Rome

Marino, then a sick elderly man, took a liking to the young Poussin and ordered him to make a series of mythological drawings illustrating the *Metamorphoses* by the ancient Roman poet Ovid. He discussed with him the theoretical fundamentals common to poetry and painting, as they were seen at that time, and tried to pave his way to Rome, recommending him to Cardinal Francesco Barberini, a nephew of the reigning pope, Urban VIII. Marino had also intended to introduce Poussin personally to several Roman connoisseurs, as he was himself about to go to Rome (1623). But his plan failed; he fell ill and had to leave for Naples before Poussin arrived from Paris. The painter was left to rely on his own abilities and again underwent a time of severe privation.

Before Poussin's journey to Rome, most art historians are inclined to insert a visit that he is supposed to have paid to Venice in 1624. This journey is refuted by other scholars for three reasons. First, it is not considered logical that a man of limited means, journeying from Paris to Rome, would, after his arrival in Italy, first travel east rather than going directly south. Second, the principal source for this supposed Venetian journey is a note of questionable reliability written by Giulio Mancini, physician of Pope Urban VIII. Third, the paintings Poussin produced immediately after his arrival in Rome, such as the battle scenes (Leningrad, Moscow, and Rome), show no influence of Venetian colourism but imitate compositions of ancient Roman sarcophagi. Yet the so-called Venetian style of some of the pictures Poussin painted in Rome during the following years was probably inspired by three relatively early pictures painted by the Venetian Titian, which then hung in the Villa Ludovisi and were greatly admired by many Roman artists: "The Feast of the Gods," "The Bacchanal," and "Bacchus and Ariadne." This evidence was first cited by the Italian art theorist and biographer Giovanni Pietro Bellori (1615?–96), who knew Poussin personally.

First Roman period. Between Poussin's arrival in Rome in 1624 and his departure for France in 1640 he came to know many of Rome's most influential people. Perhaps the most important of his new acquaintances was the commendatore Cassiano dal Pozzo, secretary to Cardinal Barberini, who had a decisive influence upon his art. Pozzo was a well-known antiquarian who owned a rich collection of antique artifacts. According to Poussin's own testimony, these antiquities taught him most of what he came to know about classical art. Through Pozzo, who became Poussin's patron, the French painter became a fervent admirer of ancient Roman civilization. Voluntarily exiled from his native country, he developed an unreasonable hatred of French culture and an artistic style that, by its thoroughly Roman character, markedly contrasted with that of the other French artists then living in Rome or France.

Poussin continued the vagrant life of a bachelor, often changing his Roman address for unknown reasons. Toward the end of the 1620s he contracted a serious illness (according to Passeri, syphilis) that forced him to change his life-style. Suffering intense pain and socially abandoned, he was finally looked after by the family of a French confectioner and his Italian wife. Possibly impelled by a sense of gratitude after his recovery in 1630 or perhaps for mere convenience, he married their daughter, Anne-Marie Dughet, who was 18 years younger than himself. With Anne-Marie he lived in a childless marriage until she, passionately lamented by the painter, died a few years before his own death. He then began to lead a modest, silent, almost hidden petit bourgeois life, in a small house to which hardly anybody was admitted.

Marriage and retirement from artistic life of Rome

Although his art was much appreciated and he was known as one of the great painters in Rome, Poussin retired from active participation in artistic circles, as well as from the French community. His retirement from the artistic life of Rome did not result from fear nor from doubts about his artistic talent; on the contrary, it was based on an enormous self-confidence. While other artists followed the general movement of the Roman Baroque with the intention of attaining a certain significance in the general trend of the day, Poussin endeavoured to overcome the fashionable style on a higher level, inspired by Roman pre-Christian architecture and Latin books on moral conduct as well as by the nobility and greatness of Raphael's works, which, as he believed, had renewed the spirit of antiquity.

Painter to Louis XIII. Between 1638 and 1639, his achievements in Rome attracted the attention of the French court. Louis XIII's powerful minister Cardinal Richelieu tried to persuade Poussin to return to his homeland and become a member of the Académie Royal de Peinture et de Sculpture that he was trying to establish. In spite of a long preliminary correspondence, including an official invitation and a personal letter written by the King, Poussin showed no intention of following the summons. For a year and a half he made all sorts of excuses, knowing only too well how little the atmosphere of the French court would be compatible with the life he was used to and with the art he produced. But he finally yielded to the continued intense pressuring of Richelieu, Louis XIII, and Fréart de Chambray, a connoisseur of high rank and a collector, who was sent to Rome to conduct the negotiations. On October 28, 1640, Poussin sailed for France with several assistants. He left his wife in Rome, a fact that shows how much he distrusted his future appointment as a *peintre du roi*, "Painter to the king."

Received with great honours, Poussin nevertheless soon found himself in trouble with the ministers of the King as well as with the French artists, whom he met with the utmost arrogance. He was offered commissions for kinds of work he was not used to nor really qualified to execute: altarpieces, book covers, designs for tapestry and for the decoration of the Grande Galérie of the Louvre palace. What he produced did not elicit the praise he expected, so he left Paris in September 1642, not triumphantly as he had arrived but defeated, alone, in the *Coche de Lyon* (*i.e.*, ordinary stagecoach), and perhaps more humiliated than ever before. Unfortunately he did not live to see his personal style glorified by the French Academy in the late 17th century.

Second Roman period. During Poussin's second Roman period his personal appearance and behaviour changed. How he wished to be seen is evident from his two self-portraits, painted 1649 (Staatliche Museen Prussischer Kulturbesitz, Berlin) and 1650 (Louvre, Paris): a noble man with an assertive air that stemmed from his self-esteem. He could be met taking walks in the gardens of the Pincio followed by a number of admirers to whom he uttered his ideas concerning art and life, which increasingly became a kind of pagan philosophy. It is only during this period of his life that he really assumed the identity of the "painter-philosopher" for which he is generally known. During this last creative period Poussin concentrated his efforts on highly sophisticated themes of classical mythology as well as on biblical subjects.

Appearance and behaviour during second Roman period

Growing rather pessimistic with respect to human existence in general, he followed only from a great distance the events of Roman life and the political difficulties of his native land. His health declined, but, the more his body decayed, the more his spiritual and artistic faculties rose. From 1660 onward Poussin's complaints about his declining health became more numerous, and continuous pain interrupted his activities. Early in 1665 he ceased to paint. He died on November 19 and was buried in San Lorenzo in Lucina, his Roman parish church.

MAJOR WORKS

"The Triumph of David" (*c.* 1626; Dulwich College Picture Gallery, London); "The Assumption of the Virgin" (*c.* 1626; National Gallery of Art, Washington, D.C.); "The Martyrdom of St. Erasmus" (1628–29; Vatican Museum, Rome); "The Plague at Ashdod" (1630–31; Louvre, Paris); "The Adoration of the Magi" (1633; Gemäldegalerie, Dresden); "The Rape of the Sabine Women" (*c.* 1635; Louvre, Paris); "The Triumph of Neptune" (*c.* 1635; Philadelphia Museum of Art, Philadelphia); "The Adoration of the Shepherds" (*c.* 1637; National Gallery, London); "The Rape of the Sabine Women" (*c.* 1637; Metropolitan Museum of Art, New York); "The Seven Sacraments" (first series; 1642; Belvoir Castle near Grantham, Leicestershire, England; also one panel, National Gallery, Washington, D.C.); "The Seven Sacraments" (second series, 1644–48; The Duke of Sutherland Collection, on loan to the National Gallery of Scotland, Edinburgh); "Holy Family on the Steps" (1648; National Gallery of Art, Washington, D.C.); "Landscape with Diogenes" (1648; Louvre, Paris); "Self-Portrait" (1650; Louvre, Paris); "Christ and the Woman Taken in Adultery" (1653; Louvre, Paris); "The Exposing of Moses" (1654; Ashmolean Museum, Oxford); "The Arcadian Shepherds" (*c.* 1655; Louvre, Paris); "The Four Seasons" (1660–64; Louvre, Paris); Apollo and Daphne" (1664; Louvre, Paris).

BIBLIOGRAPHY. SIR ANTHONY BLUNT, *Nicolas Poussin*, 3 vol. (1966–67), is the most comprehensive account of Poussin's life and art. KURT BADT, *Die Kunst des Nicolas Poussin*, 2 vol. (1969), is an interpretation of the originality of Poussin's artistic achievements.

(K.Ba.)

Poverty

Poverty is the condition that is said to exist when people lack the means to satisfy their basic needs. In this context, the identification of poor people, then, first requires a determination of what constitutes basic needs. These may be defined as narrowly as "those necessary for survival" or as broadly as "those reflecting the prevailing standard of living in the community." The first criterion would cover only those people near the borderline of starvation or death from exposure; the second would extend to people whose nutrition, housing, and clothing, though adequate to preserve life, do not measure up to those of the population as a whole.

Poverty in industrial and underdeveloped nations

The immediately apparent differences between the living standards of the poor of the past or the poor of today's underdeveloped nations on the one hand, and those of the poor of modern, industrial societies, who may actually own automobiles, television sets, and washing machines, on the other, have led some authorities to suggest that it is unreasonable to call the latter group "poor." Others point out in response that people behave in accord with community standards, expectations, and aspirations in addition to absolute physiological needs. The problem of definition is further compounded by the noneconomic connotations that the word poverty has acquired. Poverty has been associated, for example, with poor health, low levels of education or skills, an inability or an unwillingness to participate in society, high rates of disruptive or disorderly behaviour, and improvidence. While all of these attributes have often been found to exist with poverty, their inclusion in a definition of poverty would tend to obscure the relation between them and the inability to provide for one's basic needs.

Whatever definition one uses and whatever correlates one adds to one's definition, authorities and laymen alike commonly assume that the effects of poverty are harmful both to individuals and to society. The assumption has considerable justification: in the United States in the 1950s an infant boy had a life expectancy of 66 years; in the underdeveloped nations the expectation was much lower—India, 42; Guatemala, 44; Colombia, 44; Guinea, 26. A similar differential could also be observed within countries. In the United States, for example, the infant mortality rate for urban blacks was roughly twice as high as that for whites.

TYPES OF POVERTY

Poverty can be classified according to a number of dimensions: time or duration (long- or short-term or cyclical), distribution (widespread, concentrated, individual), attitudes toward the poor (sympathetic or hostile), and presumed causes. The latter two dimensions are particularly important because of their relation to the kinds of efforts that are used in the attempt to ameliorate poverty.

Cyclical poverty. Cyclical poverty refers to poverty that may be widespread throughout a population, but the occurrence itself is of limited duration. In nonindustrial societies (present and past) this sort of inability to provide for one's basic needs rests upon temporary food shortages caused by natural phenomena or poor agricultural planning.

In industrialized societies the level of life is related to fluctuations in the ability of business, industry, and government to sustain high levels of employment at adequate wages. This ability, in turn, is dependent on, among other factors, fiscal policy, international trade balances, and, in capitalist countries, the optimism or pessimism of investors. Throughout the 19th and early 20th centuries the industrialized nations of the world experienced business panics and recessions. The United States' experience in the Great Depression of the 1930s, though unique in some of its features, exemplifies this kind of poverty. The Depression was rooted in the war debts of U.S. allies, reparations payments by Germany, and the tariff policies following World War I, as well as in runaway stock market speculation. Its effects, like those of prior depressions, included high unemployment rates, lowered wholesale prices, and reduced national production and income. Although business depressions affect all segments of society, their impact is most severe on people of the lowest strata, because of their marginal resources.

Effects of panics and depressions

For many years, poverty resulting from business fluctuations was accepted as an inevitable consequence of a natural process of market regulation. Relief (either in the form of temporary public works projects or provision of food) was granted reluctantly to tide people over until the business cycle again entered an upswing. As the causes of business depressions came to be more clearly understood, relief and public works were replaced in importance by programs intended to avert business depressions or to hasten recovery from them through manipulation of fiscal policy, regulation of the stock market, and international trade agreements. The general assumption today is that these measures minimize poverty by maintaining a healthy economy.

Collective poverty. In contrast to cyclical poverty, which is temporary, "collective" poverty involves a relatively permanent insufficiency of means to secure basic needs—a condition that may be so general as to describe the average level of life in a society, or that may be concentrated in relatively large groups in an otherwise prosperous society. Both generalized and concentrated collective poverty may be transmitted from generation to generation, fathers passing their poverty on to their sons.

Generalized collective poverty. Poverty is relatively general and lasting in much of Asia, the Middle East, most of Africa, and large parts of South and Central America. Life for the bulk of the population in these societies is at a minimal level. Death from starvation is not unusual. Nutritional deficiencies cause disease seldom seen by doctors in the West. Low life expectancy, high levels of maternal and infant mortality, and poor health characterize life in these societies. The magnitude of the problem is indicated by the probability that some 1,500,000,000 of the world's population are malnourished, in contrast to 450,000,000 who are well fed.

This type of poverty is usually related to economic underdevelopment. The total resources of many of these nations would be insufficient to support the population adequately, even if they were equally divided among all of the citizens. Proposed remedies are twofold: expansion of the gross national product through improved agriculture or industrialization, or both, and population limitation. Thus far, both population control and induced economic development have proved difficult and disappointing in their results.

Problems in developing countries

Increase of the gross national product does not necessarily lead to an improved standard of living for the population at large—for a number of reasons. (1) The most apparent of these is that, in the short run, as the food supply increases, the population generally increases at a higher rate. (2) Most of the underdeveloped nations are also characterized by a long-standing system of highly unequal rewards—a system likely to continue despite marked increases in the gross national product. Some authorities suggest that for some time to come, a larger portion of any increase will be siphoned off by persons who are already wealthy or in comfortable circumstances than will trickle down to the larger part of the population living at the near-starvation level. This, for instance, has often been the fate of aid from advanced nations. (3) The problem of financing agricultural and industrial improvements is one that cannot be met through the resources of the underdeveloped nations themselves but is dependent on investments from outside. It can be anticipated that a sizable portion of any increase in productivity will be paid as interest on these investments, again reducing the surplus available to the masses of poor in these countries. Industrializing nations are also especially vulnerable to the effects of poverty due to crop failures created by droughts, floods, or other natural afflictions. Their development plans have often been aborted because they have been forced to take funds earmarked for investment and use them instead for the purchase of food from other nations.

Concentrated collective poverty. In many industrialized, relatively affluent countries, particular demographic groups are vulnerable to long-term poverty. In city ghet-

tos, in regions bypassed or abandoned by industry, and in areas where agriculture or industry is inefficient and cannot compete profitably, there are found victims of concentrated collective poverty. These people, like those afflicted with generalized poverty, have high mortality rates, poor health, and so forth, when compared with the more affluent members of society. Although concentrated poverty is discussed in greater detail later in this article, briefly it may be said here that efforts at amelioration focus on ways to bring the deprived groups into the mainstream of economic life by attracting new industry, introducing improved agricultural methods, and raising the level of skills of the employable members.

Ghetto poverty

Case poverty. Similar to collective poverty in relative permanence but different from it in terms of distribution, case poverty refers to the inability of an individual or family to secure basic needs even in social surroundings of general prosperity. This inability is generally related to the lack of some basic attribute that would permit the individual to maintain himself. How severe the physical or mental handicap or how deficient the education and skill must be to make one unemployable depends on the range of jobs available in the society. Physical strength and endurance are crucial in countries depending more on labour than on capital, whereas in industrialized societies skill, education, and social adaptability have become more important. Those who have handicaps in social adaptability have long been associated with improvidence, a label covering such behaviour as the inability to manage money, laziness, drunkenness, and producing too many children.

Physical and mental handicaps are usually regarded sympathetically, as being beyond the control of the people who suffer from them. Efforts to ameliorate poverty stemming from physical causes focus on education, sheltered employment, and, if needed, economic maintenance. The *socially* handicapped, as can be judged from the language used to identify them, are stigmatized. Their handicaps are felt, in many instances, to be within the control of the individual, if he chooses to control them, and his poverty, then, is regarded as the outcome of the individual's own failures. The person is often dealt with punitively, given financial aid only reluctantly, and pressed to accept casework services to help him change his ways. Two attitudes toward the poor—sympathy and hostility—have permeated the literature on poverty.

TRADITIONAL ATTITUDES TOWARD POVERTY

Early efforts to cope with poverty up to the latter part of the 19th century had a common undercurrent: relief should involve at most only a *minimal* provision for basic needs, reflecting the idea that poverty offers an incentive to better one's self and that a guarantee of the means of survival removes that incentive. In Scotland the Rev. Thomas Chalmers expressed the belief of many when he declared in 1823 that the "four fountains of charity" were "frozen or locked up by the hands of legislation." These fountains were the frugality of the poor themselves, the kindness of relatives, the sympathy of the wealthy for the poor, and the sympathy of the poor for one another.

Laissez-faire and Malthusian views

Laissez-faire theory added economic justifications to the attitudes of those who favoured only minimal assistance. It proclaimed that if progress were to occur, economic life should be regulated only by the natural market processes of supply and demand. Adam Smith pointed out in 1776 that the propertied classes must ultimately pay wages high enough to assure the subsistence and reproduction of the labour force. Failure to do this would lead to the downfall of the system. This process, however, would eliminate only utter destitution, not poverty. Although Smith was not unsympathetic to the plight of the destitute, he was pessimistic about the results of efforts to improve their condition through relief. Smith assumed a finite wage supply and argued that any amount removed from this supply to relieve destitution without paying for work could only reduce the amount available to those who did work, making their subsistence more difficult. In his view interference with the natural process would lead to eventual intensification of the problem of poverty.

Thomas Malthus' *Essay on the Principle of Population*, published in 1798, furthered this general attitude. Malthus contended that since any increase of a population's food supply is more than consumed by the consequent increase in population, the only antidote was the limitation of population by war, disease, and famine. Without these controls he felt that the economic growth of his time would never be reflected in improved living conditions for a large part of the population. Later economists took up Malthusian ideas and declared flatly that attempts to interfere with "natural" processes by helping the poor could only lead to perpetuating destitution by maintaining them in too large numbers.

Existence at or below the subsistence level on the part of substantial numbers of people was accepted as the necessary price for economic growth and progress. As Joseph Townsend, a clergyman, had written in his *Dissertation on the Poor Laws* (1786):

> It seems to be a law of nature that the poor should be to a certain degree improvident, that there may always be some to fulfill the most servile, the most sordid, and the most ignoble offices in the community. The stock of human happiness is . . . much increased [when] the more delicate are not only relieved from drudgery . . . but are left at liberty . . . to pursue those callings which are suited to their various dispositions.

Townsend had forecast the union of laissez-faire economic theory and some theologies.

Much of late-18th- and early-19th-century Protestant theology was dominated by the concept of predestination. Material success and the accumulation of wealth were indications that one was favoured in God's eyes. Personal attributes associated with material success, such as thrift and sobriety, were valued highly, and the failure to succeed was seen as the "will of God" that ought not to be interfered with. When it was accompanied by improvidence and drunkenness, surely the victims of poverty deserved the consequences of their folly. To protect them from the consequences of their behaviour could only encourage profligacy.

Thus economics and religion combined to create a fear of the effects of alleviating poverty. Yet the repercussions of failure to alleviate poverty were also severe: famine, disease, theft, riots, and intense suffering. The common resolution of this quandary was to take relief measures sufficient to remove only the worst and most visible of the risks.

Social Darwinism

The justifications for the minimal relief of poverty were reinforced in the late 19th century by the advent of social Darwinism, an application of the principles of biological evolution to the social sphere. Biological evolution had occurred through the unregulated competition of individual organisms and species for the scarce necessities of life. Unfit individuals and unadaptable species died or became extinct, leading to vigorous species and the eventual emergence of man. In just such a manner, reasoned the English sociologist Herbert Spencer, who coined the phrase "survival of the fittest," unregulated competition among men had led to social and economic progress and the rise of the most able to positions of power and eminence. For progress to continue, unregulated competition must continue. Those at the bottom of the social ladder were there because they were inferior or unfit to engage in social life. To aid them would retard their extinction, hence retard progress itself.

Changing attitudes in the late 19th and early 20th centuries. Something, then, had to happen to break down the biases against the poor so that public provision for them and direct financial assistance in their own homes (outdoor relief, in contrast to workhouses and almshouses) became more acceptable. Religious and economic assumptions were gradually changing. The late 19th century saw an accelerating rejection of religious dogma regarding predestination in favour of a theology that emphasized the perfectability of man. This resulted in increased expression of a view that the poor need not be relegated to the poorhouse—that both they and their situations were capable of improvement.

Furthermore, almost simultaneously with the advent of

Progressive liberalism

social Darwinism, there came those who questioned not only its conclusions but the assumptions of laissez-faire economics as well. Critics asked whether the upper classes occupied their favoured status as a result of their own fitness in the struggle for survival or through their power to regulate competition in their own interest. Those who accepted the second alternative began to propose means through which the basis for competition by the poor could be improved while reducing what they saw as the unfair advantage of the rich. This did not represent a total rejection of the laissez-faire model as an ideal. Rather it was an attempt to make it function more adequately. The changes in attitude and their conversion into practice reflect the changing experiences of people in a rapidly industrializing society. Increased exposure of the conditions and consequences of poverty and a fear of the consequences played a part in the adoption of new programs.

Exposure of poverty problems. A number of investigations of public welfare practices were undertaken in the mid-19th century, when even though society in most industrialized countries was generally prosperous, expenditures for welfare had risen sharply. Suspicions of mismanagement and charges that able-bodied people were being cared for were common. Thus, most investigations were geared to reducing public expense. Contrary to expectations, these studies revealed that almshouses and relief rolls included the mentally and physically ill, the aged, and women with children, with a very low proportion of employable persons. The almshouses were particularly reprehensible, for they were dirty and crowded, and no special treatment was afforded the insane. By the end of the 19th century there were several successful efforts to set up specialized institutions for the indigent insane, ill, and children.

The "friendly visitors"

Public outdoor relief received no such spur as a result of these investigations. Relief given to recipients in their own homes, though cheaper than almshouses, was viewed suspiciously. The policy makers of private agencies stressed that continued dependence on financial support from any source was demoralizing to the recipients and damaging to society. Therefore, the help that they were prepared to give was preventive and, in the early stages particularly, geared to removing defects of character. (Only in cases in which urgent need forestalled other efforts at improvement did they offer temporary financial assistance.) Associations designed to improve the condition of the poor utilized the services of volunteer "friendly visitors," who attempted to help the poor lift themselves up from poverty through education, kindly example, and moral suasion.

Private agencies restricted their services to those believed able to improve their own condition. Despite the reiterated policy "that it is worse to debase by alms than to withhold them; that the physical must be subservient to the moral, and the moral receive the attention its paramount importance demands" (quoted in R.E. and M.W. Pumphrey, *The Heritage of American Social Work;* Columbia University Press, 1961, p. 110), the friendly visitors were often faced with situations in which they gave priority to financial assistance. The idea of the vicious circle of poverty was born in these confrontations with the poor. Hardworking people whose income was insufficient to meet their needs, the involuntarily unemployed whose wages had been so low that they couldn't "save for a rainy day," workers felled by illness or industrial accidents, unmarried women, widows, and children whose limited employment opportunities at low wages gave little hope that they could ever become self-supporting: none of these situations could be alleviated by moral suasion. And, given such a start, the children of the poor were doomed to repeat the lives of their parents. Even in situations not so clearly outside the control of the poor, should children be penalized for their parents' failures? The voluntary agencies could not help any of these individuals. They did not deserve the label "paupers." Yet all of them were worthy of help beyond the minimal provisions of public relief. Though in practice, charity organizations often provided the necessities of life for their clients, their resources, derived from contributions, were totally inadequate to meet the need they found. It became apparent to many of the visitors, if not to their supervisors, that new means for relieving and preventing poverty were necessary.

Settlement work

Settlement houses, such as Toynbee Hall in England, Hull House in Chicago, and the Henry Street Settlement in New York, were founded to educate the poor and to provide places for social gatherings. College-educated workers from comfortable homes lived in the settlement houses located in the slums. They found that, in addition to their original purposes, their new-found knowledge—as Jane Addams, founder of Hull House, said—made these workers see the needs of their neighbourhoods as a whole and furnish data for, and use their influence to secure, legislation.

In addition to the personal experiences of the friendly visitors, settlement house workers, and journalists, other sources of information became available from 1890 to the beginning of World War I. These were the scientific social surveys starting in England with Charles Booth's *Life and Labour of the People in London* and B. Seebohm Rowntree's *Poverty: A Study of Town Life.* As Booth remarked, "a large portion of the population are ill-nourished to the point of being inefficient, but that for very many of these it is impossible that it should be otherwise; the remuneration of unskilled labour, the cost of food, and the needs of the human body being what they are."

Although, admittedly, the sentiments and findings of friendly visitors, settlement workers, journalists, and social surveyors were recorded throughout the 19th century, it is in the period 1890–1910 that they intensified and began to affect social policy. In terms of the classification of types of poverty mentioned earlier, the conception of concentrated poverty, with its presumed causes and effects, had begun to replace that of case poverty.

Discontent and protest. The change of attitudes about the poor was a significant factor in bringing about reform. Reform, however, does not necessarily follow a change of attitude alone. Another set of circumstances made it seem expedient, as well as moral, to deal with the poor in a different way. Although the period up to the outbreak of World War I in Western countries was one of tremendous economic growth and prosperity, there were large parts of the population who did not share in the growth and who were cruelly affected by the periodic business panics. Farmers vigorously protested low prices. Factory labourers working 10 to 12 hours a day for wages barely sufficient to support themselves expressed similar dissatisfaction. In times of financial depression, discontent grew as thousands of employable persons found themselves facing unemployment or reduced wages.

Radical reform movements

Discontent was expressed in many ways: third parties and movements proliferated, efforts to unionize intensified, and looting and rioting occurred. Marxist and anarchist thought provided much of the rhetoric, stressing the conflict of class interests and the eventual triumph of the workers over the capitalists. Some of the means advocated for furthering the interest of the workers alarmed the more prosperous. An anarchist, Johann Most, from Germany, wrote in 1885, "Dynamite! Of all the good stuff, that is the stuff! . . . place this in the immediate vicinity of a lot of rich loafers who live by the sweat of other people's brows, and light the fuse. A most cheerful and gratifying result will follow." (Quoted in S.H. Holbrook, *Dreamers of the American Dream;* Doubleday & Co., Inc., 1957, p. 295). Isolated acts of violence during labour disputes or political rallies, regardless of their provocation or doubt about responsibility, seemed to confirm the danger from radicals and the poor.

Measures for the relief of poverty in the late 19th and early 20th centuries. Effective change was slow in coming. Often long periods elapsed between the first proposal and discussion of a measure and its actual implementation. The first programs to be adopted bore either an indirect relation to poverty or were conceived of as preventing poverty. They did not put money into the hands of the poor. Compulsory education, better housing, improved public health facilities, and the regulation of wa-

Preventive, insurance, and direct-relief plans

ges and hours of women and children appeared to be measures that would break up the vicious circle of poverty. Advocates of public health and tenement housing regulation were relatively successful in securing early legislation. Unfortunately, the tenement laws did not alleviate the poverty-perpetuating crowding and bad sanitation, because of poor enforcement and the connection between low rent and poor housing. Child labour laws were slow to appear. Minimum-wage and minimum-hour legislation for women also proceeded slowly.

Unemployment insurance, workmen's compensation, and health insurance were planned to give money to the poor while avoiding the implication of pauperism. Proponents of workmen's compensation stressed that it would avoid prolonged litigation to establish liability and would be a means of avoiding industrial accidents, as well as providing money for injured workers. Benefit scales for all programs were initially very low for the wage earner and his dependents.

A substantial group of widows and children were not covered by insurance programs geared to previous employment, and some form of direct, government-financed assistance was required to avoid destitution and the separation of family members. Mothers' allotments were proposed to provide an adequate income. These programs were accompanied by supervision and regulation of the families' lives, however, in order to discourage dependence on public money. Allowances were small in the expectation that families would supplement them through earnings. Even so, laws permitting the allotments could not assure that local governments would implement the programs. Agitation for direct financial aid to the indigent aged began in the late 19th and early 20th centuries, leading to the German and British systems of compulsory contributory insurance. In the United States, federal programs of unemployment insurance and old-age assistance were introduced in the 1930s.

None of the plans—preventive, insurance, or direct assistance—gave sufficient money to meet the needs of the poor, and in many instances no money actually became available to them despite permissive legislation. Mothers' allotments and, to a lesser extent, old-age pensions were not granted with an eye to long-term maintenance, and the emphasis of the efforts was on making the recipients self-supporting. Nonetheless, significant changes had occurred. The undifferentiated almshouse had been rejected, issues regarding the legitimacy of various forms of aid to the poor and their financing had been settled, and a number of programs were in operation in most Western countries by the beginning of World War II. For further description of such programs, see WELFARE AND SECURITY PROGRAMS.

INDUSTRIAL POVERTY IN THE LATTER HALF OF THE 20TH CENTURY

Rediscovery of poverty amid affluence. To say that something has been "rediscovered" is to suggest that the phenomenon was virtually dormant, overlooked, neglected, or lost sight of for a period of time. This was the situation with respect to attention to poverty in affluent nations from the end of World War II to the late 1950s and beyond. Some observers contend that social scientists, social workers, and various other professionals retained a high level of interest in poverty, but that it was masked by a new frame of reference that did not use the labels "poor" and "poverty." Rather they spoke in terms of "lower class," "low-income group," and "multiproblem family." When viewed more closely, however, the situation clearly involved more than a particular set of terms falling into disrepute and disuse. No matter what the label, the interest of social scientists, professionals, and laymen in poverty dropped off markedly after the Great Depression of the 1930s and remained at a lower level for nearly two decades. Throughout most of the 1950s, social research and popular literature were concerned with such topics as the changing mores and mass culture. The lure of these subjects was enhanced by what is probably the most important single reason for the deflection of interest in the poor: most social thinkers shared with the general public the commonly held belief that the postwar prosperity of industrial nations was eliminating large-scale poverty. The few references to poverty in sociological literature, for example, portrayed the poor as comprising a small percent of population. As will be seen, this conception had little basis in reality. Nonprofessionals in the United States, most notably Michael Harrington in *The Other America* (1962), helped to shatter these illusions and redirect public attention to the paradox of extensive poverty coexisting with affluence.

The "new" poverty

After such a long period of neglect, poverty was rediscovered and action was deemed necessary because of a number of events. In the United States a set of social and economic disorders began to accelerate in the 1950s and increasingly fed into each other until their overall impact was inescapable. Among the factors involved, the most crucial were the rising costs of welfare, the belief that crime rates were soaring, and, in some countries, the sharp increase in unemployment.

Poor minorities, youth, and unskilled workers

The political consequences of the climbing rate of unemployment or partial employment were immediate and far-reaching because of the composition of the unemployed or underemployed group. Three categories posed particularly sensitive problems: minority groups, such as the nonwhites in the U.S. and Britain, whose rate was over that of the overall figure; the young, who were similarly overrepresented; and the semiskilled workers who had, until that time, been continuously employed. In the United States the increasingly large numbers of Negroes and young persons joining the ranks of the jobless represented potentially volatile forces. Equally important were the workers who had regarded themselves as having a secure claim in the job market. The "respectable working class," from which upwards of one-third of the unemployed were drawn, reacted strongly to its sudden impoverishment and began to exert political and social pressure. Through unions and informal networks of relationships, the message and implications of their discontent and anxieties got through to politicians and community leaders. Their addition to the already high welfare rolls made it impossible to continue to perceive the poor as a residual group of the old, disabled, and otherwise unemployable. A similar pattern can be detected in the rediscovery of poverty in other affluent nations, although the timetable has been somewhat different.

Life of the poor and its effects. Descriptions of the way of life and behaviour of the poor in modern industrial society are remarkably similar to depictions of the poor in industrial societies of the past. A vignette of the poor in Charles Booth's London could as well be that of a black family in a Harlem hovel, an Algerian family in the slum outskirts of Paris, an Italian family in a Naples slum, a family in the wastelands of Appalachia, or a poor Japanese family crowded in the hovels of Tokyo. The general picture that emerges, despite variations in cultural setting, is that of severe sociocultural and personality impoverishment together with pervasive family and personal disorganization. Numerous studies among the poor have documented the high incidence of family instability, emotional illness, and various types of deviant behaviour such as crime and delinquency. Indifference, apathy, fatalism, resignation, and related traits are terms typically used to describe the outlook of the poor and are at the core of what is often referred to as the "culture of poverty." Family life may be fraught with fear and violence and exhaustion and bitterness. The children typically grow up emotionally and physically stunted—a situation conducive to repeating the cycle of their parents' lives.

Anxieties of the poor

One need not seek far for an explanation of these characteristics of the poor. The chronic insufficiency of basic needs—poor nutrition, bad housing, and inadequate medical care—and the anxiety that accompanies a precarious existence are the roots of the common deterioration of personal and social behaviour of those living in poverty.

This contemporary poverty amid affluence belongs to the concentrated collective type, in that various demographic groupings in the population (*e.g.*, the aged, the unskilled, the incapacitated, the minority groups, large families) have a high incidence of relatively permanent

poverty. A common approach to explaining the sources of such poverty is to list a set of factors and characteristics associated with the poor and to suggest that the cause of poverty is to be found in some combination of these factors. A report issued by the U.S. Chamber of Commerce, for instance, uses this approach in holding that "poverty is largely a product of ignorance, changing industrial technology, prejudices that limit opportunity, worn out farms and depressed communities, inability or unwillingness to move where jobs are, lack of motivation, curable physical or mental abilities" (*The Concept of Poverty*, 1965). There is nothing untrue about this statement: all the "causes" cited do play some role in poverty. Many observers would contend, however, that this is a useless, even misleading position; that it does not isolate the salient variables; that it mixes personal attributes such as ignorance, unwillingness to move, and lack of motivation with attributes of the social structure such as changing technology, worn-out farms, and depressed communities; and that it does not say whether individual and structural attributes are to be treated equally as sources of poverty or, if not, how each should be weighed.

Any approach focussing on the poor's attitudinal characteristics (*e.g.*, motivation) or even demographic characteristics (*e.g.*, occupational status) as a *point of departure* for assessing the causes of poverty can lead to serious errors. In particular, a line of reasoning that moves from individual psychological attributes to group attributes and then to the complex phenomena of socioeconomic poverty confuses cause and effect. To be sure, lack of adequate job skills and education carry high risks for poverty, and families resigned to remaining in a depressed area and possessed of too many children are also likely candidates for poverty. Such characteristics, however, are as much the consequences of poverty as they are causes, if not more so.

Social structure of the poor

The more plausible approach may be in seeing that the social structure is the ultimate source of the bulk of poverty. This orientation allows one to encompass structural factors (for instance, technological change) that most authorities agree are implicated in poverty as well as to account for the ambiguous role of most of the personal and group attributes usually dealt with.

The key feature of the social structure pertinent to a discussion of the causes of poverty is the economic institution, with an emphasis on the relations between unemployment, employment and poverty, and the increasing impact of automation in changing the nature of these relations. (Other relevant factors in a capitalist society are its monetary and fiscal policies.) Historically, poverty and unemployment have been closely connected, but they are no longer. Admittedly, changes in the unemployment rate do still figure importantly in changes in the number of the poor today. For example, an unemployment rise of 1 percent in a labour force of over 80,000,000 could mean the addition of 1,000,000 or more (including dependents) to the ranks of the poor, with pronounced effects upon certain disadvantaged groups. But large-scale unemployment is not a prominent feature of contemporary poverty. In most industrial countries the unemployment rate in recent times has remained below 4 to 7 percent. More to the point, most of today's poor families have at least one employed person, and in the majority of instances the wage earner is the family head. The employment situation is not too different for unattached individuals under 65: the great majority of persons in this group have jobs.

Thus, although there is no denying that unemployment is important in considering such poverty groups as the aged, the incapacitated, the illiterate, the minorities, and the female-headed family, it is more important to direct attention to the bulk of the poor who are employed. Overwhelmingly, the jobs that poor persons hold in industrial society require minimal skills and bring in marginal wages. The low income from such employment, together with the high incidence of "short weeks" and periodic brief layoffs, largely explains how a situation of nominally full employment can coexist with widespread poverty.

The changed relation between unemployment and poverty seems due largely to the progressive shrinkage of job opportunities in the lower ranges of the occupational structure. Among other things, automation may be steadily reducing the need for workers doing repetitive tasks of the type required in mass production. The long-range ramifications of automation are uncertain, but for the foreseeable future it appears that the supply of workers with limited skills and education is increasingly outstripping the demand for them, with a consequent lack of pressure for a livable wage scale.

Coping with the problem of poverty. From early times, direct financial aid to the poor has been regarded as a temporary, but necessary, evil. Only the aged and severely disabled have been assured of uninterrupted, though often inadequate, income. It has been assumed that if one is of working age, reasonably healthy, and willing to work, he can locate a job that will provide for his needs. As suggested above, in recent years there has been greater awareness that this assumption is not justified. Consequently, plans to cope with poverty are increasingly geared to meeting the needs of the employed poor and those for whom there are no jobs.

Guaranteed income and job-training plans

This is the basic rationale underlying a variety of income maintenance plans falling into the category of "guaranteed annual income." Under such programs a minimum income figure is set, based on family size, and the difference between this figure and actual income is made up by a government subsidy. The problem centres obviously on determining a proper figure. If it remains too low, there will be scarcely an elimination of the welfare and relief costs. Setting the minimum appreciably higher, on the other hand, raises other problems. As the income of the unemployed, underemployed, or lowly paid poor begins to approach that of employed persons just above the minimum level, it creates a growing potential for resentment of the system. Furthermore, many observers contend that the incentive to work is reduced in both groups. Other critics argue that a guaranteed income to people who are primarily employed in jobs demanding low skills will remove pressure on employers to pay higher wages for these jobs. This would have the added effect of depressing the earnings of those who are not quite eligible for subsidies.

A number of programs, either proposed or operative in some countries, focus on semiskilled and unskilled workers, trying to upgrade their skills through education or retraining. Such relatively short-term job-training programs, however, have tended to be highly selective and to eliminate those very persons most in need of training. When programs have indeed included these persons, considerable time has had to be spent on such basic skills as reading, writing, and arithmetic, which alone do not improve employability. In short, job training as a means of eliminating poverty has been most successful only with those workers who have not required massive upgrading of skills.

Another program that has attracted many supporters is the children's allotment, which would pay a specified amount for each child to all families regardless of need. Such a program exists in such countries as France, Sweden, and the U.S.S.R. Although such a scheme neglects the childless poor, it is argued that money thus spent has the greatest impact on future poverty—improving, as it does, the future of the present generation of youth. Some critics of the plan fear that it would encourage an increase in the birthrate in a world that is already overpopulated, although this has not been the case in any Western nation granting children's allotments.

Expansion of welfare programs

Overall, there has been in most Western countries a considerable expansion of all varieties of programs aimed at equalizing incomes and reducing poverty. Sweden is a notable example. It has one of the world's most highly developed systems of labour-management mediation and conciliation, controls over incomes and production, and welfare programs. One-third of the nation's annual budget goes to welfare, financed by relatively high progressive taxes that have appreciably narrowed the gap between "rich" and "poor".

BIBLIOGRAPHY. Any account of the history of English poor laws is heavily indebted to SIDNEY and BEATRICE WEBB, *English Poor Law History*, 2 vol. (1927–29). A detailed account of the treatment of the poor in medieval times may be found in B. TIERNEY, *Medieval Poor Law: A Sketch of Canonical Theory and Its Application in England* (1959). The debate surrounding Speenhamland and the Poor Law Reform of 1834 is central to J.R. POYNTER, *Society and Pauperism* (1969). B.D. COLL, *Perspectives in Public Welfare* (1969), offers a good, brief history of changing forms of public welfare from medieval to modern times. M. ORSHANSKY describes the most commonly used standard for determining the extent and distribution of poverty in "How Poverty is Measured," *Monthly Labor Review*, 92:37–41 (1969). Information on the extent, distribution, and perpetuation of poverty in the U.S. may be found in A.B. BATCHELDER, *The Economics of Poverty* (1966); L. FISHMAN (ed.), *Poverty Amid Affluence* (1966); H. LUMER, *Poverty: Its Roots and Its Future* (1965); and H.H. MEISSNER (ed.), *Poverty in the Affluent Society* (1966). Proposals for new means of alleviating poverty, such as children's allotments and the guaranteed annual income, are discussed in A.L. SCHORR, *Explorations in Social Policy* (1969); and in J.C. VADAKIN, *Children, Poverty and Family Allowances* (1968). Problems in reducing poverty in the nonindustrial nations are dealt with in G. MYRDAL, *Asian Drama: An Inquiry into the Poverty of Nations* (1968); G.A. BORGSTROM, *The Hungry World: The Modern World at the Edge of Famine* (1967); and in G.M. MEIER, *Studies in International Poverty*, 2nd ed. (1970). "The Dimensions of World Poverty," *Scient. Am.*, 219:27–35 (1968), by D. SIMPSON, deals with the measurement, extent, and severity of world poverty and also comments on the prospects for its alleviation; as does D. HOROWITZ in *The Abolition of Poverty* (1969). J. BHAGWATI, *The Economics of Under-developed Countries* (1966), sets forth an internation framework for the relief of poverty and discusses the incidence of squalor, the economic and social structures of underdeveloped countries, and the types of domestic programs that have been undertaken. T. BALOGH, *The Economics of Poverty* (1966), concentrates on the problems, policies, and proposals for relief of poverty in the newly independent nations.

(J.R./J.K.R.)

Pragmatism

In its broadest and most familiar sense, "pragmatism" refers to the usefulness, workability, and practicality of ideas, policies, and proposals as criteria of their merit and claims to attention. Achieving results, "getting things done" in business and public affairs is often said to be "pragmatic." There is a harsher and more brutal connotation of the term in which any exercise of power in the successful pursuit of practical and specific objectives is called "pragmatic." The character of American business and politics is often so described. In these cases "pragmatic" carries the stamp of justification: a policy is justified pragmatically if it is successful. The familiar and the academic conceptions have in common an opposition to invoking the authority of precedents or of abstract and ultimate principles. Thus in law, judicial decisions that have turned on the weighing of consequences and probable general welfare rather than on being deduced from precedents have been called pragmatic.

The word pragmatism goes back to the Greek πραγμα ("action," "affair"). The Greek historian Polybius (died 118 BC) called his writings "pragmatic," meaning thereby that they were intended to be instructive and useful to his readers. In his introduction to the *Philosophy of History*, Hegel commented on this "pragmatical" approach as the second kind of reflective history, and for this genre he cited Johannes von Müller's *History of the World* (Eng. trans. 1840). As the psychologist and leading Pragmatist William James remarked, "the term is derived from the same Greek word πραγμα meaning action, from which our words 'practice' and 'practical' come." Charles Peirce, another pioneering Pragmatist, who may have been the first to use the word to designate a specific philosophic doctrine, had Kant's German term rather than the Greek word in mind: *Pragmatisch* refers to experimental, empirical, and purposive thought "based on and applying to experience." In the philosophy of education the notion that children learn by doing, that critical standards of procedure and understanding emerge from the application of concepts to directly experienced subject matters, has been called "pragmatic." In semiotics, the general theory of language, that part that studies the relation of the user to the words or other signs that he uses is called pragmatics (as distinct from semantics and syntax).

MAJOR THESES OF PHILOSOPHIC PRAGMATISM

During the first quarter of the 20th century, Pragmatism was the most influential philosophy in America, exerting an impact on the study of law, education, political and social theory, art, and religion. Six fundamental theses of this philosophy can be distinguished. It is, however, unlikely that any one thinker would have subscribed to them all; and even on points of agreement, varying interpretations mark the thought and temper of the major Pragmatists. The six theses are:

1. Responsive to Idealism and evolutionary theory, Pragmatists have emphasized the "plastic" nature of reality and the practical function of knowledge as an instrument for adapting to reality and controlling it. Existence is fundamentally concerned with action, which some Pragmatists exalted to an almost metaphysical level. Change being an inevitable condition of life, Pragmatists have called attention to the ways in which change can be directed for individual and social benefit. They have consequently been most critical of moral and metaphysical doctrines in which change and action are relegated to the "merely practical," on the lowest level of the hierarchy of values. Some Pragmatists anticipated the more concrete and life-centred philosophy of Existentialism by arguing that only in acting—confronted with obstacles, compelled to make choices, and concerned to give form to experience—is man's being realized and discovered.

2. Pragmatism is a continuation of critical Empiricism in emphasizing the priority of actual experience over fixed principles and a priori reasoning in critical investigation. For James this meant that the Pragmatist

> turns away from abstraction and insufficiency, from verbal solutions, from bad *a priori* reasons, from fixed principles, closed systems, and pretended absolutes and origins. He turns towards concreteness and adequacy, towards facts, towards action. . . . It means the open air and possibilities of nature, as against . . . dogma, artificiality, and the pretence of finality in truth.

3. The pragmatic meaning of an idea, belief, or proposition is said to reside in the distinct class of specific experimental or practical consequences that result from the use, application, or entertainment of the notion. As Peirce commented: "Our idea of anything is our idea of its sensible effects." Two propositions for which no different effects can be discerned have merely a verbal appearance of dissimilarity, and a proposition for which no definite theoretical or practical consequences can be determined is pragmatically meaningless. For Pragmatists "there is no distinction of meaning so fine as to consist in anything but a possible difference of practice." Meaning thus has a predictive component, and some Pragmatists came close to identifying the meaning of a term or proposition with the process of its verification.

4. While most philosophers have defined truth in terms of a belief's "coherence" within a pattern of other beliefs or as the "correspondence" between a proposition and an actual state of affairs, Pragmatism has, in contrast, generally held that truth, like meaning, is to be found in the process of verification. Thus truth *is* the verification of a proposition, or the successful working of an idea. Crudely, truth is "what works." Less crudely and more theoretically, truth is in Peirce's words, the "limit towards which endless investigation would tend to bring scientific belief." For John Dewey, founder of the "Instrumentalist" school of Pragmatism, these are beliefs "warranted" by inquiry.

5. In keeping with their understanding of meaning and truth, Pragmatists have interpreted ideas as instruments and plans of action. In contrast to the conception of ideas as images and copies of impressions or of external objects, Pragmatist theories have emphasized the functional character of ideas: ideas are suggestions and anticipations of possible conduct; they are hypotheses or forecasts of what will result from a given action; they are ways of organiz-

ing behaviour in the world rather than replicas of the world. Ideas are thus analogous in some respects to tools; they are efficient, useful, and valuable, or not, depending on the role that they play in contributing to the successful direction of behaviour.

6. In methodology, Pragmatism is a broad philosophical attitude toward the formation of concepts, hypotheses, and theories and their justification. To Pragmatists man's interpretations of reality are motivated and justified by considerations of efficacy and utility in serving his interests and needs; the molding of language and theorizing are likewise subject to the critical objective of maximum usefulness according to man's various purposes.

HISTORY OF PRAGMATISM

Idealism, Empiricism, Kantianism

Antecedents in modern philosophy. Pragmatism was a part of a general revolt against the overly intellectual, somewhat fastidious, and closed systems of Idealism in 19th century philosophy. These boldly speculative Idealists had expanded man's subjective experience of mind till it became a metaphysical principle of cosmic explanation. To the Idealist, all of reality was one fabric, woven from parts that cohered by virtue of the internal relations that they bore to one another; and this reality was often interpreted in abstract and fixed intellectual categories. The theory of evolution, then still new, seemed to the Pragmatists, on the other hand, to call for a new, non-Idealist interpretation of nature, life, and reason—one that challenged the long-established conceptions of fixed species. The new emphasis was on the particular variations and struggles of life in adapting to the environment. Philosophically, the fact of growth and the development of techniques for instituting changes favourable to life became the significant factors rather than the Idealist's ambitious rationalistic account of human goals and of the universe in general, and important developments in natural science and logic also encouraged a critical attitude toward earlier systems.

There were two main influences on the early formation of Pragmatism: One was the tradition of British Empiricism in the work of John Stuart Mill, Alexander Bain, and John Venn, which had stressed the role of experience in the genesis of knowledge—and particularly their analyses of belief as being intimately tied in with action and, indeed, as definable in terms of one's disposition and motive to act. The work of George Berkeley, an important 18th-century empirical Idealist, which presented a theory of the practical and inferential nature of knowledge, of sensations as signs (and thus predictive) of future experience, led Peirce to refer to him as "the introducer of Pragmatism." The other major influence came from modern German philosophy: from Kant's analysis of the purposive character of belief and of the roles of will and desire in forming belief and his doctrine of "regulative ideas," such as God or the Soul, which guide the understanding in achieving systematic completeness and unity of knowledge; from Romantic Idealists, for whom all reason is "practical" in expanding and enriching human experience; and from Hegel's historical and social conception of changing and developing subject matters. In sum, Peirce was profoundly impressed by Kant and by the Scottish philosophy of common sense, James by British Empiricism and by the voluntarisms (stressing the role of choice or will) of the genetic epistemologist James Ward and the relativistic French Personalist Charles Renouvier, Dewey by Coleridge's version of Kant's active conception of mind and by neo-Kantian and Hegelian Idealism.

Finally, to these influences must be added that of American social experience in the 19th century: the rapid expansion of industry and trade and a popular optimism, with its roots in Puritan theology, holding that hard work and virtue are bound to be rewarded. Both the precariousness of frontier life, however, and the rapidly expanding economy weakened the prevailing Calvinistic belief in a predestined future and encouraged the emergence of inventiveness, a sense of living still in the New World experiment, and adoption of the ideal of "making good."

The Metaphysical Club. It was in the critical group discussions of the "Metaphysical Club" in the 1870s in Cambridge, Massachusetts, that Pragmatism first received philosophic expression. In addition to Peirce and James, membership in the club included Chauncey Wright, F.E. Abbot, and Justice Oliver Wendell Holmes, Jr. A version of Peirce's now classic paper "The Fixation of Belief" (Nov. 1877) seems to have been presented at the club. But James also published a paper in 1878, "Spencer's Definition of Mind as Correspondence," in which his Pragmatism and analysis of thought and belief are clearly discernible. It was in a lecture delivered 20 years later, however, that James introduced Pragmatism, then fully crediting the idea to Peirce. It was primarily James's exposition that became famous and was received by the world at large.

Peirce's Pragmatism

The classical Pragmatists. The Pragmatic philosophy of Charles Sanders Peirce was part of a more general theory of thought and of signs. Thought, or "inquiry," it was held, results from doubt, a state in which habitual actions are blocked or confused and from which organic irritation and irresolution result. Resolution, unobstructed conduct, on the other hand, are products of belief, which is a form of stability and satisfaction. It is the function of scientific thought to produce true beliefs. In a prolonged effort to embed this analysis of doubt and inquiry within a more comprehensive theory of signs in which communication, thought, knowledge, and intelligent conduct could be fully understood, Peirce achieved a wealth of original insights. A sign is a socially standardized way by which something (a thought, word, object) refers man (the community of sign users) to something else (the interpretant), which, in turn, is itself another sign. Peirce's Pragmatism is thus a method for translating certain kinds of signs into clearer signs in order to surmount linguistic or conceptual confusion. Getting at the interpretant involves determining the "effects" or consequences of the signs or ideas in question.

Peirce's Pragmatism is therefore primarily a theory of meaning that emerged from his first-hand reflections on his own scientific work, in which the experimentalist understands a proposition as meaning that, if a prescribed experiment is performed, a stated experience will result. This method has two different uses:

(1) It is a way of showing that when disputes permit no resolution, the difficulties are due to misuses of language, to subtle conceptual confusions. Such questions as whether the physical world is an illusion, whether man's senses always mislead him, or whether his actions are fated are "not real problems."

(2) The method may be employed for clarification. As Peirce wrote:

> Consider what effects, that might conceiveably have practical bearings, we conceive the object of our conception to have. Then our conception of these effects is the whole of our conception of the object.

To say, for example, that an object *O* is hard means that if the operation of scratching *O* is performed, *O* will not be scratched by most substances. One thus achieves clarity when one can supply a conditional statement of this kind.

Similarly, in his theory of truth, one means by truth of belief that if a certain operation is the subject of continuous scientific inquiry by the community of investigators, assent to the belief would increase and dissent decrease "in the long run." Consequently, not only is thought purposive but meaning carries a reference to the future. Peirce's concept of the community of sign users and inquirers also has social and moral relevance, for it is nothing less than the ideal of rational democracy.

Witnessing his doctrine undergo a medly of dubious interpretations, Peirce eventually dissociated himself from these by calling his own view "Pragmaticism," a term he called so ugly as to be safe from uninformed use. Parts of the work of Dewey, of the social Pragmatist G.H. Mead, and of the conceptualistic Pragmatist C.I. Lewis are a further development of the logical Pragmatism of Peirce. The English logician F.P. Ramsey and the Italians Giovanni Vailati and Mario Calderoni also undertook significant extensions of Peirce's Pragmatism.

An alternative, though not wholly different, version of Pragmatism was developed by William James. It took a

psychological and moral form largely unforeseen and unintended by Peirce.

James's Pragmatism

A basic difference between Peirce and James is discernible in their respective conceptions of the direction to be taken by Pragmatic analysis. While Peirce construed meaning in general, conditional schema, and interpretants, James focussed upon the distinct contributions that ideas and beliefs make to specific forms of human experience on the living level of practical wants and purposes. Between the two close friends there persisted a fundamental philosophical difference in outlook that affected even their styles. While James was a Nominalist, holding that the full significance of ideas, meanings, and actions lies in their particular concrete existent occurrence, Peirce—as a scholastic Realist—sharply criticized him at this point, arguing that "a thing in general is as real as in the concrete."

The most conspicuous feature of James's writings on Pragmatism is the dominant place given to considerations of value, worth, and satisfaction—consequences of his teleological (purposive) conception of mind (*cf.* his *Principles of Psychology*). James maintained that thought is adaptive and purposive but also suffused with ideal emotional and practical interests—"should-be's"—which, as conditions of action, work to transform the world and create the future, even to "make the truth which they declare." Consequently, truth and meaning are species of value: "*The true is the name of whatever proves itself to be good in the way of belief.*"

James took meaning to be an intimate part of the use of ideas for expediting action. The notion of the difference that a proposition makes in experience was fundamental in James's Pragmatic methodology. He remarked that "it is a good rule in physiology, when we are studying the meaning of an organ," to look to the specific function that it performs. In like manner, the special difference that the presence of mind makes in observable cases, reflected in its unique functioning, defines the use of "mind": "In particular, the pursuance of future ends and the choice of means . . . are thus the *mark and criterion of the presence of mentality*."

With his training in medicine and psychology and the influence of Darwin in the background, James considered that the main function of thought is to help us establish "satisfactory relations with our surroundings." Thus man helps to mould the character of reality according to his needs and desires. Indeed, this is fundamental in James's defense of the right to believe in his famous essay "The Will to Believe" (1897). James argued that we may have a reasonable right to hold a religious or metaphysical belief (*e.g.*, that there is a perfect, eternal, and personal aspect of the universe) when the belief in question would supply a vital psychological and moral benefit to the believer, when evidence for and against the belief is equal, and when the decision to believe is forced and momentous. In James's functional conception of truth, the "working," and hence the truth, of ideas is their role in opening up valuable possible directions of thought and action—"a leading that is worth while."

James's "working" view of truth and of a reality that man in part makes by acting out and real-izing ideas, and especially his essay "The Will to Believe," were enthusiastically received by F.C.S. Schiller in England and by Giovanni Papini in Italy, and these doctrines became a cause célèbre for Pragmatists and their critics.

Schiller's Pragmatism

An admirer and friend of James, Schiller, now nearly forgotten, was once the most famous Pragmatist in England and Europe. Schiller was initially a humanist in the sense that, for him, both reality and knowledge are reflections of human activity—"the taken" rather than "the given." He first came to appreciate James's "The Will to Believe" in 1897 and subsequently acknowledged its impact on his thinking in an early and important paper, "Axioms As Postulates" (1902). He was a tireless critic of the "closed" systems of Idealism of F.H. Bradley, J.M.E. McTaggart, and Bernard Bosanquet and an advocate of the intellectual freedom that consists in "open," plural, changing, and to some extent never finished philosophical theorizing. According to Schiller, reality and truth are "man-made" rather than eternal verities. The true and the false are basically forms of good and bad and are relative to the private purposes of some particular person. He attempted to describe and analyze the "logic" of the experimental "trying" through which such needs are satisfied. For Schiller, reality is wholly plastic; and, starting out from initial postulates, one proceeds to construct his schemes for achieving a satisfactory outcome of desire, finally rendering his unformed possibilities (*hylē*) into a common world of language and action. On this view, all of science derives from and is inescapably guided by the psychological process of human thought; man is the measure of all things.

Dewey's Instrumentalism

John Dewey once noted that "Peirce wrote as a logician and James as a humanist." This distinction characterizes not only the course of Pragmatism but also the shaping of Dewey's own thought. Dewey first felt the influence of James in the 1890s while struggling to free himself from the hold of Hegelian Idealism. Later, he recognized the value of Peirce's work, which clearly prefigured certain ideas that he had developed independently.

With indefatigable effort and care Dewey reformulated Pragmatism, critically readjusting some of its conflicting doctrines, drawing upon his own work in psychology and education, and finding stimulation in the social Pragmatism of his friend George Herbert Mead. The resulting construction was Instrumentalism, which Dewey conceived as a single coherent theory embracing both the logical and humanistic currents of Pragmatism and thus integrating the methods and conclusions of scientific knowledge with beliefs about values and purposes.

While scientific, moral, and social experiences may differ in subject matter, the method of thought functioning "in the experimental determinations of future consequences" remains the same for all inquiry. Initially provoked by doubtful or problematic conditions, intelligent conduct is addressed to a resolution and settling of these conditions and to a "warranted assertion"—Dewey's version of "truth." Such is the "mediative function" of reason. "Truth" is thus identified with the outcome of competent inquiry. Actions occurring on the organic level, if they be at first confused and obstructed, can then become organized, coherent, and liberated through such inquiry.

Dewey's analysis of the organic, cultural, and formal conditions of intelligent action implies that all reflective conduct issues in an evaluation of a situation with respect to future action and consequences: thus inquiry is essentially an evaluative procedure. This method, most impressively applied in the sciences, is nonetheless a paradigm of moral activity as well. In ethics, "the action needed to satisfy" the situation is not to be found simply by the application of moral codes. The meaning

> has to be searched for [since] there are conflicting desires and alternative apparent goods. . . . Hence inquiry is exacted. . . . The good of the situation has to be discovered, projected and attained on the basis of the exact defect and trouble to be rectified.

In general, for Instrumentalism, moral ideals and "ends" function as means and hypotheses in guiding the deliberative process directed to controlling experience and attaining future goods.

> Not health as an end fixed once for all, but the needed improvement in health—a continual process—is the end and good. . . . Not perfection as a final goal, but the ever-enduring process of perfecting, maturing, refining is the aim of living . . . . Growth itself is the only moral 'end.'

Inquiry possessed a genuine religious significance for Dewey, and in its functioning as a critical, self-corrective social process of human growth he envisaged the working ethic of democracy.

Other American Pragmatists. Two important contributions to American Pragmatism, which have not yet received the attention that they deserve, came from Mead and Lewis.

Mead's orientation was social psychology. He had studied physiological psychology in Germany, had earlier worked under James and Josiah Royce at Harvard, and was also familiar with Peirce's analyses of thought and

signs. Dewey regarded him as one of the most fertile minds in American philosophy.

Mead's social and Lewis' conceptualistic Pragmatism

Mead developed the most comprehensive of the Pragmatist theories of mind. He depicted the evolution of mind and self-consciousness as emerging from social interactions and the use of gestures and "significant symbols" such as words. In contrast to other creatures, an individual regarded as having mind, engaging with others in social acts, can respond to his own gestures as others respond to them—thus taking on social roles and becoming an "other" in respect to himself. It is therefore by means of language, the use of "significant symbols," that mind emerges.

Fundamental to Mead's philosophy is his conception of the social act, in which individuals modify and direct one anothers' activities, work out their purposes, and accordingly transform their environments. In the social act the future controls present conduct, and this is distinctive of consciousness. Since the function of intelligence is to render the world "favourable for conduct," Mead viewed the development of scientific knowledge and the evolutionary process as coinciding.

Lewis's theory of "conceptualistic Pragmatism" was derived partly from his study of modern logic and partly from the influence of Royce and the classic Pragmatists. The critical results of a careful study of Kant are traceable in his work.

Lewis's Pragmatism focusses upon concepts, categories, and principles through which experience is interpreted. Though the sensuously given is "unalterable," how it is taken, how conceptually interpreted, depends on the purposes and initiatives of the mind—the a priori element in knowledge, which, functioning as categorical criteria of reality, is "true no matter what." It is by means of these that a systematic interpretation of reality is developed. However,

> there may be alternative conceptual systems, giving rise to alternative descriptions of experience, which are equally objective and equally valid. . . . When this is so, choice will be determined . . . on pragmatic grounds.

In stressing the purposive character of conceptualization, Lewis is thus in the main course of American Pragmatism.

Pragmatism in Europe. In his preface to *Pragmatism*, James commented that the Pragmatic movement was the focal expression of a number of philosophic tendencies suddenly becoming conscious of themselves and of "their combined mission." He mentioned the French thinkers Maurice Blondel, Édouard Le Roy, and B. de Sailly and the Italian, Giovanni Papini. Blondel was the author of *L'Action* (1893) and spokesman for a voluntaristic and activistic theory of knowledge. He was a founder of the "school of action," a liberal Catholic group that was part of the modernist movement (which employed the new historico-critical approach to the Bible and promoted a rationalistic interpretation of the faith). As early as 1888, Blondel appropriated the term *Pragmatisme*, only to abandon it when he learned of American Pragmatism, which was a more naturalistic philosophy than his own. Le Roy, closer to James than other French thinkers, also called his views Pragmatism. In broad respects he was like James in holding that the truth and the full significance of beliefs is found in acting them out. Le Roy was a disciple of Henri Poincaré, who had argued that scientific theories are not mere summaries of data, nor deduced from axioms, but are creative constructions, products of human thought and ingenuity, "conventions." To the question of what limits are imposed on otherwise arbitrary conventions, of what justifies them, Le Roy suggested their convenience in use. James saw similar forms of Pragmatism in Ernst Mach, Wilhelm Ostwald, Pierre Duhem, and Théodore Ruyssen in "the notion that no theory is absolutely a transcript of reality" and that "their great use is to summarize old facts and lead to new ones" so that they are a "man-made language, a conceptual shorthand . . . in which we write our reports of nature."

Blondel, Le Roy, Papini, and others

Another French thinker, Georges Sorel, undertook to reformulate James's Pragmatism into a "useful" doctrine of social criticism. Mussolini later cited Sorel and James as two of his philosophic mentors. He claimed to find in James "that faith in action, that ardent will to live and fight, to which Fascism owes a great part of its success." To the democratic James, no lesson could have been more badly learned.

A more immediate and direct form of James's Pragmatism occurred in Italy with its centre in the journal *Leonardo*, under the leadership of the iconoclastic critic Giovanni Papini. James referred to Papini as "a brilliant, humorous and witty writer." He called him a genius and was addressed in turn by him as "the Master." Papini's Pragmatism, derived from James's "The Will to Believe," became a theory of the will to action. In action, through creative power and passion, man achieves a kind of divinity. This romantic exaltation of action was appealing to artists but also to fanatics. Papini and his associate Giuseppe Prezzolini comprised the "magical" school of Pragmatism (in the sense of seeking "divinely-creative" power) in contrast to the "logical" school inspired by Peirce of G. Vailati and M. Calderoni.

Later tendencies. Certain extensions and applications of Pragmatism are to be found in current American philosophy. Sidney Hook has directed some of the critical techniques of analysis against a number of ideologies. In the tradition of the scientifically oriented Pragmatisms of Peirce and Dewey, he has explored the relation between the logic of experimental inquiry and the ethic of democracy. A converging of Pragmatism and Logical Positivism resulted in the movement of "logical empiricism" which, in addition to Dewey and Lewis, included the top-rank philosophers of science P.W. Bridgman, Rudolf Carnap, and Ernest Nagel, and the philosophical semanticist Charles W. Morris, all of whom were responsive to the Pragmatisms of Peirce, James, and Dewey. More recent and detailed studies of the structure of science, the nature of theories and explanation, by Carnap, Lewis, and Nagel, and a new interest in Instrumentalism deriving from the work of F.P. Ramsey, Ludwig Wittgenstein, W.H. Watson, and Gilbert Ryle (as well, of course, as from Dewey) exhibit a further continuation of Pragmatism.

Positivists and Nominalists

Maintaining a "more thorough Pragmatism" than that of Carnap and Lewis, who viewed choices made within scientific frameworks as pragmatic decisions, the prominent logician W.V. Quine has argued against the alleged boundary between analytic and synthetic truths. No portion of the conceptual scheme of science is exempt (as analytic truths were supposed to be) from possible revision in the flux of experience and in the light of pragmatic interests in efficacy and comprehension in predicting future experience. In important respects Quine's view of the evolution, organization, and function of the conceptual structure of science is close to that of Schiller, James, and Dewey.

Quine has also defended a methodological Pragmatism and relativism on ontological questions on the nature of being or reality. This position, also taken by another contemporary philosopher of science and language, Nelson Goodman, recalls the earlier Pragmatists' notion of the "plastic" character of reality, now seen as conceptually plastic in the sense of being expressible in a variety of systems of symbols and languages. It is conceptually misguided to seek *the* nature of objects, since what there *is* is not describable *in abstracto* from the particular language in which an ontological question has been put. Objects declared to be real, be they classes, numbers, atoms, or stones, may differ widely. But differences between the "theoretical" and the "factual" entities are basically differences of degree and purpose in evolving conceptualization. Hence reality is anything that it is truly said to be —in any of many linguistic and symbolic systems; and where differences in ways of speaking about objects call for choices, the ensuing adjudications will be pragmatic.

EVALUATION OF PRAGMATISM

Pragmatism has been vulnerable to certain criticisms. It has often been portrayed as a rationalization of the American business ethos—a portraiture perhaps inspired, but not by any scrutiny of the writings of the philosophers themselves. Similarly, the Pragmatic theory of truth has

been assailed. Concerning an idea or belief, James had held that one can say: " 'It is useful because it is true' or that 'it is true because it is useful.' " "Both phrases," he added, "mean the same thing." Most scholars, however, have denied this equivalence. His position may seem, moreover, to allow for an idea to be true (*i.e.*, useful or expedient) for one person and false (inexpedient) for others. Finally, James was accused of reducing truth to a subjective play of opinions that one happens to relish or find useful to believe. To these charges James replied that "what immediately feels most 'good' is not always most 'true' when measured by the verdict of the rest of experience." He also warned: "Woe to him whose beliefs play fast and loose with the order which realities follow in his experience."

As a single movement, Pragmatism is no longer extant; but as a body of ideas it contributes a heritage that is destined for future analysis and development. Chief among these are the interpretation of thought and meaning as forms of purposive behaviour, of knowledge as evaluative procedure in which normative and descriptive materials are integrally related, and of the logic of scientific inquiry as a norm of intelligent conduct in the affairs of men. Finally, Pragmatism has succeeded in its critical reaction to the 19th-century philosophy from which it emerged. It has influenced the current conception of philosophy as a critical method of investigating problems and clarifying communication rather than as a universal synthesis of knowledge. Pragmatism thus has certain affinities with the critical philosophizing of G.E. Moore and Bertrand Russell, as well as with the thought of the French intuitionist and vitalist Henri Bergson and his disciple Édouard Le Roy, of Blondel, of the early Positivists Mach and Duhem, of the fictionalist Hans Vaihinger, of the Vienna Circle and the philosopher of logic and language Ludwig Wittgenstein, and also of the founder of Phenomenology, Edmund Husserl, and some of the continuing forms of Phenomenology and Existentialism. It has recognized the relative, contingent, and fallible (yet still authentic) character of human reason, rather than perpetuating the dubious ideal of philosophy as a system of eternal truths. In so doing, and in thus altering the philosophical scene, Pragmatism has become vitally implicated in the practices of current intellectual life; and in the light of this fact, a more pragmatic justification of Pragmatism is difficult to imagine.

BIBLIOGRAPHY

General: H.S. THAYER, *Meaning and Action: A Critical History of Pragmatism* (1968), a study of the philosophers and the movement in America and abroad, with bibliography; "Pragmatism," in D.J. O'CONNOR (ed.), *A Critical History of Western Philosophy*, pp. 437–462 (1964); and H.S. THAYER (ed.), *Pragmatism: The Classic Writings* (1970), the basic writings in the Pragmatism of Peirce, James, Dewey, Mead, and Lewis, and further bibliographical references; John Dewey, "The Development of American Pragmatism," in *Philosophy and Civilization*, pp. 13–35 (1931); and CHARLES W. MORRIS, *The Pragmatic Movement in American Philosophy* (1970).

Special studies: P.P. WIENER, *Evolution and the Founders of Pragmatism* (1949), a basic work on the philosophical background and members of the Metaphysical Club; M.H. FISCH, "Was There a Metaphysical Club in Cambridge?" in E.C. MOORE and R.S. ROBIN (eds.), *Studies in the Philosophy of Charles Sanders Peirce*, 2nd series, pp. 3–32 (1964); A.J. AYER, *The Origins of Pragmatism* (1968); D. RUCKER, *The Chicago Pragmatists* (1969), on the "Chicago group," notably Dewey and Mead.

Collections of writings: A. RORTY (ed.), *Pragmatic Philosophy* (1966), a selection of earlier and contemporary sources, with an extensive bibliography.

Classic works: CHARLES SANDERS PEIRCE, "The Fixation of Belief," "How to Make Our Ideas Clear," and "What Pragmatism Is," in *Collected Papers*, vol. 5, ed. by C. HARTSHORNE and P. WEISS (1934); WILLIAM JAMES, *Principles of Psychology*, 2 vol. (1890), *The Will to Believe and Other Essays in Popular Philosophy* (1897), *Pragmatism: A New Name for Some Old Ways of Thinking* (1907), and *The Meaning of Truth* (1909); JOHN DEWEY, *How We Think* (1910), *The Influence of Darwin on Philosophy* (1910), *Democracy and Education* (1916), *Essays in Experimental Logic* (1916), *Reconstruction in Philosophy* (1920, 1948), *Human Nature and Conduct* (1922), *Logic: The Theory of Inquiry* (1938), *Theory of Valuation* (1939), and *Problems of Men* (1946).

On F.C.S. Schiller see R. ABEL, *The Pragmatic Humanism of F.C.S. Schiller* (1955), with a bibliography of Schiller's writings; on French and Italian pragmatists, H.S. THAYER, *Meaning and Action: A Critical History of Pragmatism*, part 3 (1968), with further bibliographical references.

(H.S.T.)

Prague

One of the finest cities of Europe, Prague (Czech Praha), the capital of Czechoslovakia and that nation's major economic and cultural centre, lies at the heart of the continent and was the home of some 1,173,000 people by the mid-1970s. Spread like Rome on seven hills, the city has a rich architectural heritage that reflects both the uncertain currents of history in Bohemia and an urban life extending back more than a thousand years. The physical attractions of the city are many. The winding course of the Vltava River, with its succession of bridges and changing vistas, contrasts with the ever-present backdrop of the great castle of Hradčany (or Prague Castle), which dominates the left-bank region of the city from behind massive walls set high on a hill. The narrow streets and little taverns and restaurants of the older quarters contrast with the broad sweep of Václavské náměstí (Wenceslas Square) and modern parks and housing developments, while the great 18th-century Baroque palaces have their own elegance and splendour. Seen from the surrounding hills, the many church towers make up a unique perspective, giving Prague its description as the "city of a hundred spires." This architectural harmony has been enhanced by post-1945 planning, which has preserved the ancient core of the city as one major monument and has carefully supervised all modern building.

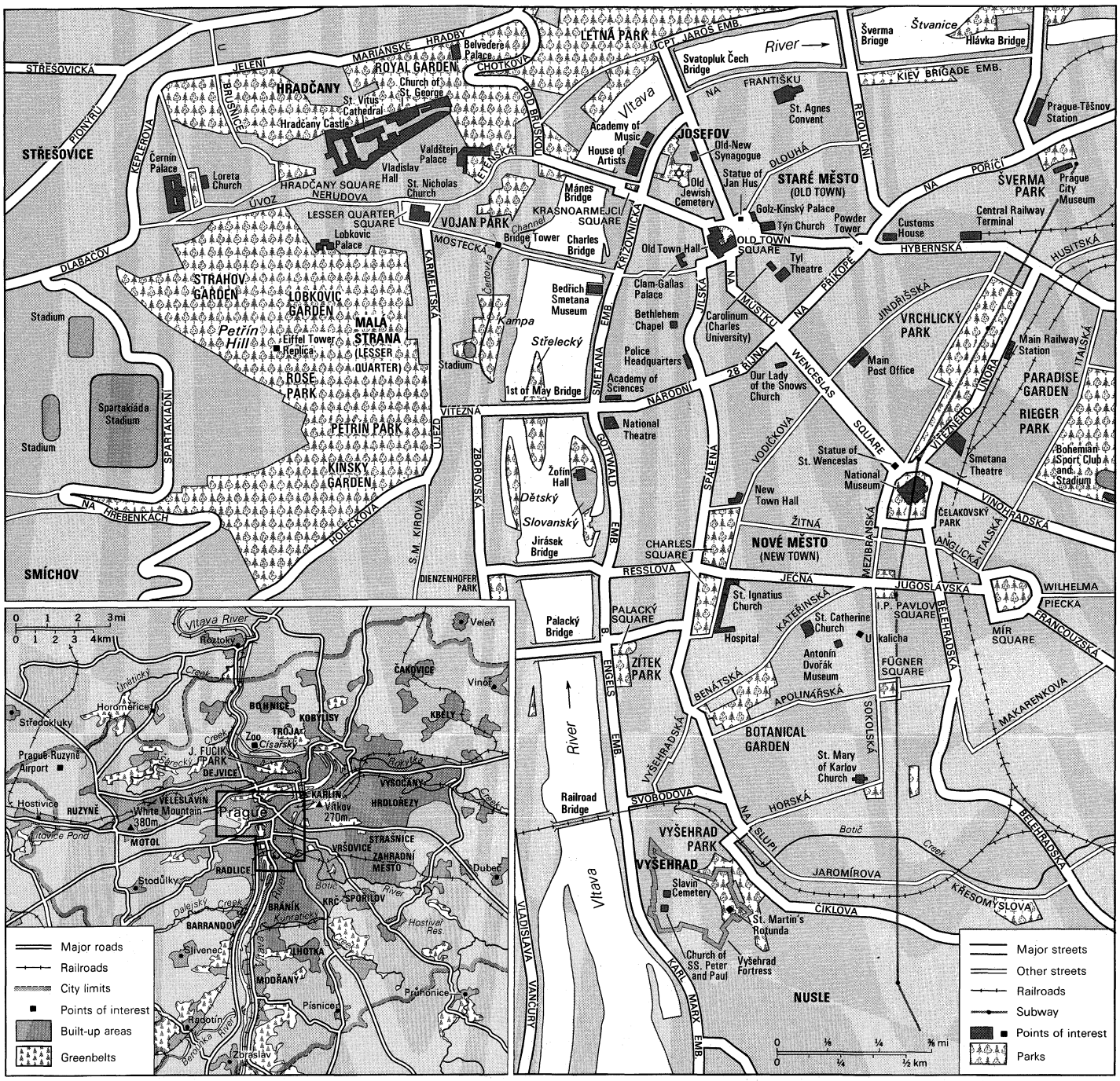
The character of the city

Prague is also famous for its cultural life: Wolfgang Amadeus Mozart lived there, and his *Prague Symphony* and *Don Giovanni* were first performed there; the lyric music of the great Czech composers Bedřich Smetana, Antonín Dvořák, and Leoš Janáček is commemorated each year in the spring music festival. In literature, the *U kalicha* ("At the Chalice") beer parlour, still popular with local residents and tourists alike, provided the setting for the humorously anti-authoritarian activities of Schweik, immortalized by the novelist Jaroslav Hašek in *The Good Soldier Schweik;* the writings of Franz Kafka, dwelling in a different way on the dilemmas and predicaments of modern man, also seem indissolubly linked with life in this city. For related information see BOHEMIA AND CZECHOSLOVAKIA, HISTORY OF; and CZECHOSLOVAKIA.

THE HISTORY OF PRAGUE

The foundation of the city. For thousands of years that portion of the Vltava's course where Prague was to rise was crossed by trade routes linking northern and southern Europe. The region is replete with Paleolithic relics, and Neolithic farmers inhabited the region from around 5000 to 2700 BC. Celts had settlements in the region from about 500 to 200 BC, including the fortified Závist, to the south of Prague. From the 4th century AD to the 6th, Slavs appeared on the Vltava banks, followed by the Avars.

The first settlement at what is now Prague has been traced to the second half of the 9th century. The oldest building was Vyšehrad Castle (*hrad*, "castle"), set on a commanding right-bank hill. It was soon followed by what was to become Hradčany Castle, set on an equally commanding left-bank site a little downstream. Legend (stirringly told in Smetana's opera *Libuše*) ascribes the foundation of Prague to a Princess Libuše and her husband Přemysl, a plowman and founder of the Přemyslid dynasty; legend notwithstanding, the Přemyslids, in power from *c.* 800 to 1306, consolidated a political base centred on Prague that was to be the nucleus of the Bohemian state and that enabled the natural trade advantages of the city site to develop under defensive protection. The dynasty included St. Wenceslas (Václav), who was murdered by his brother Boleslav *c.* 936 and whose statue now looks down upon the square to which his name was



Central Prague and (inset) its metropolitan area.

later given; and Boleslav I, whose reign (*c.* 936–967) witnessed the consolidation of power against a German threat. The little community began to flourish, and in the year 965 the Arab merchant and traveller Ibrāhīm ibn Yaʿqūb was able to describe it as a "busy trading centre." In 973 the bishopric of Prague was founded.

Medieval growth. The economic expansion of the community was soon reflected in the topography of the city. A market centre on the right bank, opposite Hradčany Castle, developed into the Staré Město (Old Town), particularly after the construction of the first stone bridge over the river in 1170. By 1230 the Staré Město had been given borough status and was defended by a system of walls and fortifications. On the opposite bank, under the walls of Hradčany, the community known as Malá Strana (Lesser Quarter; literally, Small Side) was founded in 1257. Following the eclipse of the Přemyslids, the House of Luxembourg came to power when John of Luxembourg, son of the future emperor Henry VII, became king of Bohemia. His son, Charles IV, Bohemian king and Holy Roman emperor, had his capital at Prague from 1346 to 1378 and took considerable personal interest in the development of the city. In 1348 he founded the Charles University (Universita Karlova), the first in central Europe, which was later to attract scholars and students from all over the Continent. His reign also saw the growth of the planned Nové Město (New Town) adjacent to the Staré Město; construction of the Karlův most (Charles Bridge; 1357, reconstructed in 1970) linking the Staré Město and the Malá Strana; and the beginning (1344) of the great St. Vitus' Cathedral, which was not completed until 1929. Other buildings included the Carolinum (the central hall of the university), the town hall (destroyed in 1945), and several churches and monasteries in Nové Město. The Jewish ghetto was also developed, and the bishopric was raised to an archbishopric in 1344.

The role of Charles IV in planning the city

By the 14th century Prague had become a major central

European city, with the Czech money minted at nearby Kutná Hora serving as the hard currency of the entire region. Foreign merchants, notably Germans and Italians, became economically and politically powerful in uneasy alliance with the kings. The social order, however, became less stable because of the emergent guilds of craftsmen, themselves often torn by conflicts between masters and journeymen. The town paupers added a further volatile element.

The Reformation and the Thirty Years' War. Prague played a significant role in the Reformation. The sermons of Jan Hus, a scholar at the university, begun in 1402 at the now-restored Bethlehem Chapel and carrying forward the criticisms of the church developed by the English reformer John Wycliffe, endeared him to the common people but brought him into conflict with Rome; he was burned at the stake in the town of Constance (Konstanz, West Germany) in 1415. Popular uprisings in 1419, led by the Prague priest Jan Želivský, included the throwing of city councillors from the windows of the New Town Hall in the incident known as the first Defenestration of Prague. The next year Hussite peasant rebels, led by the great military leader Jan Žižka, joined forces with the Hussites of Prague to win a decisive victory over the Catholic king (later emperor) Sigismund at nearby Vítkov Hill.

During the next 200 years, the rich merchants became ascendant once more, and the late Gothic architectural style flourished in many churches and buildings, reaching a peak in the fine Vladislavský sál (Vladislav Hall) of Hradčany Castle. In 1526, however, the Catholic Habsburgs became rulers of Bohemia and attempted to crush Czech Protestantism. The second Defenestration of Prague (1618), when the governors of Bohemia were thrown from the windows of the council room in Hradčany Castle—one of the major events precipitating the Thirty Years' War—was followed by the decisive defeat of Protestant forces at the Battle of the White Mountain (Bílá Hora), near the city, in 1620. Twenty-seven Prague commoners and Czech noblemen were executed on the Staroměstské náměstí (Old Town Square) in 1621; the city ceased to be the capital of the empire, was occupied by Saxons (1631) and Swedes (1648), and went into a decline hastened by two outbreaks of plague.

The modern period. *The 18th century.* The return of more settled conditions in central Europe was marked by renewed economic growth, and Prague's population grew from 40,000 in 1705 to more than 80,000 by 1771. In 1784 the Staré Město, the Nové Město, the Malá Strana, and the Hradčany Castle complex were administratively united into one city. The merchants and the mostly German, Spanish, and Italian nobility who were active in and around Prague in this period had an enormous effect on both architecture and cultural life. Outstanding architects created magnificent palaces and gardens, often in the shadow of the castle, and churches in the Prague version of the Baroque style sprang up throughout the city.

Effects of the Industrial Revolution

The 19th century. The onset of the Industrial Revolution had major effects in Prague. The first suburb (Karlín) was established in 1817, and in the next 20 years many factories sprang up, often in association with the coal mines and ironworks at Kladno and Králův Dvůr, not far away. The population passed 100,000 by 1837, and expansion continued after the city received its first railway eight years later. The rise of a working class and of strong nationalistic sentiments had a profound effect on the city; students, artisans, and workers took to the barricades against the ruling Austrians when revolution flared briefly in 1848. Within 20 years Czechs had won a majority on the City Council, and Czech cultural life was experiencing a renascence centred on Prague. The Neoclassical building of the National Museum and the National Theatre are only two examples of the building that took place in this period. By the 1890s the first electric streetcars (trams) were running in the city, urban services were being reorganized, and a replica of the Eiffel Tower in Paris overlooked the city from Petřín Hill.

The 20th century. In 1918 Prague became the capital of the newly independent Czechoslovak republic, and a Greater Prague of 19 districts had been established within two years. By 1930 the population had reached 850,000. The city suffered a severe setback following the surrender of large parts of Bohemia and Moravia to Germany under the Munich Agreement of 1938. The citizens rose in revolt on May 5, 1945, and held the city until the Red Army arrived four days later. The post-World War II period was marked by economic reconstruction and by careful planning with a view to restoring and preserving the historic monuments of the city centre. There is also an increasing emphasis on the development of new satellite communities. The city continued to grow (although most of its recent population growth is attributable to annexation rather than to natural growth or immigration), reaching 1,173,000 inhabitants by the mid-1970s. In spite of the political vicissitudes of the 20th century, Prague had established a growing international reputation and had become a major tourist centre.

After World War II

THE CONTEMPORARY CITY

Topography and environment. From its original small riverside settlements, Prague has spread over its seven or more hills and also up tributary valleys and along riverside terraces. By the mid-1970s, the whole administrative area had grown to 191 square miles (496 square kilometres).

The lowest point in the city is 623 feet (190 metres) above sea level and the highest is 1,247 feet (380 metres), on the White Mountain. The climate of Prague is typically midcontinental, with the July average reaching 67° F (19.5° C), while that of January falls to 31° F (−0.5° C).

Demography. Prague has a number of demographic peculiarities stemming mainly from the effects of World War II. The basic problem is that there are more women than men (53.4 percent to 46.6 percent by the mid-1970s), and a sizable proportion of the female population is past the age of fertility; for example, 26 percent of the women but only 19.6 percent of the men were over 60 years of age by the mid-1970s. The natural rate of increase is very small—2.3 per 1,000 in 1975, from a birth rate of 16.5 and a death rate of 14.2 per 1,000.

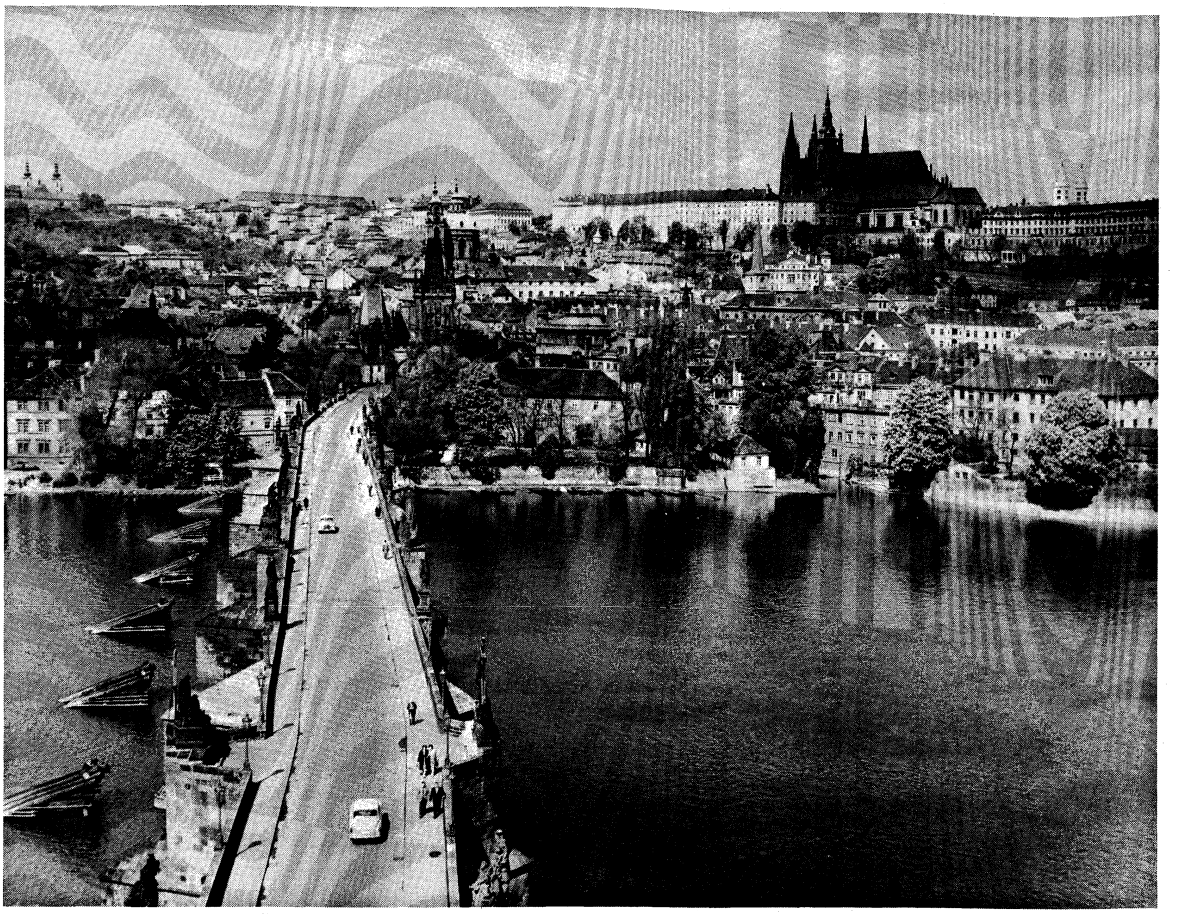
A tendency toward smaller families reflects both difficulties in housing and increased participation by both parents in the work force. Out of every 100 births in Prague by the early 1970s, about 65 percent were of a first child, 29 percent of a second child, and only 4 percent and 2 percent, respectively, of a third or later child. Future trends in the city population will depend critically on the rate of in-migration, but it appears that increase will continue through the 1970s and beyond.

Economic life and administration. Though Prague is justly renowned for its cultural life and monuments, it has played a very important role in the economic life of Czechoslovakia since the early and intensive development in the 19th century of such industries as those producing textiles and machinery. About 28 percent of the total labour force of more than 640,000 is employed in manufacturing, about 14 percent in construction, 12 percent in trade, 8 percent in education and culture and in transport and communications, and 7 percent in science and research. Nearly half the labour force is female, the proportion of women reaching 44 percent in manufacturing, 57 percent in education and culture, 66 percent in trade, and no less than 80 percent in the health field.

Women in the labour force

In manufacturing, production of machinery occupies the majority of workers, followed, in about equal numbers, by food production (including brewing of the famous Prague beer), textiles, and printing. By the end of the 1960s, 38 percent of manufacturing production was of consumer goods—a 50 percent increase over the beginning of the decade.

Since 1945 the municipal administration has been carried out by the 150-member National Committee of Prague, with power residing in an 18-member council headed by a mayor with cabinet rank in the national government. The mayor has three deputies in charge of city planning, scientific development, and cultural life. A dozen or so specialized committees deal with particular aspects of city administration and services. A similar ad-



View of Prague (Malá Strana district) across Charles Bridge and the Vltava River, with Hradčany Castle at upper right and the spires of St. Vitus' Cathedral rising behind it.
H. Armstrong Roberts

ministrative structure operates, on a smaller scale, in each of the 10 city districts, which range in size from the smaller inner-city districts—which are losing population to the suburbs but nevertheless maintain a high population density and hold most of the employment opportunities—to the much larger outer residential districts.

Municipal services. The standard municipal services—the supply of gas, electricity, and water—have been consolidated under state control since World War II and have been considerably modernized and expanded as part of overall urban planning. The high percentage of employed women has caused the authorities to turn their attention toward the provision of nurseries for the children of working mothers; there were 138 of these by the mid-1970s, supplying places for more than 7,400 children. Other facilities include swimming pools, often run in conjunction with sports organizations. On the river, the city provides mooring positions for pleasure boats.

Housing shortage

The housing problem continues to cause difficulties in the city, in spite of a building program. Of the more than 380,000 separate housing units in the city by the beginning of the 1970s, about a quarter had been built since 1945. Tourist accommodation has also expanded, and more than a million people a year (half of them from abroad) now visit the city.

As in other cities of the Communist bloc, all retail establishments—food stores, department stores, and self-service establishments—are publicly owned and part of the municipal system. There are numerous small restaurants and taverns, many of which—especially in the Malá Strana—have an intimate and historic atmosphere and offer fine views over the city and the winding river.

Public transportation is handled by buses and streetcars, or trams. More than 1,800,000 people are carried daily on this system, which is cheap and subsidized; only 12 to 15 percent of commuters use their own cars. The defects of this long-established public system have been evident for some time, and in 1966 construction started on a subway system, very deep in places; the first section was opened in 1974. Under a related reorganization, two of the large Prague railway stations were to be abolished and the main station reconstructed and enlarged to handle all rail traffic entering the city. The international airport at nearby Ruzyně, because of its favourable location in the centre of the continent, has become of major significance. A fleet of passenger boats on the Vltava is also an element in transportation.

Health services

Health services in the city are provided by some 270 health districts, each serving some 3,500 people. The 13 major hospitals in the mid-1970s (including a new hospital complex in the Motol district) had about 8.5 beds per 1,000 inhabitants.

Education and cultural life. There are nine institutions of higher education in Prague, including two universities with more than 34,000 students, but by far the most famous is the Charles University, founded in 1348, the oldest in central Europe. The Academy of Arts and the Academy of Music (with a conservatory founded in 1811) are also important. The other notable university institution is for foreign students, the University of the 17th of November (a reference to the killing of a number of Czech students by occupying forces on that date in 1939), founded in 1961, which has about 4,000 students from more than 90 countries, many of them from the developing world. The activity of the Academy of Sciences is supplemented by many specialized institutions; the academy sponsors an increasing number of international congresses. Higher education in the city benefits from a tradition that can count among its scholars and teachers the great 17th-century astronomers Tycho Brahe and Johannes Kepler and the noted modern physicist Albert Einstein, who taught in Prague in 1911–12.

Prague has a renowned and active musical life, which reaches a high point each year in the internationally

known spring music festival. The city's fine orchestras —the Prague Symphony and the Prague Philharmonic— have won reputations abroad. Theatrical traditions are also strong, with about 15 well-attended theatres in the city. There are also many museums and galleries.

Architecture

Perhaps the greatest treasures of the city, however, are the more than 1,700 officially recognized architectural and artistic monuments, ranging in period from the Romanesque through the Gothic to the Baroque, Rococo, Classical, and Neoclassical. The interiors of these buildings, which often house major art collections, have been restored since 1945. The most notable Romanesque monument is probably the simple but striking 10th-century Church of St. George, behind the north wall of Hradčany Castle. To the west of it is its more massive successor, the basically Gothic St. Vitus' Cathedral, the twin spires of which dominate the city skyline. Other Gothic monuments include the Týn Church on Old Town Square; the elegant Prašná brána (Powder Tower), marking the former city walls in what is now the busy Příkopy shopping area; the restored Bethlehem Chapel, where Hus preached; and the St. Agnes Convent, built in 1234 and notable for its collection of 14th-century paintings. The Old-New Synagogue and the tumbling, crowded gravestones of the Old Jewish Cemetery—Europe's oldest—serve as a reminder of the strong Jewish tradition in Prague life.

Baroque buildings are the city's greatest single artistic treasure, among them the splendid Valdštejn and Clam-Gallas palaces, St. Nicholas Church, and the Antonín Dvořák Museum. The geometric tiling of the Golz-Kinský Palace facade provides a distinctive glimpse of the Rococo style. Classical buildings include the Bedřich Smetana Museum on the riverside and the elegant Belvedere Palace (the former Royal Summer Palace). The National Museum and the National Theatre are the main Neoclassical buildings.

The beauty of the city is enhanced by its many parks and gardens, including a major cultural and entertainment centre in the park named for Julius Fučík (a slain resistance leader of World War II) and a large zoo in suburban Troja.

Recreational facilities also include the vast Strahov sports complex—containing three stadiums, the largest of which, Spartakiáda Stadium, holds 250,000 spectators and is used for the mass gymnastic display known as the Spartakiáda—as well as a number of other sports and cultural centres, with emphasis on facilities for youth. The film studios at Barrandov, on the city outskirts, have produced a number of high-quality motion pictures and there is a museum of modern sculpture at Zbraslav.

PROSPECTS FOR THE FUTURE

The 1976–80 Five-Year Plan, approved at the 15th Congress of the Czechoslovakian Communist Party in 1975, is primarily a long-term plan, dealing with developmental problems of the city (through the year 2000 in some respects, although for the most part only through 1980 or 1990). Housing is a matter of critical concern, and the plan calls for construction of some 58,000 new apartments to be completed by 1980. These are to be built as 11 new residential developments on the outskirts of Prague; remodelling and reconstruction are planned for the older, more historic parts of the city. The other areas of strongest concern are public services, especially health care, education, public buildings, and transportation. It

Hospitals

was expected that by 1980 major new hospital facilities would be completed at Bulovka and Motol, the Vysočany Hospital would have been reconstructed, and 10 lesser facilities would be built or expanded. Some 32 schools, mostly in the newer, outlying districts of the city where the younger population is concentrated, were to be built. Among the projects involving public buildings, two of the most important are the new Czechoslovak Palace of Congresses and a modernized National Theatre. The principal transportation projects include completion of some 12 miles (20 kilometres) of new subway lines and the construction of the first links in a major system of circular and radial roads.

BIBLIOGRAPHY. BOHUMIL LANDISCH, *Praha* (1970) and *Praha, matka měst*, 3rd ed. (1970), are annotated (in Czech, German, English, and French) photographic publications dealing with Prague's history. JIRI DOLEZAL and EVZEN VESELY, *Památky staré Prahy* (1966), is a popularized photographic survey of historical Prague, with brief commentary. EUGEN VASILIAK, *Nad Prahou (Prague Seen from Above)* (1966), is a book of aerial photographs of old and new Prague with accompanying text. JOSEF CIBULKA, *The Crown Jewels of the Bohemian Kingdom* (Eng. trans. 1969), provides a description of the Czech coronation jewels (well illustrated), discussing their history and artistic value. EMANUEL POCHE, *The Golden Lane on Prague Castle* (Eng. trans. 1969), describes the picturesque street of Hradčany Castle. See also JAROMIR NEUMANN, *The Picture Gallery of Prague Castle* (1967). ALOIS SVOBODA, *Prague* (1968), is a guide for the tourist (in English). JOSEF JANACEK (ed.), *Dějiny Prahy* (1964), is a comprehensive collective publication covering the detailed development and history of the city from the earliest times to 1960.

(Ja.K./R.H.O.)

Prayer

Prayer is an act of communication by man with the sacred or holy—God, the gods, the transcendent realm, or supernatural powers. Found in all religions in all times, prayer may be a corporate or personal act utilizing various forms and techniques. Prayer has been described in its sublimity as "an intimate friendship, a frequent conversation held alone with the Beloved" by St. Teresa of Ávila, a 16th-century Spanish mystic.

NATURE AND SIGNIFICANCE

Prayer is a significant and universal aspect of religion, whether of primitive peoples or of modern mystics, that expresses the broad range of religious feelings and attitudes that command man's relations with the sacred or holy. Described by some scholars as religion's primary mode of expression, prayer is said to be to religion what rational thought is to philosophy; it is the very expression of living religion. Prayer distinguishes the phenomenon of religion from those phenomena that approach it or resemble it, such as religious and aesthetic feelings.

Prayer as the primary mode of expression in religion

Historians of religions, theologians, and believers of all faiths agree in recognizing the central position that prayer occupies in religion. According to the American philosopher William James, without prayer there can be no question of religion. An Islāmic proverb states that to pray and to be Muslim are synonymous, and Sadhu Sundar Singh, a modern Christian mystic of India, stated that praying is as important as breathing.

Of the various forms of religious literature, prayer is considered by many to be the purest in expressing the essential elements of a religion. The Islāmic Qur'ān is regarded as a book of prayers, and the book of Psalms of the Bible is viewed as a meditation on biblical history turned into prayer. The *Confessions* of the great Christian thinker St. Augustine (354–430) are, in the final analysis, a long prayer with the Creator. Thus, because religion is culturally and historically ubiquitous, if prayer were removed from the literary heritage of a culture, that culture would be deprived of a particularly rich and uplifting aspect.

From its primitive to its mystical expression, prayer expresses a desire on the part of men to enter into contact with the sacred or holy. As a part of that desire, prayer is linked to a feeling of presence (of the sacred or holy), which is neither an abstract conviction nor an instinctive intuition but rather a volitional movement conscious of realizing its higher end. Thus, prayer is described not only as meditation about God but as a step, a "going out of one's self," a pilgrimage of the spirit "in the presence of God." It has, therefore, a personal and experiential character that goes beyond critical analysis.

Prayer is also linked to sacrifice, which seems to support prayer as a cultic—as well as a personal—act and as a supplement to the bare word of man in his attempts to relate to the sacred or holy. In any case, the sacrificial act generally precedes the verbal act of prayer. Thus, the presentation of an offering often prolongs prayer and is viewed as a recognition of the sovereignty and benefi-

cence of the deity or supernatural powers. The word of man (in prayer), however, apart from a concomitant sacrificial act, is itself viewed as the embodiment of sacred action and power.

When prayer becomes dominating and manipulative in its intent, it becomes magic. With words and songs, man thus believes that he can ask, conjure, and threaten the sacred or supernatural powers. Imprecation and incantation become, in effect, "oral talismans" (charms). The effectiveness of such magical prayer is believed to depend on the recitation of a precise formula, or rhythm, or on the saying and repeating of the divine name. Manipulation by magic, however, is neither the explanation nor the essence of prayer but rather its deviation and exploitation, a tendency that is to be noticed whenever prayer departs from its basic and essential meaning—*i.e.*, the expression of a desire to enter into contact with the sacred or holy.

ORIGIN AND DEVELOPMENT

Theories of the origin of prayer

During the 19th century, when various evolutionary theories were in vogue, prayer was viewed as a stage in the development of religion from a magical to a "higher" stage. Such theories, which saw in prayer no more than a development of magic or incantation, failed to recognize the strictly personal characteristics of prayer. Even if a scholar could prove the chronological precedence of magical incantations to prayer—which has thus far not been done—he would be derelict in his scholarly duty if he saw in such a precedence the only explanation of prayer. The origin of prayer is to be found—essentially and existentially—in the recognition and invocation of the creator-god, the god of heaven.

Though some scholars, such as Costa Guimaraens, a French psychologist in the early 20th century, have attempted to trace prayer back to a biological need, the attempt, on the whole, has been unsuccessful. If sometimes —especially with exceptional subjects or subjects with fragile nervous systems—the act of prayer is accompanied by corporal phenomena (*e.g.*, bleeding, shaking), such phenomena can accompany it without having provoked it and without explaining its deep inspiration. In order to analyze normal prayer psychologically it is especially important to choose normal subjects. Affective sources such as fear, joy, and sadness doubtless play a role in prayer. Such affectations are expressed in prayers recorded in various religions and particularly in the Psalms of the Bible; but they do not explain the recourse to prayer itself, which is explained by a motivation deeper than affective elements. The cause and occasion of prayer must not be confused.

Moral sentiments also are integrating elements, but they are accidental to the development of prayer; virtue is not necessarily expressed in the act of praying because there exist atheists of incontestable morality. Morality is more a consequence than a cause of prayer; and it follows more than it prepares for the development of the religious man.

William James and psychologists such as Joseph Segond describe prayer as a "subconscious" and "emotional effusion," an outburst of the mind that desires to enter into communication with the invisible. Experiences of prayer very often, in fact, do include "cries from the heart," "inexpressible laments," and "spiritual outbursts." The psychological explanation has the advantage of probing the subconscious, of describing the various forces that act within men's psyches; but the emergence of the subconscious in the act of prayer is not the essence of prayer since it minimizes the role of intelligence and the will. Among what are called the higher religions (*e.g.*, Judaism, Christianity, Islām, Hinduism, Buddhism), divine action, which is the object of the human action of prayer, violates neither man's consciousness nor his freedom.

Sociologists often explain prayer in terms of the religious environment, which plays an indubitable role in spiritual behaviour. Though prayer supposes a personal belief, that belief is, to a great extent, provided by society. Society creates and regulates social and religious rites and liturgies to express its beliefs, but to explain the origin of prayer solely in terms of an environmental context would be to neglect the inner, personal origins of prayer. That belief is transmitted by society is incontestable, but the channel is not to be viewed as the source. Society itself is, so to speak, a tributary of beliefs that are both received from and given to the collective whole and also from and to each of its members. The collective forms may influence personal prayer, but they do not explain it.

The vertical (divine–human) as well as horizontal (social) dimension of prayer is also expressed in the alternation between speech and silence. Whereas magical formulas are used to coerce the supernatural, liturgical language, even when incomprehensible to the congregation, seeks to lead the participants into an apprehension of the mystery of the divine. In the presence of the mystery of the divine, man often discovers that he can only stammer or that his speech often falters. When this occurs, he frequently expresses his "fear and love" (Luther) or "*tremendum et fascinans*"—*i.e.*, fear and attraction (according to Rudolf Otto, a modern German historian of religion), in apophatic (negative) formulas. Speech with the divine is, in such cases, followed by silence before men, as one apprehends the inexpressible (*i.e.*, the sacred or holy). Religious language, like silence, thus expresses the distance and inadequacy of man in relation to the divine mystery.

TYPES OF PRAYER

Because the various types of prayer are connected and permit a flow from one type to another, it is difficult to conceive of them in terms of rigid classifications. They are enumerated here more on the basis of psychology than on history.

The central position of request in religion

Petition. The role of the request in religion has played such a central part that by metonymy (using a word for another expected word) it has given its name to prayer. However contestable this may sometimes be, it is impossible to refuse to recognize the importance of request, whether it be for a material or spiritual gift or accomplishment. The requests that occur most often are for preservation of or return to health, the healing of the sick, long life, material goods, prosperity, or success in one's undertakings. Request for such goals may be tied to a magical invocation; it may also be a deviation from prayer when it takes the form of a bargain or of a request for payment due: "In payment of our praise, give to the head of the family who is imploring you glory and riches" (from the Ṛgveda, a sacred scripture of Hinduism). Christianity has never condemned material requests but rather has integrated them into a single providential order while at the same time subordinating them to spiritual values. Thus, in essence though not always in practice, requests are only on the fringe of prayer. As a religion adopts more spiritual goals, the requests become more spiritual: in the *Choephori*, a play written by Aeschylus (a Greek tragic poet of the 6th–5th centuries BC), Electra, the daughter of King Agamemnon, prays, "Grant that I may be a more temperate and a more pious wife than was my mother." Other examples of the transformation to spiritual goals may be seen in the prayers of the ancient Babylonian and Assyrian kings who asked for the fear of God, rather than material benefits, and that of a priest of the Ewe (a West African people) who even asks of his god "That I remain near you and that you remain near me."

Confession. The term confession expresses at the same time an affirmation of faith and a recognition of the state of sin. In Mazdaism (Zoroastrianism and Parsiism), as in ancient Christianity, the confession of faith accompanies the renunciation of demons. The *Confessions* of St. Augustine also illustrate this dual theme. In a similar fashion, ancient and primitive men recognize that their sins unleash the anger of the gods. To counter the divine wrath, a Ewe, for example, throws a little bundle of twigs —which symbolizes the confessor's sins—into the air and he says words symbolizing the deity's response, "All your sins are forgiven you."

The admission of sin cannot be explained only by anguish or by the feeling of guilt; it is also related to what is deepest in man—*i.e.*, to what constitutes his being and his

action (as noted by Karl Jaspers, a 20th-century philosopher). The awareness of sin is one of the salient features of religion, as, for example, in Hinduism: "Varuṇa is merciful even to him who has committed sin" (Ṛgveda). Confession is viewed as the first step toward salvation in both Judaism and Christianity; in Buddhism, monks confess their sins publicly before the Buddha and the congregation two times every month.

Situated at the most personal level of man, sin places him directly before God, who alone is able to grant pardon and salvation. The Miserere ("Lord, have mercy," Psalm 51) of the ancient Israelite king David expresses repentance for sin with an intensity and depth that has a universal value. One of the results of such a dialogue with God is the discovery of the dark depths of sin.

Intercession. Members of primitive societies have a clear sense of their solidarity in the framework of the family, the clan, and the tribe. This solidarity is often expressed in intercessory prayer, in which the needs of others are expressed. In such societies, the head of the family prays for the other members of the family, but his prayers also are extended to the whole tribe, especially to its chief; the primitive may pray even for those who are not members of his tribe (*e.g.*, strangers or Europeans).

The purposes of prayer

Intercessory prayers are also significant in Eastern and ancient religions. In the hymns of the Ṛgveda, the father implores the god Agni (god of fire) for all of those who "owe him their lives and are his family." In the Greek play *Alcestis* by Euripides (5th century BC), the mother, on her death, entrusts the orphans she is about to leave to Hestia, the goddess of the home. Among the Babylonians and the Assyrians, a priesthood was established primarily to say prayers of intercession.

Prayers of intercession to the divine are supported by mediatory minor gods or human protectors (alive or dead), marabouts (dervishes, or mystics, believed to have special powers) in Islām, or saints in Christianity, whose mediation ensures that the prayer will be efficacious.

In biblical religion, intercession is spiritualized in view of a consciousness of the messianic (salvatory) mission. Moses views himself as one with his people even when they fail in their duty: "Pardon your people," he prays, "or remove me from the Book of Life." Such solidarity finds its supreme form in the prayer of Christ on the cross —"Father, forgive them, for they know not what they do"—which St. Stephen (the first recorded Christian martyr) and other martyrs repeated in the course of their sufferings.

Praise and thanksgiving. Praise, in the prayer of primitive peoples, can be traced to salutations, such as in the prayer of the Hottentots (of South Africa) to the New Moon—"Welcome." Praise among most of the ancient peoples was expressed in the hymn, which was primarily a prayer of praise (whether ritual or personal) for the gift of the created world. Israel praises its Creator for "his handiwork," as does the Qur'ān. Contemplation of the majesty of the universe thus often gives rise to a prayer that is not always completely free from pantheism (the divine in all things) and that can be found all the way from the nature hymns of Oriental religions to the effusions of J.-J. Rousseau, the 18th-century French moralist, embracing the trees and contemplating the sunrise.

Praise—in addition to concerns for the created world—plays an important role in the prayer of mystics, for whom it is a form of adoration. Praise in this instance constitutes an essential element of the mystic experience and celebrates God, no longer for his works, but for himself, his greatness, and his mystery.

When the great deeds of God are the theme of praise, it becomes benediction and thanksgiving. Even when words denoting thanksgiving are not present, the substance of thanksgiving is manifest, even, for example, for the Pygmy of Central Africa, who says to his god, "Waka (meaning God), you gave me this buffalo, this honey, this wine." Mealtime prayers, frequently enunciated in both ancient and modern religions, give thanks for the goods of the earth and are linked to the giving of an offering.

In Christianity, Christ is discovered as the gift of God and in his mission the economy (or mode of operation) of salvation. Thus, the giving of thanks is viewed as man's response, as a spiritual reaction to the benefit received—*i.e.*, the mediatory work of Christ. Because of the cultivation of this expected response, praise and thanksgiving occupy a central position in Christian prayer and in the liturgy, so much so that its name is given to the Eucharistic Prayer (*i.e.*, the Prayer of Thanksgiving).

Adoration. Adoration is generally considered the most noble form of prayer, a kind of prostration of the whole being before God. Even if the prayer of request is predominant among primitives, they are seized with the feeling of fear and trembling before the numen (spiritual power) of all that is mana (endowed with the power of the sacred or holy) or taboo (forbidden because of association with the sacred). Names given to the divinity in prayers of adoration express dependency and submission, as, for example, in the prayer of the Kekchí Indians of Central America: "O God, you are my lord, you are my mother, you are my father, the lord of the mountains and the valleys." To express his adoration man often falls to the ground and prostrates himself. The feeling of submissive reverence also is expressed by body movements: raising the hands, touching or kissing a sacred object, deep bowing of the body, kneeling with the right hand on the mouth, prostration, or touching the forehead to the ground. The gesture often is accompanied by cries of fear, amazement, or joy; *e.g.*, *has* (Hebrew), *hū* (Islām), or *svāhā* (Hindu).

Motivation behind adoration

Adoration takes on its fullest meaning in the presence of the transcendental God who reveals himself to man in the religions of revelation (Judaism, Christianity, and Islām). In the Old Testament prophet Isaiah's vision of the holy (Isa. 6:3), the seraphim (winged creatures) chant to Yahweh: "Holy, holy, holy is the Lord of hosts; the whole earth is full of his glory." This hymn of adoration became a part of the Christian liturgy. The supreme form of adoration, however, is generally considered to be holy silence, which can be found in primitive religion and in ancient religions, as well as in the "higher" religions, and among mystics it expresses the most adequate attitude toward the immeasurable mystery of God: "I am in a dark sanctuary, I pray in silence; O silence full of reverence" (Gerhard Tersteegen, an 18th-century Protestant mystic). Silent adoration is often viewed as the introduction or the response to an encounter with the sacred or holy.

Unitative: mystical union or ecstasy. Ecstasy is literally a departure from, a tearing away from, or a surpassing of human limitations and also a meeting with and embracing of the divine. It is a fusion of being with being, in which the mystic experiences a union that he characterizes as a nuptial union: "God is in me and I am in him." The mystic experiences God himself in an inexpressible encounter because it is beyond the ordinary experiences of man. The mystical union may be a lucid and conscious progression of contemplative prayer, or it may take a more passive form of a "seizing" by God of the one who is praying.

Prayer as an expression of the mystical experience

The mystic, by his goals and actions, is removed from both the world and himself. He discovers in the light and majesty of the divine his own poverty and nothingness and is thus torn between the contemplation of the greatness of God and his own meagreness. St. Francis of Assisi exemplified this dichotomy in his prayer: "Who are you, O God of sweetness, and who am I, worm of the earth and your lowly servant?" Ecstatic prayer goes beyond the frame of ordinary prayer and becomes an experience in which words fail. Mystics speak in turn of unity (*e.g.*, the 3rd-century-AD Greek philosopher Plotinus), of great pleasure (Augustine), or of intoxication (Philo). It is found in the accounts of Hindu, Persian, Hellenistic, and Christian mystics. "You are me, supreme divinity, I am you," says Nimbāditya. The Ṣūfī (mystic) of Islām Jalāl ad-Dīn ar-Rūmī sighs in the same words as a Christian mystic, Angelo de Foligno: "I am you and you are me." Mechthild von Magdeburg develops the same kind of reciprocity: "I am in you and you are me. We cannot be closer. We are two united, poured into a single form by an

eternal fusion." Such reciprocity that is so complete that it becomes identity is the supreme expression of ecstatic prayer. It is found in all of the mystic writings, from the Orient to the West.

FORMS OF PRAYER IN THE RELIGIONS OF THE WORLD

The forms that prayer takes in the religions of the world, though varied, generally follow certain fixed patterns. These include: benedictions (blessings), litanies (alternate statements, titles of the deity or deities, or petitions and responses), ceremonial and ritualistic prayers, free prayers (in intent following no fixed form), repetition or formula prayers (*e.g.*, the repetition of the name of Jesus in Eastern Christian Hesychasm, a quietistic monastic movement, or the repetition of the name of Amitābha Buddha in Japanese Buddhism), hymns, doxologies (statements of praise or glory), and other forms.

Primitive religions. Prayer is one of the most ancient expressions of religion. Practically the only evidence of early forms left, however, is that to be noted among the most primitive peoples of today. Together with his dependency in relation to his tribe, the primitive man is aware of his dependency in relation to the Supreme Being. He often addresses his prayers, however, to various numina (spiritual powers): the dead, the divinities of nature, protective gods or actor gods, the Supreme Being localized somewhere in heaven, or a feminine divinity linked to the earth (*i.e.*, the great mother). It is impossible to determine the historical precedence of one over the others, and it is difficult to describe the most primitive prayer because certain forms escape modern scholars, so much so that it has been assumed by some that prayer was absent in earliest religion. The first form may have been a cry, then brief formulas repeated as incantations, such as "Come . . . hear me . . . have pity." (*e.g.*, Algonkin Indians of North America).

Internalized prayer is found among the primitive Eskimos of North America, the Algonkin tribes, the Semangs (of the Andaman Islands in the Indian Ocean), and the Aboriginals of Australia. Prayer in gestures is found among the Semangs and other primitive peoples. Another form is spontaneous prayer, without any precise formulation, which is found, for example, among the Negritos of the Philippines and the Alacaluf (Halakwulups) of Tierra del Fuego. More developed liturgies and prayer vigils are found among the Negritos and the Pygmies of Gabon.

The concerns of prayers of primitive peoples

The prayers of peoples of a primitive culture generally are concerned with the self (egoistic) and concerned with well-being (eudaemonic) at the same time; they are clearly pragmatic, concerned above all with food, protection, and posterity. But the higher forms of adoration and recognition of obligations, of confidence, and self-abandonment are to be noted. Among the Australian Aboriginals are prayers on tombs for the dead, so they may be received in heaven, and prayers are also addressed to the spirits of ancestors. Request and pardon accompany sacrifices in the propitiatory rites of the Semangs. Of special interest is the fact that the Wiradjuri-Kamilaroi of Australia practice public prayer on only two occasions: the burial of a man and the consecration of puberty. They believe that excessive prayer serves no purpose.

Ancient civilizations. From the 3rd millennium BC to the beginning of the Christian Era, forms of prayer changed little among the Assyrians and Babylonians and their descendants. The oldest forms are composed of hymns and litanies to the moon goddess Sin and to the god Tammuz. Though some songs of joy have been found, most are adjurations. Some hymns of thanksgiving tell of gratitude to the divinity for victory over an enemy. One such hymn, addressed to Marduk (the Babylonian sun god), apparently goes back to the 12th century BC. A number of hymns of later date celebrate the king, but their intent is to request divine protection first for him and his country. Preserved in the library of Ashurbanipal (7th-century-BC Assyrian king) at Nineveh is a rather long hymn to the goddess Nana (queen of the world and giver of life), the consort of the god Nabu, son of Marduk and a god of wisdom and science. There also is a long acrostic poem in praise of the god Marduk, creator of heaven and earth, and hymns that the Babylonians recited at the new year, at the beginning of spring, and at the celebration of Marduk.

Other hymns accompany sacrifices, such as in the offering of a young gazelle in place of humans. A most important form of prayer, however, is found in the conjurations and exorcisms of a priest or believer and in lamentations, which are particularly numerous and which often end in a refrain similar to a litany.

Ancient Egyptian piety is preserved in numerous precepts engraved on the backs of scarabs. These engravings sometimes include praises of the divinity ("All good fates are in the hand of God"), statements of confidence, or requests for protection for the one praying and for his whole family ("God is the protector of my life; the house of one favoured by God fears nothing"). Hymns of thanksgiving, such as that of the artist Nebre, who obtained from the god Amon the healing of his son who had been struck with illness because of Nebre's fault, are numerous in ancient Egyptian religion. Protective magic, widely practiced, also utilized formulas of incantation, recited or written, and amulets (charms). Some of the incantation formulas (anonymously written) come from the earliest times, and others, more recent but no less efficient, were composed by magicians. In order to increase their authority and efficacy, several, such as those composed by pharaoh Ramses III (12th century BC) and preserved in Cairo, were attributed in origin to the gods themselves.

Collections of formulas, such as the Egyptian Pyramid Texts and the Book of the Dead, were compilations of magical prayers that allowed the dead to forestall all the dangers and meet all the eventualities. In particular, they contain negative confessions in which the dead man justifies himself before the court of Osiris (god of the dead). The funeral liturgies of the ancient Egyptians have preserved lamentations that echo the family in mourning. Hymns written on papyrus that are compositions in honour of a divinity and that were recited during sacred ceremonies have also been preserved. Such are the hymns of the pharaoh Akhenaton (Amenophis IV, 14th century BC) to the god Aton and the hymns in honour of the god Amon-Re that boast of divine benefits and sometimes confess misery and sin.

Ceremonial prayers

In Greece, poetic prayer can be distinguished from ceremonial prayer. The first, like all of the liturgical prayers, contains three essential parts: the invocation of the god, a justification for fulfillment (*e.g.*, sacrifices offered, favours given and received), and a conclusion that formulates the request, such as in the prayer of Diomedes to the goddess Athena in the *Iliad* (written by the Greek poet Homer in the 8th century BC). Generally, the ceremonial prayer followed a ritual pattern: washing of the hands, the prayer proper, then sacrifice and libations. The prayer initiated the liturgical action; without it there could be no ceremony. Prayers often were transformed into hymns, a characteristic of Greek religion. One of the oldest known Greek hymns is that of women devotees of Dionysus, the god of wine and fertility. That such hymns were not always sublime in character is attested to by the comment of a 6th-century-BC Greek philosopher in regard to a Dionysiac festival.

> If they did not hold a procession and sing a hymn to the genitals, it would be an outrageous performance. Hades and Dionysus, in whose honour they rage and celebrate the Bacchus rites, are one and the same.

Another ancient hymn is a morning hymn to Asclepius, the god of healing. All the hymns begin with invocations of the names of the gods to whom they are addressed. The invocation was believed to have an almost magic value. Though there are many individual Greek hymns in existence, the only official collection remaining contains the Orphic Hymns (addressed to the ancient hero Orpheus); it dates from the Greco-Roman period (*c.* 3rd century BC–*c.* 4th century AD).

Roman prayers begin with an invocation to the divinity. Addressing the god is of capital importance and one must be careful not to address the wrong god. In order to avoid this error, there were litanies of 15 gods and goddesses.

The prayer itself generally takes two forms, depending on whether it implies a request or is simply limited to praise. The prayer of request has a juridical pattern in which the offering, as a contractual element, dominates. The offering is what jurists call bail bond, a guarantee. The prayer of request's effectiveness depends on a precise formulation, with parallelisms, solemn repetition, and accumulation of synonyms. The verb *precor* ("I pray") is reinforced by many synonyms. Prayers of praise developed out of meditation or experiences of religious elevation and utilized various patterns in both public and private ceremonies. An example of collections of prayers of praise is preserved in the *Verba pontificalia* ("Priestly Words").

Another form of prayer is the *votum* ("vow"), in which a person undertakes to offer to the divinity, in exchange for divine favour, a sacrifice, the building of a temple, or other such offerings. It is a kind of bargain in which is still felt the prudence of the peasant who has experienced failure. These *vota* ("vows") become more numerous than other prayers the farther one goes from the historical origins of Rome. The most solemn form of the vow is the *devotio* ("act of devotion"), by which a chief offers himself to the divinity in order to obtain victory.

Religions of the East. Although the religion of the Vedas contains private prayers, it gives importance and hieratic stature to liturgical prayer, which may or may not include sacrifice. There exists a whole series of hymns, such as the morning hymn addressed to Agni (the god of fire), who brings light, and to the two Aśvin (twin gods of light). There is also an evening prayer, the *sāvitū,* more precisely a prayer for dusk, which the disciple of the Brahmins (priestly teachers) says at nightfall until the stars appear, and a benediction formula. The gestures of adoration (*upasthāna*) in effect give more intensity to the prayer. The prayers that accompany sacrifices and the numerous hymns of the Ṛgveda (a collection of sacred ritualistic lyrics), which were composed by the members of the priestly caste according to a stereotyped and schematic form, are addressed to the greatness of the divinity in exaltation of his great deeds.

In Hinduism there is an elementary form of prayer—*i.e.*, an affirmation of homage and refuge with the divinity. More frequent is a more elaborate prayer in two forms: *dhyāna* ("meditation") and the *stotra* ("praise"). The *stotra* occurs in a variety of subforms and generally opens with an invocation. It is often characterized by a sort of litany of the titles given, for example, to Viṣṇu (the preserver god) or Śiva (the destroyer god). The *Śivasahasranaman* ("The Thousand Names of Śiva") lists 1,008 titles. In this hymn, each strophe ends with the same refrain. When recited with concentration and pure heart, these prayers are believed to achieve remission of sins.

Hindu mystical prayer

Hindu mysticism gives great importance to spoken prayer, which, by progressive absorption, leads to ecstasy. The scale of the prayer of Hindu mystics is exemplified in the five stages of *bhakti* ("devotion") as taught by the Hindu mystic Caitanya (15th–16th century AD), who uses the metaphor of love in social relationships: *śānta* (peaceful love), *dāsya* (servant of God love), *sakhya* (friendship with God), *vātsalya* (filial attitude toward God), and *mādhurya* (love of God as one's lover). "When I was no longer capable of recognizing, I said me and mine. I am you and you are mine" (*Nalayiram*).

In Chinese Buddhism and Taoism, in addition to prayer that accompanies sacrifice, there is the monastic prayer (mu-yu), which is practiced morning, noon, and night to the sound of a small bell. There is also a prayer for the dead, related to the transmigration of souls, which is recited at funerals, the 30th day, the anniversary of the death, and the celebration of the deceased's day of birth. Taoism gave increased importance to this latter form.

Private prayer prepares the way for liberation and illumination. The *tsai-fei* is a prayer—to accompany abstinence—that monks will recite for a believer on payment of alms. Other prayers accompany vows and pilgrimages. Both Buddhist monks and laymen use a string with 108 beads, which monks always carry in their hands.

Religions of the West. In Judaism is one of the best known collections of prayers, the 150 psalms in the Bible. In these psalms, which always presuppose a collective witness, though they may be used by an individual privately, praise is descriptive (God is . . .) or narrative (God does . . .) in nature. Also included are hymns, exhortations to praise God, and supplications. The psalms of request include lamentations and songs of confidence or gratitude. Whether individual or collective, the psalms have a rather similar structure: a cry to God, a confession of sins, a protestation of innocence, and imprecations against one's enemies.

To the prayers of the Bible, the rabbis (religious teachers and leaders) added the Shema ("Hear"), which is a confession composed of three quotes from the Bible (Deut. 6:4–9, 11:13–21, Num. 15:37–41) with attendant blessings and which the Israelite recites daily. At the time of Christ, there appears the prayer *par excellence*, the *tefilla* or *ʿamida* (standing prayer), also called *shemone ʿesre* ("18 Benedictions"), which every Israelite recites two or three times a day. To these must be added the benediction before eating that raises the meal to the level of the dignity of a religious act.

Relationship of Christian to Jewish prayers

Christianity preserves the doxologies and benedictions from its Jewish heritage. It adds to them the Lord's Prayer, psalms, hymns, and canticles, the first specimens of which are furnished by the New Testament (*e.g.*, the *Nunc dimittis*, "Now let your servant depart"). Christian prayer, like that of other religions, includes liturgical prayer and personal prayer. Liturgical prayer frames and explains more especially the sacraments of Baptism and the Eucharist (Lord's Supper).

The liturgical collection, for Sundays as well as other days, includes readings from the Bible, collects (brief prayers including an invocation, petition, and conclusion in which the name of Jesus is called upon), and a litany (general prayer) for the intentions of the universal church. During the Eucharist, there is a consecration of the bread and wine to be used in the sacred meal. This consecration prayer is called the Eucharistic (Thanksgiving) Prayer, a long prayer in which the element of thanksgiving is dominant. Addressed to the Father, through the mediation of the Son, and in the Holy Spirit, this prayer develops, like the Jewish liturgies, from praise, to thanksgiving, to the memorial (or anamnesis), and finally to an invocation of the Spirit (epiclesis). Originally improvised and spontaneous, this liturgical prayer became fixed in stereotyped forms, first in the West, then —though with more flexibility—in the East.

The first Christians retained the custom of praying three times a day, reciting the "Our Father" (Lord's Prayer). Special times for prayer are morning and evening. Christ's custom of praying at meals (as a devout Jew) is also maintained. This framework can and does favour the life and spirit of prayer that make a Christian existence, according to the words of Clement of Alexandria, a 2nd–3rd-century theologian, "an uninterrupted celebration." Bible readings, silent prayer (in the West especially), brief, fervent invocations, and the repetition of formulas like the Kyrie eleison ("Lord, have mercy") in the East have enriched spiritual life and have led monks and laymen to contemplative prayer, as is shown by the growth of mysticism in both the West and East.

From its beginning in the 7th century AD, the most important part of Islāmic liturgy has been the ritual prayer called the *ṣalāt* (daily prayer), in which both Christian and Jewish influences can be seen. This minutely detailed prayer is recited while the suppliant turns toward Mecca (in Saudi Arabia) five times a day. On Friday, the *ṣalāt al-jūmʿah* (Friday prayer) replaces the noon prayer. It is celebrated by the community in the principal mosque and includes preaching and a *ṣalāt* of two ritual bowings. Twice a year, at the end of Ramaḍān and the 10th month, a solemn *ṣalāt* is celebrated, similar to Friday's.

Islāmic prayer is an act of adoration of Allāh (God) and thus it would not be suitable to add a request. Before adoring God the believer must purify himself by means of ablutions in pure water or, failing this, in sand. The prayer is accompanied by a meticulous ceremonial with

prostration of the body (*rak'ah*). The sense of adoration and conversation with Allāh has led many spiritual Muslims to the heights of mysticism (Ṣūfism).

In Mazdaism, Avestan (scriptural) prayer, sacerdotal prayer, and the prayer common to priests and laymen alike can be distinguished. In the very first poem of the Avesta, Zoroaster (Zarathushtra) presents himself to Ahura Mazdā (the Good Lord) in a prayer that ends with these words: "I will sing for you again praises of great value." What is characteristic of these hymns is that man proceeds almost exclusively by questions and answers. Only priests can understand the ceremony of the *Yasna* (the sacrifice), during which they recite verses from the Avesta, adding to it the *Visp-rat* (shorter liturgy), with or without the *Vidēvdāt* ("Law against the Demons"), which is concerned with ritual purity. Songs (involving light symbolism) accompany the five fire ceremonies that are celebrated daily. There are also ceremonies in which both priests and laymen participate. The great Bāj, a ritual offering of consecrated bread, grain, and butter, begins with a long preface: "In the name of God, Lord Ormazd, may your power and glory increase." The Satum, in praise of the dead, is recited at the beginning of a meal prepared in their honour every month for the first year after a death and then on each anniversary. Other prayers accompany benedictions, especially those used at the consecration of fire, initiation, and marriage. To these must be added the prayers of great purification.

Conclusion. Though historians of religion, psychologists, and anthropologists debate various theories concerned with the origin of prayer, the act of prayer itself is of great significance to the believers of all religions, whatever their inspiration, revealed or otherwise. Ludwig Feuerbach, a 19th-century German philosopher, summed up the significance of prayer when he stated, "The most intimate essence of religion is revealed by the most simple religious act: prayer."

As a religious phenomenon, prayer—in terms of its evolution—appears to be neither universally progressive nor progressively regressive. Its great moments and the appearance of men of prayer at various times, whether simple men or men of genius, are found throughout its long history, which thus marks it as a significant and characteristic element of most, if not all, religions. Whether halting or mystical, ceremonial or personal, prayer expresses the experience of a mystery that envelopes and surpasses man. In the presence of that mystery, prayer seeks and establishes dialogue.

BIBLIOGRAPHY. The classical work is still that of F. HEILER, *Das Gebet*, 5th ed. (1923), which provides the bibliography of the subject up to 1923 and defends the theory of religious syncretism. R. BOCCASSINO (ed.), *La Preghiera*, 3 vol. (1967), is a historical and psychological study of the whole subject and includes an up-to-date bibliography. G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans. *Religion in Essence and Manifestations*, 1938), is an excellent overview and general introduction of the psychology of prayer. Also of a psychological bent are WILLIAM JAMES, *The Varieties of Religious Experience* (1902); and H.U. VON BALTHASAR, *Das betrachtende Gebet* (1955; Eng. trans., *Prayer*, 1961). For prayer in the various religions, in addition to *La Preghiera*, one should refer to M.P. NILSSON, *Geschichte der griechischen Religion*, 2nd ed., 2 vol. (1955–61; Eng. trans., *A History of Greek Religion*, 2nd ed., 1949); A. FALKENSTEIN and W. VON SODEN, *Summerische und akkadische Hymnen und Gebete* (1953); G.C. LOUNSBERY, *La Méditation bouddhique* (1935; Eng. trans., *Buddhist Meditation in the Southern School*, 1950); E. CONZE, *Buddhist Meditation* (1956); R.C. ZAEHNER, *The Teachings of the Magi* (1956) and *The Dawn and Twilight of Zoroastrianism* (1961). For a recent bibliography on biblical prayer, see the article "Prière," in the *Dictionnaire de la Bible*, suppl. 8, pp. 604–606 (1968). I. ELBOGEN, *Der jüdische Gottesdienst in seiner geschichlichen Entwicklung*, 4th ed. (1962), is a classic on Jewish prayer, to which should be added A.Z. IDELSOHN, *Jewish Liturgy and Its Development* (1967). On the origins of Christian prayer, see A. HAMMAN, *La Prière*, 2 vol. (1959–63).

(A.G.Ha.)

Precambrian Time

Precambrian time covers the interval between the formation of the crust of the Earth about 4,600,000,000 years ago and the beginning of the Cambrian Period, which started about 570,000,000 years ago. The Cambrian marks the earliest stage in the history of the Earth when extensive and varied forms of life were preserved as fossil remains in sedimentary rocks.

It was supposed formerly that life originated early in Cambrian times and that there was a fundamental distinction in this respect between the Precambrian and all later geological periods. It is now known that primitive forms of life were in existence over 3,000,000,000 years ago and that living matter was present throughout most, if not all, of Precambrian time. The nomenclature originally designed to bring out this supposed distinction persists, however, although it is now recognized that the Precambrian Era covers the greater part of geological time and that the Cambrian and all later geological periods represent little more than a tenth of the time that has elapsed since the formation of the Earth's crust.

There is, as yet, no generally accepted scheme for the division of Precambrian time. The terms Archeozoic and Proterozoic are widely used to denote the older and younger Precambrian. The boundary between the two divisions is not yet settled but is placed by many authorities at 2,500,000,000 years. Local divisions have been established in each of the principal regions where Precambrian rocks are exposed.

General nature of the Precambrian record

The most extensive exposures of Precambrian are those that outcrop over roughly equidimensional areas many hundreds of kilometres across. Other Precambrian rocks are known to be concealed by younger deposits, as are those below the North American plain, the lowlands of Germany, and the region south and east of the Baltic. Many of the buried Precambrian rocks covered by flat-lying younger deposits have undergone little change since Precambrian times and can be readily identified by age determinations made on specimens obtained by deep drilling. A third type of occurrence is found in mountain chains formed since the Precambrian, where highly modified Precambrian rocks altered by high temperatures and pressures may be found within belts of metamorphic schists and gneisses. Examples occur in western Norway and in the Appalachians. Occurrences of the first type, in which Precambrian rocks form extensive outcrops at the Earth's surface, are known as continental shields.

The largest shields occur in Canada, Greenland, Scandinavia and adjoining parts of the U.S.S.R. (the Baltic Shield), Siberia, China, peninsular India, Guyana, Brazil, Antarctica, most of Africa south of the Atlas Mountains,

Adapted from "Time and Place in Orogeny," special publication no. 3 (1969); Geological Society of London

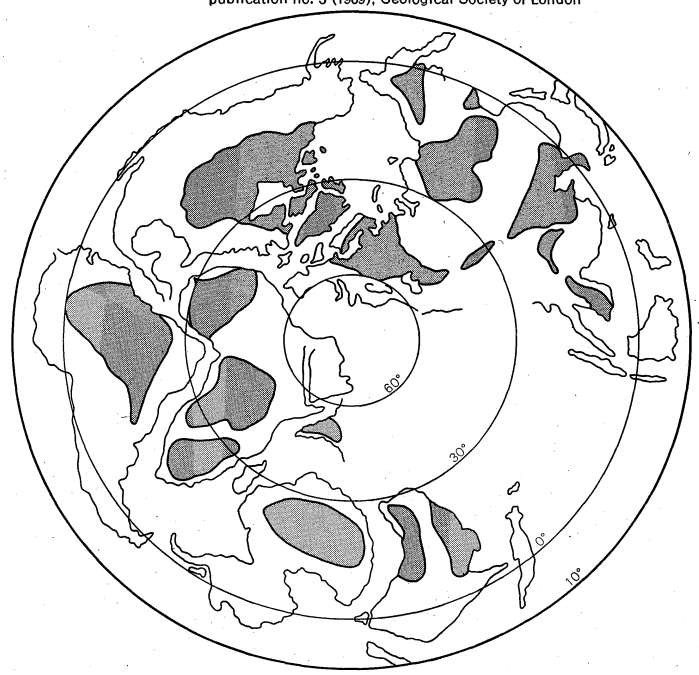


Figure 1: Precambrian shields and platforms that are more than 1,000,000,000 years of age.

Some Precambrian Events and Their Chronology

North America	British Isles	Africa	Australia	ages in 000,000 of years
Diabase dikes	Moinian metamorphism, deposition of Torridonian		Moonlight Valley Glaciation	800
Grenville plutonism		pegmatites, Orange River	pegmatites, Western Australia	1,000
Duluth Gabbro, Keweenawan Lava			pegmatites, Fraser Range, Western Australia	1,200
		Karagwe/ Ankolian plutonism		1,400
	Late Laxfordian Granites			1,600
		Ubendian plutonism		
Start of Hudsonian plutonism	start of Laxfordian plutonism		Halls Creek Granites	1,800
Diabase dikes	Scourie Diabase Dykes	end of Limpopo Belt metamorphism		2,000
	Late Scourian Pegmatites			2,200
		Great Dyke	diabase dikes, Western Australia	2,400
		Shamvaian plutonism	close of formation of Yigara Block	2,600
Kenoran plutonism	start of Scourian plutonism	Bulawayan plutonism		2,800
Keewatin Greenstones		Dominion Reef	granites, Pilbara, Western Australia	3,000
Early Basement, Montana, Ontario		metamorphism of Barberton Greenstone Belt		3,200
		Swaziland Older Granite		3,400
				3,600

and in Western Australia (see Figure 1). The term platform can be applied to a mass of Precambrian rock overlain by a cover of flat-lying younger deposits. The Russian and North American platforms, which are in effect buried extensions of the Baltic and Canadian shields, are examples. It follows from the wide distribution of the three types of Precambrian occurrences, the shields, the platforms, and the belts of altered Precambrian rocks found in younger metamorphic chains, that much of the continental crust of the Earth dates from Precambrian times. In contrast, Precambrian rocks are absent from the floors of major oceans except on such islands as the Seychelles, which can be regarded as fragments of a continent broken up during continental drift (*q.v.*). The abundance of Precambrian rock is one feature that distinguishes the crust of the continents and continental shelves from the crust below the deep oceans.

It was formerly supposed that each continent had grown around a central nucleus of Precambrian rock, about which successively younger belts of continental crust had accreted. Although it is true that regions such as the Mississippi or Niger deltas represent geologically recent additions to older land masses and that mountain chains of Paleozoic and younger ages tend to be peripheral to shield areas, the hypothesis that continents grow by the addition of a succession of younger rocks is somewhat inconsistent with the known distribution of Precambrian formations. These do not get systematically younger toward the margins of continents but form an irregular network of rocks of varying age.

On a worldwide scale the distribution of Precambrian outcrops can be seen to follow certain rules. With few exceptions, the Precambrian rocks that outcrop near the shores of the Pacific and along the Alps and Himalayas form parts of mountain chains. Any Precambrian rock in these regions is likely to be highly altered. On continents flanking the comparatively youthful Indian or Atlantic oceans, which have opened since Mesozoic times, the principal Precambrian outcrops occur in shields close to these young oceans. Thus, the Brazilian, Guyanaian, Canadian, Baltic, and West Africa shields border the Atlantic, and the East African, Indian, West Australian, and Antarctic shields lie around the Indian Ocean. Platforms of Precambrian rock concealed below flat sediments usually lie between the shields and the younger Alpine or circum-Pacific mountain chains with their highly modified Precambrian remnants.

Geological significance

Taken as a whole, the Precambrian rocks available for study do not provide a continuous record of Precambrian time. They suffice to demonstrate that processes of erosion and sedimentation similar to those that operate at the present day were affecting the crust at least as far back in time as 3,200,000,000 years ago. At that time, primitive forms of life such as the blue-green algae and certain bacteria had already evolved. The nature of the Precambrian rocks also shows that from time to time parts of the Earth's crust were heated and deformed, so that crystalline metamorphic rocks were produced. These rocks resemble those found in the deeper parts of more recent chains, but it is not yet clear whether the Precambrian crystalline rocks were connected with mountain building or whether the deformation of the crust took other forms during the early history of the Earth.

This article treats Precambrian rocks, life, and environments and the dating and chronology of Precambrian events. For further information on the place of the Precambrian in Earth history, see EARTH, GEOLOGICAL HISTORY OF; for additional detail on Precambrian life and environments, see FOSSIL RECORD; ATMOSPHERE, DEVELOPMENT OF; OCEANS, DEVELOPMENT OF; and PALEOGEOGRAPHY; see also MOUNTAIN-BUILDING PROCESSES; IGNEOUS ROCKS; and METAMORPHIC ROCKS for discussions of Earth processes and rock formation that are of relevance to the Precambrian record.

ABSOLUTE AGE DETERMINATION OF PRECAMBRIAN EVENTS

Isotopes and minerals involved. The isotopes (species of an element with the same atomic number but different atomic weight) that have proved useful for dating Precambrian material are those with a long half-life; in other words, those in which the time required for half of the original radioisotope to decay may be as long as several thousand million years. The decay of uranium and thorium to give lead and helium, the decay of the potassium isotope potassium-40, and the decay of the rubidium isotope rubidium-87 provide suitable data. The methods require that the decay constant of the radioactive nuclide be accurately known and that a proper correction be made for

Radiometric age and the thermal history of rocks

any radiogenic nuclide incorporated in the rock or mineral when it was formed or subsequently added or lost through processes other than radioactive decay. Radiometric measurements do not necessarily establish the age of a rock or mineral, but they do determine the time that has passed since the material last cooled to the temperatures below which the products of radioactive decay remain trapped within the mineral lattice.

Many crystalline (igneous and metamorphic) rocks have been repeatedly heated through their geological history, and if for this reason the products of the decay of an unstable isotope have been dissipated, accurate age determinations of any but the latest period of heating may be impossible. By a suitable combination of methods, however, these difficulties can be overcome. Sedimentary rocks may be dated if they contain a mineral, formed at the time of deposition of the sediment, that includes within its lattice a suitable radioactive isotope. In general, however, sedimentary rocks are dated by indirect methods, as, for example, when the age of a lava flow within a succession of sediments gives the approximate age of adjacent strata (see DATING, RELATIVE AND ABSOLUTE).

The two most widely used dating methods are those utilizing the decay potassium-40 and rubidium-87. The first method depends on the decay of potassium-40 to argon-40. The decay of potassium-40 also gives rise to calcium-40. The rubidium–strontium method is based on the decay of rubidium-87 to strontium-87. The precise value of the half-life of this decay system—namely, the time required for half the parent to break down—is not readily measurable. Despite this uncertainty, the method is particularly useful for determining the early history of rocks and minerals that may have been subsequently reheated. Any potassium-bearing minerals, such as the feldspars, micas, or glauconite, can be used for potassium–argon age determinations. Micas and potassium feldspars are among the most suitable minerals for rubidium–strontium dating. Because feldspars and micas are widespread, these methods can be applied to most varieties of crystalline rock.

Uranium- and thorium-bearing minerals, though much rarer, provide several methods of determining geological age utilizing the various radioactive isotopes that may be present. Comparison of the amounts of lead isotopes present in lead-bearing minerals can also be used in certain situations to determine the passage of time.

Distribution and significance of absolute dates. The dating of Precambrian rocks has shown that certain types of geological structure characterize different stages in the Precambrian evolution of the Earth. Very few existing rocks and minerals date from the first thousand million years of the history of the Earth (between 4,600,000,000 and 3,500,000,000 years ago). It would appear that most rocks formed during that span of time subsequently were destroyed or so modified that the time of their original formation cannot be determined by radiometric measurement. Well-authenticated records of rocks dated at more than 3,600,000,000 years in age are exceedingly rare. The oldest recognizable sedimentary and volcanic rocks range from about 3,800,000 years onward. It is probable that still older rocks were present, but this cannot be verified by radiometric age determinations. Sedimentary and volcanic rocks formed on the Earth's surface between about 3,800,000 and 2,700,000,000 years ago are always deformed and metamorphosed. The oldest unmetamorphosed flat-lying successions of sedimentary and volcanic rocks are those found in the lower parts of the Witwatersrand Basin in South Africa (see Figure 2). The development of this sedimentary basin, initiated about 2,700,000,000 years ago, appears to have marked a new stage in the evolution of the Earth, for it demonstrates that a section of the crust had become sufficiently rigid and cool for sedimentary rocks to accumulate and survive with little alteration.

The Earth's oldest rocks and early crustal history

Increasing rigidity within the crust also is suggested by large numbers of diabase dikes (intrusive rocks that penetrate pre-existing rocks and structures), which filled fracture systems in the Earth's crust that extend for many

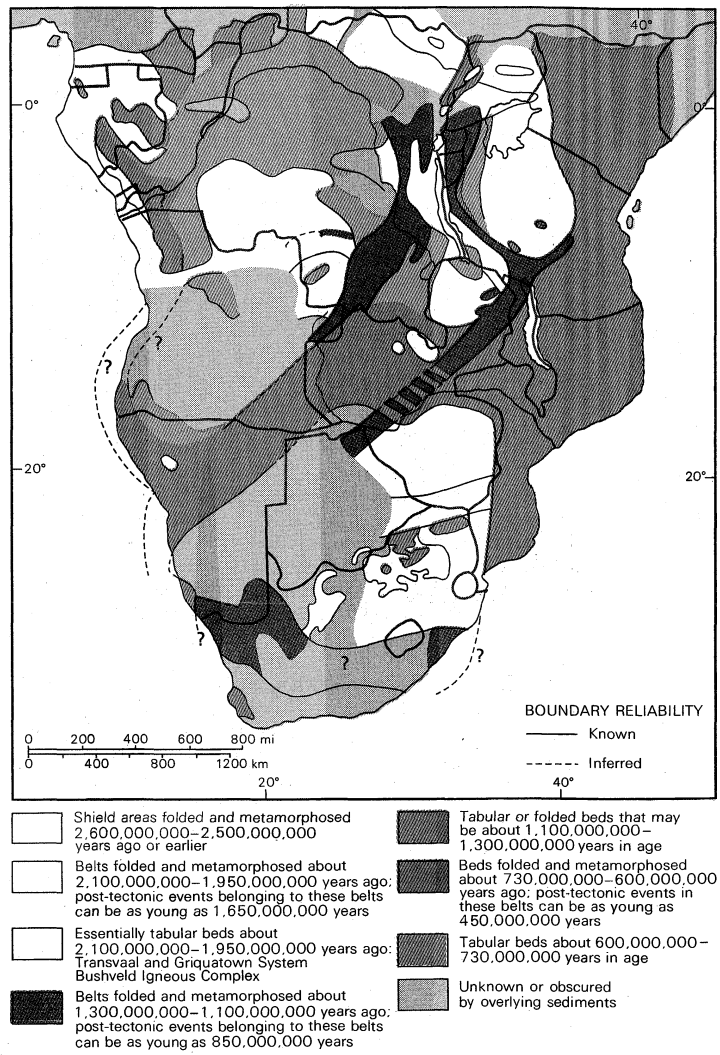


Figure 2: Distribution of Precambrian rocks in central and southern Africa.

Adapted from L. Cahen and N. Snelling, *The Geomorphology of Equatorial Africa* (1966); North-Holland Publishing Company, Amsterdam

hundreds of kilometres. Some of the earliest Precambrian dike systems date from between 2,400,000,000 and 2,500,000,000 years ago, as for example, those in the Transvaal of South Africa and in Western Australia. The Great Dyke of Rhodesia, which extends for several hundred kilometres, is 2,500,000,000 years old, as are somewhat similar structures in the western part of Australia. The development of these dikes and the survival of unaltered deposits in sedimentary basins show that by that time the Earth had attained a condition in which certain rigid areas were marked off from other regions of greater mobility. A mobile crust, in this connection, means a crust in which vertical and horizontal movements have occurred, in which melting or partial melting temperatures of rocks were reached, and in which all rocks show, to a greater or lesser degree, evidence of recrystallization or deformation.

It is not possible to distinguish individual mobile belts in the Early Precambrian because the entire crust was mobile, but once rigid areas had developed, the Earth's crust at any later point in geological time can be regarded as composed of stable rigid blocks set in a network of mobile belts.

Recognition of Archeozoic and Proterozoic events

The subdivision of Precambrian time into an older Archeozoic and younger Proterozoic Age was accomplished by recognition of the change from a wholly mobile crust to one composed of stable blocks bounded by mobile belts. The exact age of this change, which may not have occurred simultaneously in different parts of the crust, is not precisely known. It appears to have occurred about

2,500,000,000 years ago, which can be taken provisionally as marking the end of Archeozoic time and the start of the Proterozoic. In Late Precambrian times, about 1,000,000,000 years ago, another major change occurred in the structural behaviour of the Earth when a new system of mobile belts and rigid blocks superseded the system that had been established early in the Proterozoic. This Late Precambrian system has continued to influence the structural development of the Earth until the present time. Many Late Precambrian formations laid down since the system of structures was established are more closely linked with the younger Phanerozoic deposits that overlie them than with any older Precambrian rocks. These comparatively young Precambrian deposits, which lie for the most part outside the Precambrian shields, tend to foreshadow the arrangement of sedimentary basins and troughs that emerged in the Early Paleozoic.

PRECAMBRIAN ROCKS

As previously mentioned, Precambrian rocks outcrop over large areas to form continental shields, occur below flat-lying sediments, where they constitute continental platforms, and exist as more or less modified remnants within Phanerozoic (post-Precambrian) mountain chains.

Continental shields. Most of the knowledge of the older Precambrian comes from the study of the major shields, which are largely composed of rocks of Archeozoic and Early and Middle Proterozoic age. The shields are regions of the Earth's crust that have undergone little modification since Late Precambrian time. Many of the youngest Precambrian rocks, dating from the last 400,000,000 years of Precambrian time, occur outside the shields in close proximity to Phanerozoic rocks, into which they may pass without conspicuous unconformity. Many such rocks were later incorporated into Phanerozoic mountain belts, but some have survived as unaltered flat-lying formations such as the Scottish Torridonian. Good examples of folded Late Precambrian deposits are provided by the Adelaidian of southern Australia, the Sparagmites of Scandinavia, and the Late Precambrian rocks of the western Congo. The great majority of the Precambrian rocks, however, occur either within the shields or within their buried extensions on the continental platforms.

The Canadian Shield

The largest of the principal shields is the Canadian Shield, which extends for nearly 8,000,000 square kilometres (3,000,000 square miles) over eastern, central, and northwestern Canada from the Great Lakes to the Canadian Arctic and into Greenland. This constitutes the largest mass of exposed Precambrian rock in the world, and it is known to continue into the United States as basement, largely concealed by the more recent deposits of the North American plains.

It has been established that the Canadian Shield, like other shields, can be divided into several structural provinces, each of which evolved during a restricted period of Precambrian time. These provinces fall into three age groups. The oldest formed more than 2,500,000,000 years ago, the second between 1,800,000,000 and 1,400,000,000 years ago, and the youngest less than 1,200,000,000 years ago. Each province records a long history of sedimentation and erosion, volcanic activity, deformation, and recrystallization of igneous and metamorphic rocks.

The younger provinces contain relics of older rocks, and their presence demonstrates that the oldest structural provinces of the Canadian Shield were once more extensive and may have covered much of the region where younger Precambrian provinces have subsequently developed. Studies of the Canadian Shield have revealed a succession of changes in the behaviour of the Earth's crust during Precambrian time. The oldest provinces within the Canadian Shield result from the Kenoran mountain-building events, which ended about 2,500,000,000 years ago. The three principal Kenoran provinces are the Superior, or Keewatin, Province, which lies east, south, and southwest of Hudson Bay; the Slave Province in the northwest of the Canadian Shield; and the Atlantic Province. This last Kenoran Province extends from the Labrador coast eastward across Greenland to terminate on the continental shelves of the Atlantic, between Greenland and Scotland.

These three provinces contain belts of sedimentary and volcanic rocks many kilometres in thickness. They rest on a basement of older crystalline rocks, parts of which are at least 3,000,000,000 years in age. It is probable that this basement contains still older rock that may represent a highly modified portion of the Early Precambrian crust. The Kenoran volcanic and sedimentary rocks typically form irregular belts that may attain 400 kilometres (250 miles) in length. Although these rocks have been folded and metamorphosed to some extent, many original features are usually preserved. Long successions of pillow lavas as much as several kilometres in thickness, for example, may retain structures that formed when lava was rapidly chilled by eruption into water. Evidence of subaqueous volcanism is widespread, and it would appear that these belts represent regions of the crust that subsided as much as ten to 15 kilometres (six to nine miles) as volcanic rocks and sediments accumulated. The belts, which are often known as greenstone belts because of the appearance of the slightly altered volcanic rocks, have no exact equivalent among contemporary active volcanic regions, although they show some resemblances to present-day volcanic island arcs (*q.v.*).

Greenstone belts

In Canada, as in other parts of the world, gold mineralization is associated with the greenstone belts. The rocks around the greenstone belts are often predominantly granitic. It is probable that the tight folding that has been referred to is at least in part connected with the development of the roughly circular granite masses, which are typical of this association. Age determinations in other parts of the world have shown that the development of greenstone belts and granitic rocks characterizes many other areas of equally old crust (over 2,500,000,000 years in age). Comparable associations, for example, occur in Western Australia, southern India, and in central and southern Africa. They represent one aspect of the highly mobile crust that characterized Early Precambrian times.

The second group of provinces of the Canadian Shield contains structures developed between 1,800,000,000 and 1,400,000,000 years ago, in a complex of rocks composed partly of Kenoran rock more than 2,500,000,000 years in age and partly of younger successions, formed between about 2,300,000,000 and 1,800,000,000 years ago. Much of this Early Proterozoic, or Huronian, succession was involved in a series of metamorphic events that developed in a system of mobile belts. Within these belts both the old Archeozoic rocks and the Proterozoic cover were folded and metamorphosed so that distinctions between the two groups often are difficult to make. In contrast, regions outside these mobile belts were not so affected and in such regions unmetamorphosed Proterozoic rock can be identified resting unconformably upon Archeozoic basement.

The Proterozoic mobile belts contrast with greenstone belts of the older structural provinces in several ways. They are very much longer, extending for thousands of kilometres as a network of belts that still range across the Precambrian crust as it survives today. These networks probably were worldwide in extent when they originally formed. Another contrast is the fact that Proterozoic mobile belts were bounded by rigid areas of crust, whereas in Archeozoic times no such stable regions of crust appear to have been present.

Overview of rock types. There are few, if any, types of rock that are restricted to Precambrian formations. Despite this, certain kinds of rocks occur more abundantly in Precambrian terrains than in any younger groups. Furthermore, certain combinations of rocks are found only in the Precambrian. The uniqueness of some Precambrian terrains, therefore, lies in the extent to which certain associations of rock are present rather than in the existence of distinctive rock types.

Many Archeozoic provinces consist of strongly folded, steeply dipping volcanic successions compressed between granite bodies. These volcanic rocks characteristically occur in sequences that begin with ultramafic (rich in iron

and magnesium minerals) varieties and pass with time through mafic to felsic (granitic-type rock, relatively rich in potash and soda) rocks. Such sequences may be repeated several times within a single greenstone belt. They form a most distinctive feature of Early Archeozoic times.

Anorthosites, charnockites, and glaucophane schists

The igneous rock anorthosite, composed largely of plagioclase feldspar, is not confined to the Precambrian but was developed on a larger scale in the Precambrian than in any subsequent interval of geological time. Anorthosites older than 3,000,000,000 years of age have been found on the Moon (*q.v.*). Rocks of the Charnockite Series, which are distinguished by the presence of hypersthene and by the dark colour of the feldspars and quartz, were widely developed by crystallization from melts and by metamorphism of older rocks. Conversely, certain types of metamorphic rock that are common in the Phanerozoic, such as the glaucophane schists, have not yet been recorded from any region of Precambrian age. Because glaucophane schists are widespread in Mesozoic and Tertiary terrains, their absence from the Precambrian, taken in conjunction with the abundance of charnockitic rocks, suggests a change in the conditions under which regional metamorphism occurs within the crust. These facts are consistent with a higher heat flow during the Precambrian than at the present day. Most Precambrian shield terrains contain abundant gneisses and schists that do not differ significantly from their younger counterparts. These rocks have developed on an extensive scale, however—one which cannot be matched in more recent post-Cambrian metamorphic belts. This also may indicate a high flow of heat during the Early Precambrian (see EARTH, HEAT FLOW IN).

Precambrian sediments show several distinctive features. Chemical deposits such as carbonates, banded iron formations, and cherts are conspicuous. The earliest iron formations date from the Archeozoic, but there was a remarkable increase in the abundance of banded iron formations in Early Proterozoic times, about 2,000,000,000 years ago. These deposits provide important sources of iron ore in North America, around the Great Lakes and in Labrador; in the Hamersley region of Western Australia; and in Mauritania, in northwest Africa. Some of these iron deposits owe their origin to the action of bacteria or blue-green algae that promote precipitation of iron and silica. Hence, they can be regarded as biochemical deposits. Precambrian cherts were widely developed, usually as thin bands no more than a few centimetres in thickness, though locally, as in the Hamersley Basin, in much thicker sequences. Organic remains recently have been found in many such bands, the oldest of which were deposited more than 3,200,000,000 years ago and contain the oldest fossils yet discovered on Earth (Barberton area in South Africa).

Such Early Precambrian sediments are closely associated with volcanic rocks in the unstable belts that characterize much of the Archeozoic. In addition to the chemical and biochemical cherts, and carbonates and ironstones, thicker successions of graywackes and conglomerates derived by rapid erosion are present. Some were derived from the volcanic successions, whereas others represent the products of erosion of granites. The extensive ironstones of the Early Proterozoic lie within basins and troughs that are the earliest forerunners of the more recent geosynclines. In South Africa, a quartz-bearing conglomerate, or banket, of the Witwatersrand Basin is the host rock of the world's largest deposits of gold. The banket occurs in a sedimentary basin dating from Late Archeozoic and the Early Proterozoic.

Graywackes and conglomerates

In Late Precambrian time, from about 1,000,000,000 years onward, extensive troughs filled with clastic sediments (consisting of silt, sand, and larger particles) developed in many parts of the world. The deposits in these troughs mark the start of sedimentation that continued into the Phanerozoic. They provided much of the material to be subsequently folded and metamorphosed during periods of Paleozoic mountain building.

PRECAMBRIAN LIFE

During most of the 19th century, it was supposed that no life existed in Precambrian times and that the oldest fossils appeared at the base of the Cambrian (about 570,000,000 years ago). Fossil bacteria and blue-green algae are now known to occur in the oldest sedimentary rocks so far recognized, however. The age of these rocks exceeds 3,200,000,000 years. This occurrence raises two possibilities: either life evolved on Earth in some still older period of Precambrian time, in which case it is unlikely that sedimentary rocks old enough to record the first steps in this evolution have survived, or life came to Earth from some extraterrestrial source, and evidence as to the earliest forms of life will have to be sought outside the Earth.

Development of multicellular life forms. Over most of the Precambrian the predominant forms of life were microscopic unicellular organisms, and only in the late Precambrian, about 800,000,000 years ago, did a great variety of multicellular, or metazoan, forms appear. It is of considerable interest that Early Precambrian blue-green algae closely resemble forms living today, which demonstrates the extreme stability of these forms of life. The living organism *Kakabekia barghoorniana*, for example, isolated from soil at the base of Harlech Castle on the Welsh coast, is a form of an organism preserved in the Gunflint Chert of North America deposited about 2,000,000,000 years ago. This stability provides a possible clue to the contrasting fossil record of Precambrian and later times. The microfossils found in Precambrian rock more than 3,000,000,000 years old appear to be derived from procaryotic organisms, forms of life, principally represented by the bacteria and blue-green algae, that lack a distinct nucleus enveloped in a membrane and that are unable to reproduce by division and recombination of such a nucleus.

The Beok Spring Dolomite of eastern California, formed between 1,400,000,000 and 1,200,000,000 years ago, and the Bitter Springs Formation in Australia, deposited about 1,100,000,000 years ago, contain the oldest examples of the Eucaryota, more advanced organisms than the Procaryota, which contain a nucleus and can reproduce sexually. In these formations the Eucaryota are represented by green algae. The exact date of the evolution of the first Eucaryota is not known. The earliest multicellular, or metazoan, organisms are preserved in rocks about 800,000,000 years old.

Oldest eucaryotic organisms

After the appearance of the metazoans, there was a rapid burst of evolution, and by the end of the Precambrian a great variety of life forms had evolved; by the beginning of the Cambrian Period many of the existing invertebrate groups had come into being. Single-celled organisms without a nucleus evolved 3,200,000,000 years ago, and single-celled nucleate organisms capable of sexual reproduction appeared between 1,200,000,000 and 1,400,000,000 years ago. Multicellular animals were present about 800,000,000 years ago. All these dates should be regarded as minima likely to be extended farther back in Precambrian time by future research.

Fossiliferous deposits. Bacteria, blue-green algae, and other procaryotic organisms, as well as primitive fungi, are preserved in cherts, banded iron formations, shales, and limestones dating from the Archeozoic. The oldest examples, from the Barberton region in South Africa, exceed 3,200,000,000 years in age. Some of the best preserved fossils occur in cherts. The bacteria-like organisms are less than a micron in diameter, but the algae are larger. Geochemical investigations have shown an enrichment in the carbon-12 isotope, and this can be taken as an indication of organic activity. The presence of phytin and pristine indicates that photosynthesis took place. Mats of blue-green algae can trap carbonate sediments to build large structures known as stromatolites. Many Precambrian limestones and dolomites contain stromatolites, the oldest of which, in Rhodesia, are some 2,700,000,000 years in age.

The early Proterozoic deposits, about 2,000,000,000 years old, contain a number of fossiliferous cherts, of which the Gunflint Iron Formation of North America is probably the best known (Figure 3). This formation, including shales, limestones, and cherts, together with iron deposits, contains a variety of microfossils. Most are

Gunflint Chert and other Proterozoic deposits



Figure 3: Fossil of rod-shaped bacteria dating from Precambrian time, found in the Gunflint Chert, a sedimentary rock type in southern Ontario. Electron micrograph (magnified about 7,900 ×).
By courtesy of Morton D. Maser

closely similar to modern blue-green algae, though some of the filamentous forms resemble fungi. The Proterozoic Bitter Springs Formation of central Australia contains more advanced microfossils, derived from eucaryotid forms. These resemble present-day green algae, and it may be from these that the multicellular organisms of the Late Precambrian subsequently developed. Examples are preserved in South Australia, at several locations in the U.S.S.R., and as isolated forms such as the fossil *Charnia* in the Late Precambrian of Leicestershire.

PRECAMBRIAN ENVIRONMENTS

Paleogeography. The distribution of land and seas over much of Precambrian time is not known. During the Archeozoic the sediments and volcanic rocks of the greenstone belts were accumulating underwater. This is demonstrated by the nature of the abundant graywackes and by the widespread development of pillow lavas. At the present time they cannot be dated precisely, and all that is known is that a large number of such belts formed on the crust between about 3,400,000,000 and 2,500,000,000 years ago. In Canada, southern Africa, and Western Australia the remnants of such belts form roughly parallel systems in which each belt is separated from its neighbour by an irregular mass of granitic rock, perhaps derived partly from still older crust. This suggests that for unknown lengths of Archeozoic time, a number of belts of continental crust were probably submerged below the sea, and less probably below freshwater lakes.

Sedimentary basins

From about 2,700,000,000 years onward, individual sedimentary basins were formed. These basins have undergone little subsequent deformation and their original shape can therefore be discerned at the present time.

The Witwatersrand Basin of southern Africa is one of the oldest of these basins. The nature of the detritus suggests that sediments within this basin were derived from the erosion of older greenstone belts and granitic rocks. The gold in the conglomerates may well have been produced by the erosion of gold-bearing volcanics. The nature of the banded ironstone formations that occur in the upper part of the Witwatersrand Basin and in regions such as Western Australia indicates accumulation of chemical or biochemical deposits under conditions of extreme crustal stability. Other ironstone deposits, also dating from Early Proterozoic times, about 2,000,000,000 years ago, contain oolites and other structures that are indicative of shallow-water deposition. These deposits mark some of the oldest shorelines recognized in Earth history. Examples of such shorelines are present in Labrador and in Karelia, near the borders of Finland and the U.S.S.R. In these two regions shoreline deposits can be traced laterally into thicker sedimentary successions that appear to mark the sites of Early Proterozoic sedimentary troughs. Such troughs occur in eastern Canada along the coast of Labrador, in southern Greenland, and over much of Sweden and Finland. It is not yet possible to map these deposits on a worldwide scale. In Late Precambrian time, sequences of rocks up to ten or 15 kilometres (six or nine miles) in thickness accumulated in a network of troughs, many of which still are well preserved. In some areas it has been possible to identify the source areas from which these Late Precambrian deposits were derived. The Late Precambrian Torridonian rocks of Scotland, for example, formed between about 900,000,000 and 750,000,000 years ago and originated from the erosion of crystalline rocks to the west of the present coast of Scotland. The Late Precambrian is the earliest period for which some generalized distribution of sedimentary troughs and uplifted source areas can be established.

In general, it appears that a network of troughs extended over many thousands of kilometres but were no more than 100 or 200 kilometres (60 or 120 miles) in width. These troughs overlay subsiding continental crust and surrounded highland regions, possibly many hundreds of kilometres in diameter, from which detritus reached the sedimentary basins. Such basins occur in western Norway, northwestern Ireland and Scotland, Newfoundland, eastern Greenland, Spitsbergen, the Urals, eastern Australia, the western Congo, southwestern Africa, and in extensive regions in China.

Mountain building and deformation. Since the end of the Precambrian, a succession of mountain chains has formed. The oldest of these have been largely removed by erosion (Figure 4), but it is possible to identify their deep-

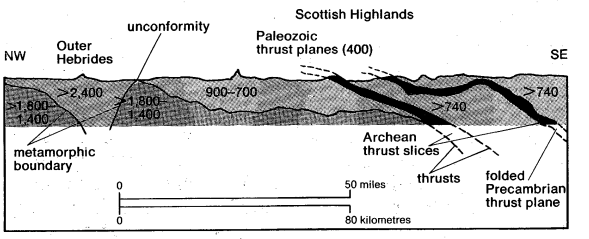


Figure 4: Geological section through Precambrian rocks of northwest Scotland, showing the remnants of an ancient mountain system. Numbers shown designate millions of years.

ly dissected roots, in which folded and metamorphosed rock were produced during mountain-building processes (*q.v.*). Somewhat similar metamorphosed and deformed rocks are widespread in Precambrian terrains, and this suggests that these also represent the roots of mountains, although no chains of Precambrian age survive as mountains at the present time. The possibility that Precambrian folded and metamorphic rocks developed in connection with processes other than mountain building cannot be eliminated, and perhaps the term orogenic should not be applied to the Precambrian belts of metamorphic and igneous rock until it can be demonstrated that these were in fact the roots of true mountain chains.

Geosynclinal deposition and deformation

The oldest belts of metamorphic and deformed rock are the greenstone belts, most of which date from between 3,400,000,000 and 2,500,000,000 years ago. The infilling of volcanic rock of many kilometres in thickness formed below sea level, and of shallow water sediments, indicates long-continued subsidence in such belts. The steep dips, thrusts, and folds show that intense deformation occurred. In some belts, such as those in Rhodesia (Figure 2), the presence of coarse detritus suggestive of rapid deposition indicates the possible occurrence of highlands subject to denudation. Generally, however, the surface relief of such Archeozoic structures is unknown.

Early in the Proterozoic, about 2,300,000,000 years ago, great sediment-filled troughs began to form. The

rocks of the Labrador trough of eastern Canada and the sediments of Sweden and Finland are examples of such deposits. In due course these sediments were folded and deformed, and it is possibly at this stage in the history of the Earth that a global system of mountain chains comparable to that of today was first developed. Certainly these Proterozoic structures were very much more extensive than the limited Archeozoic greenstone belts. To a large extent, regions where sediments accumulated in Early Proterozoic times continued as mobile belts, regions of unusual crustal movement, for many hundreds of millions of years. About 1,000,000,000 years ago a new system of mobile belts was developed; much of this system persisted into Paleozoic and more recent time to give the worldwide system of mountain chains that formed during the Caledonian and Hercynian orogenies.

Igneous activity

Extensive areas of igneous rocks were formed in each of these Precambrian periods of activity; that is, first, during and after the deformation of the Archeozoic greenstone belts; second, within the Early Proterozoic sedimentary troughs; and, finally, in the Late Precambrian troughs. It is uncertain whether this threefold structural division of the Precambrian based largely on data from North America and northern Europe can usefully be applied to other Precambrian shields.

Climatic conditions. Climatic conditions during most of Precambrian time are largely unknown. In Late Precambrian time, about 700,000,000 years ago, however, extensive glaciation is known to have occurred. Glacial deposits of Late Precambrian age are preserved in the Antarctic, in southern Australia, in southern and western Africa, in northwestern Europe, in Greenland and certain other Arctic islands, and in China. In some of these regions, glaciation was repeated. For this reason, the correlation of widely separated glacial deposits is not always clear. In general terms, however, it can be safely assumed that some time before the end of the Precambrian a widespread glaciation affected areas that now lie in both the Northern and Southern hemispheres. The suggestion has been made that this Late Precambrian Ice Age could be used to define the base of another geological period, which would extend from the time of these glaciations to the start of the Cambrian.

CORRELATION OF PRECAMBRIAN EVENTS

Problems of correlation. Despite the discovery of many fossiliferous formations, the rate of evolution throughout most of Precambrian time appears to have been so slow that it is unlikely that fossils can be used to establish the age of any but the youngest Precambrian. Similarly, the study of magnetic reversals during the Precambrian has not reached the stage where these can be employed to correlate rocks of this age. Widespread evidence of climatic change such as glaciation may be employed to correlate widely separated rocks, however. The existence of Late Precambrian glaciation is now well documented. It seems probable, though, that there was more than one Precambrian glaciation, and this promising method of subdividing the very latest Precambrian rocks may require further refinement before it can be adopted. In the near future the Late Precambrian glaciation and the Late Precambrian fossil assemblages, now known from more than a dozen localities, will provide a means of classifying the latest Precambrian rocks. It seems unlikely, however, that classical stratigraphic methods can be applied to rocks much older than 800,000,000 years in age. The problem of subdividing the main mass of Precambrian rocks will have to be dealt with on some other basis.

Radiometric data and correlation

Since 1960 the numbers of radiometric measurements on Precambrian rocks and minerals have been sufficient to establish local successions and to indicate the possible correlation of events on a worldwide basis. The data available at present, however, are insufficient to establish anything comparable to the detailed stratigraphical subdivisions that geologists have developed to mark the passage of Phanerozoic time. There are several reasons for this, but most important is the fact that sedimentary rocks seldom can be dated from the time of the formation of their contained sediments, and most age determinations on crystalline rocks indicate the time that has elapsed since the last time of cooling. It is often possible to correlate this cooling with some other geological event and thus determine the time when a region of metamorphic rock was deformed or was intruded by igneous rocks. In this way, a chronology of isolated geological events has been built up for many Precambrian regions, but from its very nature it is not a continuous record. It would appear that for a very long time to come, Precambrian time is likely to be measured by the dating of a number of distinct geological events and that no continuous stratigraphical record will generally be available. A number of sedimentary formations in the Precambrian can be placed within certain limits of time, although they cannot be precisely dated. It is often possible to show that a deposit must be younger than the youngest crystalline rock on which it rests and older than the oldest rock that intrudes it; but in many instances a very considerable uncertainty remains, and this may reflect several hundred million years. Intercontinental correlations from one sedimentary basin to another are hardly possible at present. On the other hand, intercontinental correlations of metamorphic and magmatic events can be made. It has been demonstrated, for example, that 2,000,000,000 years ago large numbers of diabase dikes (tabular igneous masses that intrude pre-existing rocks) were formed in many parts of North America, Greenland, and northwestern Europe. It can also be shown that certain widespread metamorphic events recorded in more than one continent came to an end over large regions at times that can be quite closely defined. The uncertainty may be of the order of 1 percent of the total time elapsed. It follows from this that the most extensive correlations are those concerned with the ending of igneous or metamorphic activity.

The potassium-bearing mineral glauconite has been used to date Late Precambrian deposits with some success, particularly in the Soviet Union. In any area where volcanism occurred within a sedimentary basin, it may be possible to date volcanic rocks and so decrease the range of uncertainty as to the age of adjoining sediments. The problems of intercontinental correlation of Precambrian structures have been greatly simplified by the recognition of the manner in which the continents fitted together before the last episode of continental drift. If concern is with earlier Precambrian time, however, the possibility that there may have been other periods in the Earth's history when the relative position of continents changed rapidly must be considered. Knowledge of Precambrian historical geology has not yet reached the stage where the effects of possible Precambrian continental drift can be assessed.

Summary of Precambrian events. The following general statement about the Precambrian chronology can be made (see Table). It is consistent with existing knowledge but will undoubtedly be modified as new information becomes available:

1. Formation of the Earth's crust about 4,600,000,000 years ago.
2. Modification of the original crust by the intrusion of igneous rocks, the extrusion of lavas, metamorphism, deformation, and erosion. At an unknown time between 4,600,000,000 and 3,500,000,000 years ago, the earliest sedimentary rocks were formed, almost certainly by the erosion of then-existing igneous material. During this period life either evolved on Earth or reached the Earth from some extraterrestrial source.
3. Between 3,500,000,000 and 2,500,000,000 years, the formation of a succession of greenstone belts occurred. Sediments and volcanic rocks accumulated in these zones and were subsequently deformed and metamorphosed. The belts lie on a crust of older rock that was itself invaded by many granites during the time when the greenstone belts were evolved. All greenstone belts did not form at the same time, but their chronology is not yet known in detail.
4. From 2,700,000,000 years onward the first stable areas to survive until the present day were developed on the continental crust. Flat-lying sedimentary deposits and un-

deformed diabase dikes survive within such regions; the Witwatersrand Basin of South Africa is an example. Important features dating from this period include the earliest stable sedimentary basins, formed about 2,700,000,000 years ago, and the earliest diabase swarms, intruded between 2,500,000,000 and 2,400,000,000 years ago. Extensive diabase swarms approximately 2,000,000,000 years old are known in northwestern Canada, eastern Canada, Greenland, and northwestern Europe.

5. The accumulation of Early Proterozoic sedimentary deposits marks the development of a new system of sedimentary troughs and basins. These date from approximately 2,000,000,000 years ago. Many banded ironstone formations, such as those in North America, South Africa, and Western Australia, date from this time.

6. From about 1,800,000,000 years onward, the Early Proterozoic sediments and the older Archeozoic rocks that underlie them were metamorphosed and deformed to produce new complexes of crystalline rock. Local names for such complexes include the Hudsonian of Canada, the Ketilidian of Greenland, the Laxfordian of Scotland, and the Svecofennid of Scandinavia.

7. Among the most important events toward the end of the Precambrian was the accumulation of the Late Precambrian deposits formed from about 1,000,000,000 years onward. These lie in a network of Precambrian troughs, many of which received further sediments during the Paleozoic and became involved in mountain building during periods of Paleozoic orogeny.

BIBLIOGRAPHY. General works that provide information on Earth history include: R. DEARNLEY, "Orogenic Fold-Belts and a Hypothesis of Earth Evolution," in *Physics and Chemistry of the Earth*, vol. 7 (1966); A. HOLMES, *Principles of Physical Geology*, 2nd ed. (1965); G. GASTIL, "Continents and Mobile Belts in the Light of Mineral Dating," *Proc. Int. Geol. Congr. 21st Sess.*, pt. 9 (1960); and T.M. CLIFFORD and I.B. GASS (ed.), *African Magmatism and Tectonics* (1970). Life of the Precambrian is well covered in the articles by: E.S. BARGHOORN and S.A. TYLER, "Microorganisms from the Gunflint Chert," *Science*, 147:563–577 (1965); and A.E.J. ENGEL *et al.*, "Alga-like Forms in the Onverwacht Series, South Africa: Oldest Recognized Life-like Forms on Earth," *Science*, 161:1005–1008 (1968). Regional coverage of Precambrian rocks is afforded by the several volumes of *The Geologic Systems*, ed. by K. RANKAMA (1963–70). See vol. 1 on the Precambrian of Denmark, Norway, Sweden, Finland; vol. 2 on the Precambrian of Spitsbergen, the British Isles, Greenland, Canada; vol. 3 on the Precambrian of India, Ceylon, the Seychelles, Madagascar, the Congo; and vol. 4 on the Precambrian of the southeast, south central, and northwest United States and Mexico. Other works of a regional nature include: L. CAHEN and N.J. SNELLING, *The Geochronology of Equatorial Africa* (1966); T.C.R. PULVERTAFT, "The Precambrian Stratigraphy of Eastern Greenland," *Proc. Int. Geol. Congr. 23rd Sess.*, sect. 4 (1968); and C.R. ANHAEUSER *et al.*, "The Barberton Mountain Land: A Model of the Elements and Evolution of an Archaean Fold Belt," *Trans. Geol. Soc. S. Afr.*, annexure to vol. 71, pp. 225–254 (1968).

(J.Su.)

Precipitation

Atmospheric precipitation includes all forms of water particles, whether solid or liquid, that fall from clouds and reach the ground. It thus includes drizzle, rain, snow, snow pellets, ice crystals, and hail. All forms of precipitation are produced as the result of the condensation of atmospheric moisture to form clouds in which some of the particles, by growth and aggregation, attain sufficient size to fall out of the cloud and reach the ground.

Precipitation is one of the three main processes (evaporation, condensation, and precipitation) that constitute the hydrologic cycle, the continual exchange of water between the atmosphere and the surface of the Earth. Water is evaporated from ocean, land, and freshwater surfaces, is carried aloft as vapour by the air currents, condenses to form clouds, and ultimately is returned to the Earth's surface as precipitation. The average global stock of water vapour in the atmosphere is equivalent to a layer of water 2.5 centimetres (one inch) deep covering the whole Earth. Because the Earth's average annual rainfall is about 100 centimetres (40 inches), the average time that the water spends in the atmosphere, between its evaporation from the surface and its return as precipitation, is about 1/40 of a year or about nine days. Of all the water vapour that is carried at all heights across a given region by the winds, only a few percent is converted into precipitation and reaches the ground in that area. In deep and extensive cloud systems, the conversion is more efficient, but even in thunderclouds the quantities of rain and hail released amount to only some 10 percent of the total moisture entering the storm.

Measurement of amount and intensity

In the measurement of precipitation, it is necessary to distinguish between the amount—defined as the depth of precipitation (calculated as though it were all rain) that has fallen at a given point during a specified interval of time—and the rate or intensity—which specifies the depth of water that has fallen at a point during a particular interval of time. Persistent moderate rain, for example, might fall at an average rate of five millimetres (0.2 inch) per hour and thus produce 120 millimetres, or nearly five inches, of rain in 24 hours. A thunderstorm might produce this total quantity of rain in 20 minutes, but at its peak intensity the rate of rainfall might become much greater—perhaps 120 millimetres per hour, or two millimetres (0.08 inch) per minute, for a minute or two.

The amount of precipitation falling during a fixed period is measured regularly at many thousands of places on the Earth's surface by rather simple rain gauges. Measurement of precipitation intensity requires a recording rain gauge, in which water falling into a collector of known surface area is continuously recorded on a moving chart or a magnetic tape. Investigations are being carried out on the feasibility of obtaining continuous measurements of rainfall over large catchment areas by means of radar.

Apart from the trifling contributions made by dew, frost, and rime and by desalination plants, the sole source of freshwater for sustaining rivers, lakes, and all life on Earth is provided by precipitation from clouds. Precipitation is therefore indispensable and overwhelmingly beneficial to mankind, but very heavy rainfall can cause great harm: soil erosion, landslides, and flooding; and hailstorm damage to crops, buildings, and livestock occasion great economic loss each year.

This article treats the general causes of precipitation, its several different types, and its world distribution and effects. For detailed information on snow see the article SNOW AND SNOWFLAKES, and for related information on weather and climate see CLIMATE; CLOUDS; THUNDERSTORMS; and WINDS AND STORMS. See also WEATHER MODIFICATION for coverage of fog, storm, and hail suppression, as well as cloud seeding; HYDROLOGIC CYCLE and WATER RESOURCES for treatment of precipitation within the global water budget; and SEDIMENT YIELD OF DRAINAGE SYSTEMS and RIVERS AND RIVER SYSTEMS for further consideration of precipitation effects.

ORIGIN OF PRECIPITATION IN CLOUDS

Cloud formation. Clouds are formed by the lifting of damp air, which cools by expansion as it encounters the lower pressures existing at higher levels in the atmosphere. The relative humidity increases until the air becomes saturated with water vapour, then condensation occurs on any of the aerosol particles suspended in the air. A wide variety of these exist in concentrations ranging from only a few per cubic centimetre in clean maritime air to perhaps 1,000,000 per cubic centimetre (16,000,000 per cubic inch) in the highly polluted air of an industrial city (see URBAN CLIMATES). For continuous condensation leading to the formation of cloud droplets, the air must be slightly supersaturated. Among the highly efficient condensation nuclei are sea-salt particles and the particles produced by combustion (*e.g.*, natural forest fires and man-made fires). Many of the larger condensation nuclei over land consist of ammonium sulfate. These are produced by cloud and fog droplets absorbing sulfur dioxide and ammonia from the air. Condensation onto the nuclei continues as rapidly as water vapour is made available through cooling; droplets about 0.01 mil-

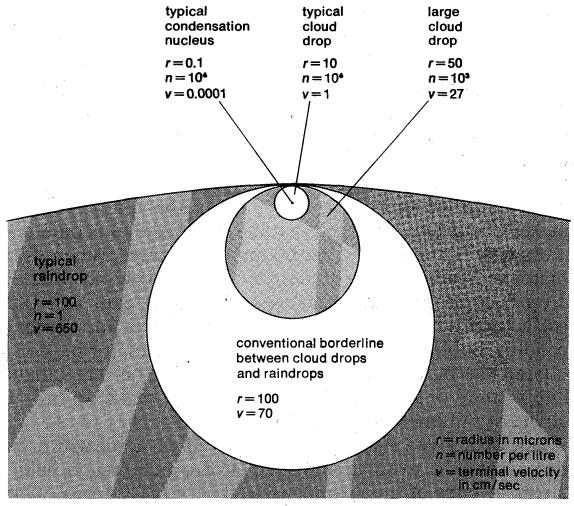


Figure 1: Comparative sizes, concentrations, and falling velocities of some particles involved in the condensation and precipitation process (see text).
Adapted from H.E. Landsberg and J. Van Mieghem, *Advances in Geophysics*, vol. 5 (1958); Academic Press Inc.

limetre, or ten microns (0.0004 inch), in diameter are produced in this manner. These droplets constitute a nonprecipitating cloud.

Mechanisms of precipitation release. Growing clouds are sustained by upward air currents, which may vary in strength from a few centimetres per second to several metres per second. Considerable growth of the cloud droplets (with falling speeds of only about one centimetre [0.4 inch] per second) is therefore necessary if they are to fall through the cloud, survive evaporation in the unsaturated air below, and reach the ground as drizzle or rain. The production of a few large particles from a large population of much smaller ones may be achieved in one of two ways. The first of these depends on the fact that cloud droplets are seldom of uniform size because droplets form on nuclei of various sizes and grow under slightly different conditions and for different lengths of time in different parts of the cloud. A droplet appreciably larger than average will fall faster than the smaller ones and so will collide and fuse (coalesce) with some of those which it overtakes. Calculations show that, in a deep cloud containing strong upward air currents and high concentrations of liquid water, such a droplet will have a sufficiently long journey among its smaller neighbours to grow to raindrop size. This coalescence mechanism is responsible for the showers that fall in tropical and subtropical regions from clouds whose tops do not reach the 0° C (32° F) level and therefore cannot contain ice crystals. Radar evidence also suggests that showers in temperate latitudes may sometimes be initiated by the coalescence of waterdrops, although the clouds may later reach to heights at which ice crystals may form in their upper parts.

The coalescence process

The second method of releasing precipitation can operate only if the cloud top reaches elevations at which temperatures are below 0° C and the droplets in the upper cloud regions become supercooled. At temperatures below −40° C (−40° F) the droplets freeze automatically or spontaneously; at higher temperatures they can freeze only if they are infected with special, minute particles called ice nuclei. The origin and nature of these nuclei are not known with certainty, but the most likely source is clay-silicate particles carried up from the ground by the wind. As the temperature falls below 0° C, more and more ice nuclei become active, and ice crystals appear in increasing numbers among the supercooled droplets. But such a mixture of supercooled droplets and ice crystals is unstable. The cloudy air is usually only slightly supersaturated with water vapour with respect to the droplets and is strongly oversaturated with respect to ice crystals; the latter therefore grow more rapidly than the droplets. After several minutes the growing crystals acquire falling speeds of tens of centimetres per second, and several of them may become joined together to form a snowflake. In falling into the warmer regions of the cloud, this flake may melt and hit ground as a raindrop.

The ice crystal process

The deep, extensive, multilayer cloud systems, from which precipitation of a widespread, persistent character falls, are generally formed in cyclonic depressions (lows) and near fronts. Such cloud systems are associated with feeble upcurrents of only a few centimetres per second that last for at least several hours. Although the structure of these great raincloud systems is being explored by aircraft and radar, it is not yet well understood. That such systems rarely produce rain, as distinct from drizzle, unless their tops are colder than about −12° C (10° F) suggests that ice crystals are mainly responsible. Such a view is supported by the fact that the radar signals from these clouds usually take a characteristic form that has been clearly identified with the melting of snowflakes.

Precipitation from layer cloud systems

Showers, thunderstorms, and hail. Precipitation from shower clouds and thunderstorms, whether in the form of raindrops, pellets of soft hail, or true hailstones, is generally of great intensity and shorter duration than that from layer clouds and is usually composed of larger particles. The clouds themselves are characterized by their large vertical depth, strong vertical air currents, and high concentrations of liquid water, all these factors favouring the rapid growth of precipitation elements by the accretion of cloud droplets.

In a cloud composed wholly of liquid water, raindrops may grow by coalescence with small droplets. For example, a droplet being carried up from the cloud base would grow as it ascended by sweeping up smaller droplets. When it became too heavy to be supported by the vertical upcurrents, the droplet would then fall, continuing to grow by the same process on its downward journey. Finally, if the cloud is sufficiently deep, the droplet will emerge from its base as a raindrop.

Adapted from J. McDonald, *Scientific American*, February, 1954, Scientific American, Inc.

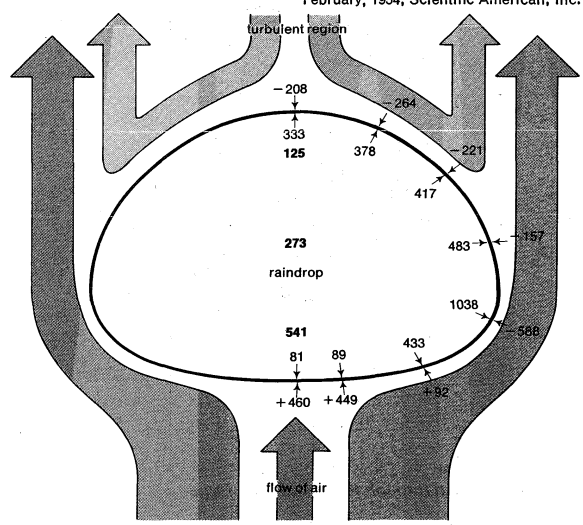


Figure 2: Pressure of air around a raindrop. Numbers outside the drop indicate the difference between the pressure around the drop and that of the atmosphere. Inside numbers denote the difference in surface pressure; boldface numbers indicate difference in internal pressure. Units are dynes per square centimetre.

In a dense, vigorous cloud several kilometres deep, the drop may attain its limiting stable diameter (about six millimetres [0.2 inch]) before reaching the cloud base and thus will break up into several large fragments. Each of these may continue to grow and attain breakup size. The number of raindrops may increase so rapidly in this manner that after a few minutes the accumulated mass of water can no longer be supported by the upcurrents and falls out as a heavy shower. The conditions that favour this rapid multiplication of raindrops occur more readily in tropical regions. In temperate regions, where the 0° C level is much lower in elevation, conditions are more favourable for the ice crystal mechanism.

Growth of hailstones

The hailstones that fall from deep, vigorous clouds in warm weather consist of a core surrounded by several alternate layers of clear and opaque ice. When the growing particle traverses a region of relatively high air temperature or high concentration of liquid water, or both, the transfer of heat from the hailstone to the air cannot occur sufficiently rapidly to allow all of the deposited water to freeze immediately. This results in the formation of a wet coating of slushy ice, which may later freeze to form a layer of compact, relatively transparent ice. If the hailstone then enters a region of lower temperature and lower water content, the impacting droplets may freeze individually to produce ice of relatively low density with air spaces between the droplets. The alternate layers are formed as the stone passes through regions in which the combination of air temperature, liquid-water content, and updraft speed allows alternately wet and dry growth.

It is held by some authorities that lightning is closely associated with the appearance of precipitation, especially in the form of soft hail, and that the charge and strong electric fields are produced by ice crystals or cloud droplets impacting on and bouncing off the undersurfaces of the hail pellets. For a detailed discussion of the cause and occurrence of electrical effects in clouds, see LIGHTNING.

By courtesy of K.A. Browning

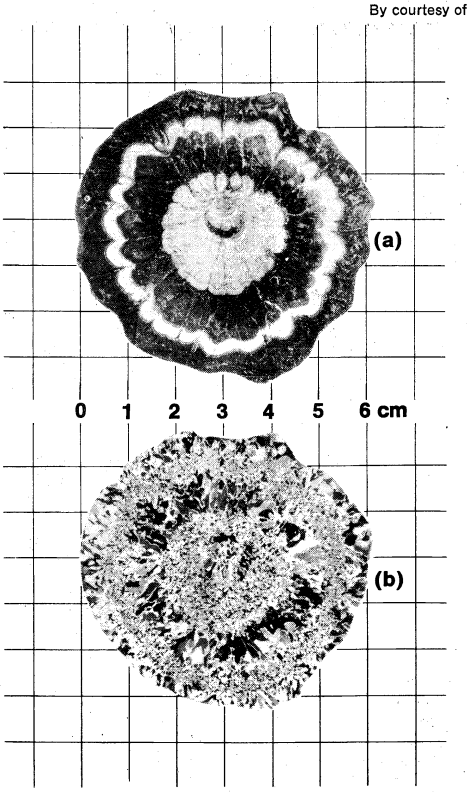


Figure 3: *Section through the centre of a large hailstone.*
(A) shows alternate layers of opaque (white) and transparent (black) ice seen in reflected light. (B) shows the individual crystals seen in polarized light. The opaque layers are composed of small crystals and the transparent layers of much larger crystals.

TYPES OF PRECIPITATION

Drizzle. Liquid precipitation in the form of very small drops, with diameters between 0.2 and 0.5 millimetre (0.008 and 0.02 inch) and terminal velocities between 70 and 200 centimetres (30 and 80 inches) per second, is defined as drizzle. It forms by the coalescence of even smaller droplets in low layer clouds containing weak updrafts of only a few centimetres per second. High relative humidity below the cloud base is required to prevent the drops from evaporating before reaching the ground; drizzle is classified as slight, moderate, or thick. Slight drizzle produces negligible runoff from the roofs of buildings, and thick drizzle accumulates at a rate in excess of one millimetre (0.04 inch) per hour.

Rain. Liquid waterdrops with diameters greater than those of drizzle constitute rain. Raindrops rarely exceed six millimetres (0.25 inch) in diameter because they become unstable when larger than this and break up during their fall. The terminal velocities of raindrops at ground level range from two metres (seven feet) per second for the smallest, to about ten metres (30 feet) per second for the largest. The smaller raindrops are kept nearly spherical by surface-tension forces, but, as the diameter surpasses about two millimetres (0.08 inch), they become increasingly flattened by aerodynamic forces. When the diameter reaches six millimetres, the undersurface of the drop becomes concave because of the airstream and the surface of the drop is sheared off to form a rapidly expanding bubble or bag attached to an annular ring containing the bulk of the water. Eventually the bag bursts into a spray of fine droplets and the ring breaks up into a circlet of millimetre-sized drops.

Rain of a given intensity is composed of a spectrum of drop sizes, the average and median drop diameters being larger in rains of greater intensity. The largest drops, with diameter greater than five millimetres (0.2 inch), appear only in the heavy rains of intense storms.

Freezing rain. When raindrops fall through a cold layer of air (colder than 0° C) and become supercooled, freezing rain occurs. The drops may freeze on impact with the ground to form a very slippery and dangerous "glazed" or "black" ice that is difficult to see because it is almost transparent.

Snow. Solid precipitation composed of small ice crystals at temperatures well below 0° C—but also of larger snowflakes at temperatures near 0° C—is termed snow. Sometimes the snow crystals become covered with small frozen cloud droplets that give them a rimed or jewelled appearance. Continued riming may lead to the formation of small flattened pellets of granular snow that rarely exceed one millimetre in diameter.

Snowfall is classed as slight, moderate, or heavy if the rate of accumulation (in the absence of drifting or melting) is less than 0.5 centimetre (0.2 inch) per hour, 0.5 to four centimetres (0.2 to 1.5 inches) per hour, or greater than four centimetres (1.5 inches) per hour, respectively. The water content of 30 centimetres (12 inches) of freshly fallen snow approximately equals that of 2.5 centimetres (one inch) of rainfall. For a description of snow crystals and snowflakes, see SNOW AND SNOWFLAKES.

Sleet. In British terminology, sleet is the term used to describe precipitation of snow and rain together or of snow melting as it falls. In America, it is used to denote partly frozen ice pellets.

Hail. Solid precipitation in the form of hard pellets of ice that fall from cumulonimbus clouds is hail. It is convenient to distinguish among three types of hail particles.

Types of hail

The first is soft hail, or snow pellets, which are white, opaque, rounded or conical pellets as large as six millimetres in diameter. They are composed of small cloud droplets frozen together, have a low density, and are readily crushed.

Second is small hail (ice grains or pellets), transparent or translucent pellets of ice, spherical, spheroidal, conical, or irregular in shape, with diameters as large as a few millimetres. They may consist of frozen raindrops, of largely melted and refrozen snowflakes, or of snow pellets encased in a thin layer of solid ice.

True hailstones, the third type, are hard pellets of ice, larger than five millimetres in diameter, that may be spherical, spheroidal, conical, discoidal, or irregular in shape and often have a structure of concentric layers of alternately clear and opaque ice (see Figure 3). A moderately severe storm may produce stones a few centimetres in diameter, whereas a very severe storm may release stones with maximum diameters of ten centimetres (four inches) or more. Large damaging hail falls most frequently in the continental areas of middle latitudes (*e.g.*, in the Nebraska–Wyoming–Colorado area of the United States, in South Africa, and in northern India) but is rare in equatorial regions. Hailstones of at least walnut size fall at least two days per year in the Denver area of

Colorado. The largest hailstones ever reported from the United States fell in Potter, Nebraska, on July 6, 1928, the largest being 13.8 centimetres (5.4 inches) in diameter and weighing 680 grams (1.5 pounds). Terminal velocities of hailstones range from about five metres (16 feet) per second for the smallest stones to perhaps 40 metres (130 feet) per second for stones five centimetres (two inches) in diameter.

WORLD DISTRIBUTION OF RAINFALL

Regional and latitudinal distribution. The yearly rainfall averaged over the whole Earth is about 80 centimetres (30 inches), but this is distributed very unevenly. The regions of highest rainfall are found in the equatorial zone and the monsoon area of Southeast Asia. Middle latitudes receive moderate amounts of precipitation, but little falls in the desert regions of the subtropics and around the poles.

If the Earth's surface were perfectly uniform, the long-term average rainfall would be distributed in distinct latitudinal bands, but the situation is complicated by the pattern of the global winds, the distribution of land and sea, and by the presence of mountains. Because rainfall results from the ascent and cooling of moist air, the areas of heavy rain indicate regions of rising air, whereas the deserts occur in regions in which the air is warmed and dried during descent. In the subtropics, the trade winds bring plentiful rain to the east coasts of the continents, but the west coasts tend to be dry. In high latitudes, on the other hand, the west coasts are generally wetter than the east coasts. Moreover, rain tends to be abundant on the windward slopes of mountain ranges but sparse on the lee sides.

Equatorial and desert regions

In the equatorial belt, the trade winds from both hemispheres converge and give rise to a general upward motion of air, which becomes intensified locally in tropical storms that produce very heavy rains in the Caribbean, the Indian and southwest Pacific oceans, and the China sea, and in thunderstorms that are especially frequent and active over the land areas. During the annual cycle, the doldrums move toward the summer hemisphere so that, outside a central region near the Equator, which has abundant rain at all seasons, there is a zone that receives much rain in summer but a good deal less in winter.

The dry areas of the subtropics, such as the desert regions of North Africa, the Arabian Peninsula, South Africa, Australia, and central South America, are due to the presence of semipermanent, subtropical anticyclones in which the air subsides and becomes warm and dry. These high-pressure belts tend to migrate with the seasons and cause summer dryness on the poleward side and winter dryness on the equatorward side of their mean positions. The easterly trade winds, having made a long passage over the warm oceans, bring plentiful rains to the east coasts of the subtropical landmasses, but the west coasts and the interiors of the continents, which are often sheltered by mountain ranges, are very dry.

Middle and high latitudes

In middle latitudes the weather and the rainfall are dominated by travelling depressions and fronts that give a good deal of rain in all seasons and in most places except the far interiors of the Asian and North American continents. Generally the rainfall is more abundant in summer, except on the western coasts of North America, Europe, and North Africa, where it is higher during the winter.

At high latitudes and especially in the polar regions, the low precipitation is caused partly by subsidence of air in the high-pressure belts and partly by the low temperatures. Snow or rain occur at times, but evaporation from the cold sea and land surfaces is slow, and the cold air has little capacity for moisture; for this reason precipitation amounts are small.

The influence of oceans and continents on rainfall is particularly striking in the case of the Indian monsoon. During the Northern Hemisphere winter, cool, dry air from the interior of the continent flows southward and produces little rain over the land areas. After the air has travelled some distance over the warm tropical ocean, however, it releases heavy shower rains over the East Indies. During the northern summer, when the monsoon blows from the southwest, rainfall is heavy over India and Southeast Asia. These rains are intensified where the air is forced to ascend over the windward slopes of the Western Ghāts and the Himalayas.

The combined effects of land, sea, mountains, and prevailing winds show up in South America. There the desert in southern Argentina is sheltered by the Andes from the westerly winds blowing from the Pacific Ocean, and the west-coast desert is not only situated under the South Pacific subtropical anticyclone but is also protected by the Andes against rain-bearing winds from the Atlantic.

Amounts and variability. The long-term average amounts of precipitation for a season, or a year, give little information on the regularity with which rain may be expected, particularly for regions where the average amounts are small. At Doorbaji, in the Thar Desert, for example, where the annual amount is about 13 centimetres (five inches), a rare storm produced 86 centimetres (34 inches) in two days. At Iquique, in northern Chile, four years once passed without rain, whereas the fifth year gave 1.5 centimetres (0.6 inch); the five-year average was therefore 0.30 centimetre (0.12 inch). Clearly such averages are of little practical value and the frequency distribution or the variability of precipitation also must be known.

Relation of variability and average amount

The variability of the annual rainfall is closely related to the average amounts. For example, over the British Isles, which have a very dependable rainfall, the annual amount varies by less than 10 percent above the long-term average value. A variability of less than 15 percent is typical of the middle-latitude cyclonic belts of the Pacific and Atlantic oceans and of much of the wet equatorial regions. But in the interiors of the desert areas of Africa, Arabia, and Central Asia, the rainfall in a particular year may deviate from the normal long-term average by more than 40 percent. The variability for individual seasons or months may differ considerably from that for the year as a whole, but again the variability tends to be higher where the average amounts are low.

The heaviest annual rainfall in the world was recorded at Cherrapunji, India, where 2,647 centimetres (1,042 inches) fell between August 1860 and July 1861. The heaviest rainfall in a period of 24 hours was 187 centimetres (73.62 inches) recorded at Cilaos, La Reunion, in the Indian Ocean on March 15–16, 1952. The lowest recorded rainfall in the world occurred at Arica, in northern Chile. The annual average, taken over a 43-year period, is only 0.05 centimetre (0.02 inch).

Although past records give some guide, it is not possible to estimate very precisely the maximum possible precipitation that may fall in a given locality during a specified interval of time. Much will depend on a favourable combination of several factors, including the properties of the storm and the effects of local topography. Thus it is possible only to make estimates based on analyses of past storms or on theoretical calculations that attempt to maximize statistically the various factors or the most effective combination that are known to control the duration and intensity of the precipitation. For many important planning and design problems, however, estimates of the greatest precipitation to be expected at a given location within a specified number of years are required.

In the design of a dam, the highest 24-hour rainfall to be expected once in 30 years over the whole catchment area might be relevant. For dealing with such problems, a great deal of work has been devoted to determining, from past records, the frequency with which rainfalls of given intensity and total amount may be expected to reoccur at particular locations and also to determining the statistics of rainfall for a specific area from measurements made at only a few points.

EFFECTS OF PRECIPITATION

Raindrop impact and soil erosion. Large raindrops, up to six millimetres (0.2 inch) in diameter, have terminal velocities of about ten metres (30 feet) per second and so may cause considerable compaction and erosion of the soil by their force of impact. The formation of a com-

pacted crust makes it more difficult for air and water to reach the roots of plants and encourages the water to run off the surface and carry away the topsoil with it. In hilly and mountainous areas, heavy rain may turn the soil into mud and slurry, which may produce enormous erosion by mudflow generation. Rainwater running off hard, impervious surfaces or waterlogged soil may cause local flooding.

Surface runoff. The rainwater that is not evaporated or stored in the soil eventually runs off the surface and finds its way into rivers, streams, and lakes or percolates through the rocks and becomes stored in natural underground reservoirs. A given catchment area must achieve an overall balance such that precipitation (P) less evaporation of moisture from the surface (E) will equal storage in the ground (S) and runoff (R). This may be expressed: $P - E = S + R$. The runoff may be determined by measuring the flow of water in the rivers with stream gauges, and the precipitation measured by a network of rain gauges, but storage and evaporation are more difficult to estimate.

Of all the water that falls on the Earth's surface, the relative amounts that run off, evaporate, or seep into the ground vary so much for different areas that no firm figures can be given for the Earth as a whole. It has been estimated, however, that in the United States 10 to 50 percent of the rainfall runs off at once, 10 to 30 percent evaporates, and 40 to 60 percent is absorbed by the soil. Of the entire rainfall, 15 to 30 percent is estimated to be used by plants, either to form plant tissue or in transpiration.

BIBLIOGRAPHY. There are three general works that treat the processes of condensation and precipitation from clouds from theoretical and experimental bases, and which provide peripheral information on such aspects as cloud seeding as well. These are: N.H. FLETCHER, *The Physics of Rainclouds* (1962); and B.J. MASON, *Clouds, Rain and Rainmaking* (1962), and *The Physics of Clouds,* 2nd ed. (1971). The last is quite comprehensive and each of these books is a source for additional references.

(B.J.M.)

Prefabrication and Shop Fabrication

Prefabrication as used here encompasses the assembly of buildings or their components at locations other than their final position in the completed structure. Prefabrication may take place at the building site itself or may be carried out at remote sites ranging up to several thousand miles away—occasionally even crossing international boundaries. Prefabrication methods may be used for items such as residential kitchen cabinets and appliances, wall panels, floor panels, roof trusses, window-wall elements, structural rigid frames, room-sized components, and even total buildings.

Basic types of preassembly

There are two basic types of preassembly associated with the construction industry: custom fabrication and general fabrication. The first method is used widely in commercial and institutional construction when the design calls for several identical components within a single structure or a complex of buildings. Hospitals, dormitories, apartment buildings, offices, and schools are examples of possible applications for custom prefabrication. On the other hand, the general prefabrication class of components is designed for broad usage and is marketed for application in a wide range of building types and designs. Some of the elements in this classification are roof trusses for residences and small commercial and industrial buildings; rigid frames for warehouses, bowling alleys, and small factories; and bathroom and kitchen "core" modules for residential use.

Although there is evidence to indicate that the enormous stone blocks of the Pyramids of Egypt were cut roughly to size in the quarries and hewn to exact shape on site prior to their final positioning, a process that would be called precutting in present-day terminology, the history of prefabrication starts in 1905 when a design was made for the construction of a factory to manufacture precast concrete box-shaped modules to be hauled by truck for consumer use in Chicago. From World War I on, prefabrication has made rapid, if irregular, strides in western Europe, the U.S.S.R., and the United States.

ECONOMICS OF PREFABRICATION

Time. The promise of reduced costs has been the prime, but not the only, motivation for the application of prefabrication techniques. The use of prefabricated modules in Russian apartment blocks, in the Palacio del Rio Hotel in San Antonio, Texas, and in Habitat '67 in Montreal, as well as in several other cases, was in response to a time requirement and not in anticipation of reduced costs.

Weight and durability of materials. The economics of buildings differs fundamentally from the economics of other manufactured items. Since buildings are not moved but must stand against the weather for 60 or more years, material selection for buildings follows far different criteria than for transportation equipment. Building materials can be heavy since they are transported to the building site only once and, after that, never relocated. Furthermore, building weight aids in resisting the corrosion that is associated with a long life of minimal maintenance.

Many manufactured items consist of a great number of small, intricate parts; buildings are made up of a relatively few large, simple parts. There is little need for malleability in the structural and architectural elements of buildings. Even in earthquake zones, buildings are subjected to only slight vibrations compared to those constantly experienced by most manufactured products. The need for highly efficient joining techniques is not as critical in buildings as it is in other products.

When the economic factors are well understood and properly dealt with in design, the prefabrication of major architectural elements up to and including whole buildings is feasible. A climate more favourable to prefabricated construction is currently developing. The use of low-cost plastics as binders of natural aggregates shows promise of providing materials of proper strength, workability, and cost to satisfy the specific needs of prefabrication and its associated transportation. In many countries, rising labour costs are increasing the demand for alternate methods of construction.

ADVANTAGES OF PREFABRICATION

Auxiliary equipment. The inherent advantages of prefabrication in construction are especially noteworthy in the area of auxiliary equipment. Complex duct systems for heating, ventilating, and air conditioning are prefabricated in the shops of subcontractors and delivered to the construction site for final erection. Networks of electrical raceways and heating- and cooling-system ducts are also shop fabricated. Concrete roof, floor, and decorative exterior wall panels are precast and frequently prestressed in large factories devoted specifically to those purposes. Open-web steel joists—featuring light metal framework and steel or even wood runners—rigid frames, and wooden roof trusses represent well-established general-purpose prefabricated items. All of these building elements are based on standard designs with dimensions varying for the specific project at hand.

Modular construction. Modularized prefabricated components are seen more in architectural, mechanical, electrical, and structural design uses than in readily visible applications. The term module has two meanings in building design: (1) It indicates a uniform linear dimensional guide for coordinated architectural and engineering design. For example, the architect may select a four-foot module for the floor plan of a particular building. The structural, mechanical, electrical, and other consulting engineers will design connections, joints, columns, etc. to fall on these four-foot grid lines. In this scheme wall panels may be four feet wide, columns may be spaced at 16 feet, and contraction joints of exterior concrete plazas may be on a 2-foot spacing. All of these dimensions are consistent with the four-foot module even though some elements are integral multiples of the basic module and some are integral divisions of it. (2) The

second common use of the term module refers to three-dimensional boxlike architectural–structural elements, generally with all utilities included. These building components also are designed on a modular basis as described under (1) above but are prefabricated in plants and delivered to the site as units to be "plugged" into the building. The usages of the term can be distinguished by the context. Modularized prefabricated components are represented by such items as kitchen cabinet and appliance units, electrical panels and switchgear, furnaces, water heaters, and bathroom elements. Habitat '67 in Montreal and the Palacio del Rio in San Antonio are examples of this concept in room-sized prefabricated components.

Economic considerations

Cost reduction through prefabrication is based principally on the use of lower cost labour and mass-production techniques in factories to supplant higher rate labour and slower manual methods at the construction site. Since the basic materials of construction are generally less expensive than those of any other industry, there is little hope of reduced material cost when using factory assembly methods. Indeed, it may be necessary to select materials more expensive than those presently used in on-site construction to take advantage of the available differences in labour cost. Development of materials tailored specifically for prefabricated construction use may be one of the keys to realizing the promise of factory-assembled building modules.

Interchangeable parts. Factory assembly cuts down the time needed for final erection of the building at the site. The influence of weather is markedly reduced. Realistic design of interchangeable subassemblies and components permits a wide variety of products to be made up from a few standard construction elements. Such an approach is conducive to manufacturing efficiency and at the same time makes it possible to avoid the bane of prefabrication: aesthetic monotony. In current projects to achieve the maximum use of standardized, interchangeable parts, the structure of the building merely provides a framework for the attachment of utility modules such as bathrooms, kitchens, sleeping elements, laundry and utility rooms, and entertainment alcoves. Space for living and working is divided by movable partition and closet units. Electric power, water, gas, and waste drainage are made available in a modular pattern so that utility units can be located in a wide variety of ways within a building. If a bathroom unit were to deteriorate or become obsolete, it could be removed and replaced in a few hours without major disruption of activities within the overall building.

Blankmeyer system

In its present state of development, this concept, called the Blankmeyer system after its inventor, Donald C. Blankmeyer of the United States, applies principally to residential construction. The principal value of the system is that it extends the function of existing appliance manufacturers into the construction of units containing their appliances but also incorporating small wall, floor, and roof elements. Manufacture of the modules would require standardization similar to that of the auto industry, but beyond that units could be competitively manufactured and distributed in the same ways that autos and appliances are now purveyed.

Prefabrication in construction is aimed primarily at the future, since the past is not full of successes and public acceptance. The system provides for production efficiency through mass manufacture of standard products, it allows for the proper selection of materials in direct response to needs, requires shipment of components small enough for easy handling without special escorts and truck sizes, reduces expensive on-site labour, allows the user to choose from among the offerings of many competitive manufacturers, and provides for orderly updating of components in direct response to normal deterioration and obsolescence.

Land and building considerations. Modular designs respond to the various desired lifetimes of the constituents of buildings and sites. For example, the land has an infinite life, except that its value may change in response to community needs for service. That is, a piece of land occupied by an apartment building may at some point in time be more valuable to the community if it supports a commercial building or factory. But compared to the life of a building, the life of its site can be considered as infinite. A properly designed and constructed foundation has an expected life of as much as 120 years. A structural framework is expected to be serviceable for 60 or more years and—with routine maintenance—the life of a roof is in the same range. But mechanical, electrical, and other utility systems have expected useful lives of only 5 to 15 years because of the greater susceptibility of these elements to deterioration and to obsolescence from the rapid advance of technology. Therefore, to retain the durability, safety, and variety required and at the same time to realize the economies inherent in the mass manufacture of components, any system must take into account the various lifetimes of the constituents of a building.

PRODUCTION PLANNING AND PROCEDURES

Specialized processes. The available market for prefabricated elements is nearly as broad as the facets of the construction industry itself. By such techniques as precasting and prestressing, and lift-slab construction, massive concrete components have been prefabricated and raised into place. Some lift-slab floors and roofs in a structure have approached one acre in total area and 2,000 tons in weight. Even domed roofs have been constructed through lift-slab operations.

Prestressing, precasting concrete

Prestressing and precasting is a technique wherein high-strength steel wires are run through concrete molds or formwork and tensioned. While the wires are in tension, concrete is poured into the mold and allowed to cure. After cure the wire tension is released; as the tension relaxes, the wires shorten, applying a compressive stress to the concrete (see also CONCRETE).

This technique is useful in concrete construction because concrete can tolerate only light tensile loads but can support extremely high loads in compression. Solid steel structural members can resist high tensile loads, but wires or rods alone can resist almost no compressive load because the small diameter of wire and rod introduces a tendency to buckle. By combining the best characteristics of the two materials, concrete and steel, prestressing provides efficient, lightweight, fire-resistant building units.

Lift-slab construction is used principally for the erection of floors and roofs. In this operation foundations are prepared in the usual fashion. Columns are erected to a height of as much as 40 to 50 feet (12 to 15 metres). After the base section of a column is set, the bottom floor slab is cast; this may be a basement floor or a grade-level slab. A special, liquid release compound, designed to prevent bonding of the concrete slabs, is applied to the top surface of the slab in preparation for other floors to be cast directly on it. Structural collars for each floor level are placed over the columns and lowered to the surface of the floor slab. These collars provide guidance for the next slab and establish clearance from the column so that the concrete will not adhere and restrain the upward movement of the slab during lifting. Lower slabs are used as the undermolds for the next-higher slabs; the casting process is continued at the ground level until all floors and the roof slab are cast. After the roof slab is suitably cured, the lifting process is initiated by placing one jack atop each column and attaching it to its collar by two long, threaded rods. These rods extend down from the jack on either side of the column and engage the collars by use of "lifting nuts" threaded onto the lower end of each rod. The jacks slowly pull the slabs up the columns by applying tension to the lifting rods. Power is provided by a central hydraulic unit and control is through a central electronic system. By alternately lifting in short increments and automatically taking up the lift by screwing down nuts engaged on the threaded rods, the slab is accurately raised either directly to its final position or first to an intermediate position for structural inspection before final lifting. As soon as any slab has reached its final position, it is welded permanently. The openings for engagement of the lifting rods in the collars also penetrate through all the slabs so that the rods extend through the upper slabs and attach to the next-lower slab.

Prehung doors

At the other end of the size and weight scale for prefabricated elements are prehung doors and windows, prefabricated electrical distribution centres, prefabricated plumbing assemblies, and prefabricated chimneys and vent stacks. The market scope is broad and the demand for construction is high, in response to expanding world population with its associated need for products and employment. In the case of these components, the speed offered by prefabrication is a factor in its further use even when there is no cost advantage.

Building codes, inspections. In general, governments have encouraged, or at least have not resisted, prefabrication provided that safe quality standards have been maintained, but inspection to assure adherence to local standards is made difficult by the fact that factories are frequently outside the cities of intended delivery. In prefabrication, on-site inspectors cannot observe hidden construction elements; in some countries this problem has created difficulties. The solution visualized is standardization of quality requirements and cooperative inspection at the factory and at the site.

Use of modern management techniques. Factory production enhances the value of systems engineering, and modern management techniques of planning and scheduling are fundamental to the achievement of the economic promise of prefabrication. Management methods developed for space exploration programs have demonstrated that it is possible to unify the planning and control of extremely broad, multifaceted projects. Computer analysis of progress and of plan changes provides the mechanism for quick response and sensitive control of complex operating systems. The development of resource allocation techniques and the understanding of construction logistics are contributing to improved efficiency. These management methods are the key to success of the prefabrication industry.

Logistics of prefabrication. Contemporary management techniques assist in decision making regarding proper compromises between hauling costs and labour saving through factory production. Large components such as room-sized modules require special design consideration to render them rigid enough for transportation. When rigidity is provided, the units become heavy and require special handling equipment for erection at the job site. When units are so massive that a truck can accommodate only one, economics dictates that the distance they must be hauled preferably be no farther than the round trip that the truck can make in one day, including allowance for unloading time. The economics of rail shipment or barging of very large units is often complicated because additional handling is required to get the units to the construction site.

Stockpiling

Another advantage of prefabrication lies in the possibility of having large volumes of materials delivered to a single storage point. Suitable warehouse space allows the purchase and stockpiling of materials at advantageous prices.

The scope of definition of component materials includes subassemblies and subsystems manufactured in plants owned either by the same producer or by others.

MATERIALS AND TECHNIQUES

Wood-framed units. *Precut lumber and prefabricated walls.* One of the earliest excursions into prefabrication was the adoption of precut framing lumber pieces for walls and partitions. The next evolutionary step involved prefabricating exterior wall and interior partition panel framing on the floor of the building, tipping the frames upright, and securing them. In recent years a few lumber suppliers have equipped themselves with fixtures and tooling to build these frames in their warehouses and have offered them for sale as prebuilt units. Many sophisticated and versatile systems were devised to facilitate the installation of insulation, exterior siding, interior finish, windows, and doors. One of these, the Queensbury system, employs a versatile fixture for quickly positioning studs, plates, and sills for wall framing. Framing for doors and windows at any location is accounted for within the scope of the system design.

In all of the evolutionary steps discussed thus far, the components are custom built for specific orders and no stockpiling of completed units is considered. This is also true in large fabrication organizations. The difference between the mode of operation of the large prefabrication company and the small lumber supplier lies mainly in the approach to design. The lumberman builds in response to a design brought to him by the building contractor, while the prefabrication company prepares designs intended specifically to facilitate factory production.

Truss designs. Prefabricated roof trusses are widely used in light construction. The early trusses used metal split-ring connectors, resembling short sections of pipe, which were driven into holes aligned within the lumber chords (horizontal units), and web (diagonal) members. To use these connectors, the lumber elements were offset, or overlapped, so that the members could pass alongside one another to accommodate the special characteristics of split rings. This made the trusses twice as thick as necessary and introduced undesirable secondary bending stresses. The next development placed the truss members in the same plane by abutting the ends and used small plywood gusset plates nailed or glued on either side of the joints to connect the members. This development was immediately followed by a multitude of metal gusset-plate designs and fabrication systems. Sophisticated fixtures and automatic saws were designed to automate the truss-making process. Presently, trusses are manufactured and stockpiled in standard sizes.

Plywood components

During the 1960s considerable research was done on the design and manufacture of prefabricated plywood structural components. In such components as stressed skin panels, box beams, folded plates—flat slabs, abutted and connected at their edges—and curved roof panels and arches, plywood and lumber are glue laminated to take best advantage of the structural characteristics of each material, a significant departure from merely prefabricating conventional framing in the shop.

Packaged structures. Another important new concept is that of the "folded plate." A typical folded plate structure consists of flat-plane (or plate) wall and roof members, interconnected and delivered to the construction site as a package (8 x 9 x 40 feet). With the assistance of a deck lift mounted on the delivery truck, the accordion-like package unfolds to form a building (40 feet wide by 72 feet long with a 14-foot ceiling height at the lowest point). All walls, doors, and windows are included. Such a building is not at present economically competitive in the commercial market but is of value in fulfilling certain special needs for quick erection and relocatability—*e.g.*, military. Yet the sophistication of the technique offers promise for the future.

Steel construction. Steel construction has always involved a degree of prefabrication in the sense that almost every steel structural member is cut to exact length and has connection holes punched in the shop. Complex framing systems are fabricated into subassemblies in sizes suitable for shipping, and before shipment these subassemblies are test connected to assure proper matching of connection holes. Steel columns and other compression members are cut to exact length and milled square in the shop. Steel reinforcing rods are cut to length and bent in the factory. Open-webbed steel joists are welded into complete structural units in fabrication plants. But in addition to these traditional fabrication processes, the approach presently used in the steel industry of the world goes much farther.

Steel rigid frames with spans up to 120 feet are listed in the catalogs of fabricators. Relocatable wall panels made of two layers of thin sheet steel separated by an insulating material are available for use in schools and office buildings. Porcelainized-steel exterior-curtain wall panels are in evidence in large volume. Commercial stairs are made of prefabricated steel with concrete (terrazzo) poured in metal pans to form the treads.

Metal deck panels for floors and roofs incorporate raceways and ducts for electrical work and the top surfaces provide forms for concrete or other finish material. Often floor tile or roofing is attached directly to the steel.

Whole systems of steel prefabrication have been designed in the U.S. for such purposes as school construction, in which a few well-designed modules can be fitted together in a variety of ways.

Civil-engineering works. Prefabrication, or "shop fabrication" as it is more generally known in this case, plays an increasingly important role in the construction of civil-engineering works, notably bridges. Instead of erecting roadway and stiffening trusses separately, in modern bridge construction whole units, weighing as much as 400 tons and comprising roadway skeleton sections and truss sections joined together, are fabricated "in the shop" and transported to the site. On the Verrazano-Narrows Bridge in New York, for example, the bridge is made up of 60 such sections.

Bridges

An even more striking innovation was tried in the late 1960s—prefabrication of suspension cables, increasingly used for suspending roofs as well as bridges.

Use of concrete. Precast concrete components range from sewer tile and septic tanks to the huge lift slabs discussed earlier. Although wooden components are used extensively in Scotland and the Scandinavian countries, concrete is the major material turned to for prefabrication in both hemispheres and on both sides of the Iron Curtain. Structural elements include slabs, beams, columns, panels, modular boxes, and stairways. Frequently, these elements are cast in combinations, as in tees and double tees designed as roof components up to 12 feet wide and over 100 feet clear span; *i.e.*, without support from columns or walls except at the ends. The structural components are termed tees because in section their profile resembles the letter T. Similar tees, standing upright, may form exterior walls. Several schemes for reducing the weight of concrete structural members are in evidence. Unnecessary material can be eliminated by casting closely spaced longitudinal cylindrical holes through slabs or by casting reinforced ribs or waffle patterns in the undersides of thin slabs. The performance characteristics of the basic material can be improved by prestressing as in the tees, double tees, and flanged beams, for buildings or highway bridge construction. Prestressed concrete can be made flexible enough for diving boards.

Flat concrete panels are used in England for large suburban apartment projects, for structures up to five stories in height. Walls and partitions are load bearing. Slabs are simple rectangles of uniform thickness. The Techcrete system developed by architect Carl Koch of the United States and the Balenci system of Italy are similar to the British system. Techcrete has been used in structures up to nine stories high. The Jespersen system of Denmark is also a wall-bearing panel and slab design. It differs from the other systems in that Jespersen incorporates stairs, beamed slabs, and many other special features.

Various national construction systems

Immediately following World War II, several precast factory buildings were constructed using rigid frames and slabs. A Danish box system also uses precast rigid frames, but the frames support box modules for residential occupancy. The Danish plant is subsidized by the government.

Several other nations experimented with similar systems but as of 1970 all had been abandoned. Russia developed two box systems: the lightweight and the massive. These units were both architectural and structural since the attractive faces of the box also were load bearing. The units used in the Palacio del Rio Hotel in San Antonio, Texas, and in Habitat '67, Montreal, were also load bearing boxes.

The most intricate precasting system appears in the Swedish Heart system of precast bathroom and kitchen elements. A wall is cast with piping and vents formed directly in the concrete. Sink, lavatories, bathtub, and water closets are all molded integrally with the wall and floor sections. The bathroom is on one side of the dividing wall and the kitchen on the opposite side in the usual back-to-back relationship. Various concrete finishes are made so that fixture material surface is different from wall and floor material. The fixtures are terrazzo-like while the walls and floors are of ordinary concrete. The workmanship is excellent but the rough surfaces may fall short of contemporary comfort standards.

Precast concrete is also widely used in shop fabrication for civil-engineering projects. A notable example in the 1970s was a double viaduct built in Switzerland to carry a European highway past Chillon Castle on Lake Geneva. An automated plant supplied concrete, preheated in winter, to the five casting beds in the prefabrication yard just south of the construction site. A total of 1,380 boxes were required to make up the two roadways, with many variations in form owing to the curve of the viaduct and the cantilevering from the piers, which meant deeper boxes directly over the piers and shallower ones farther out.

Other techniques and materials. *Utility cores.* A set of prefabricated elements not classified by basic material has had impact of various degrees upon prefabrication. The most important of these is the utility core in which kitchen, bath, and laundry are arranged along opposite sides of a common wall containing electric power, water, and drainage. The degree of fabrication varies among builders. Some fabricators assemble only the service elements attached directly to a wet wall. Others include the walls, floors, and ceilings of adjacent rooms. In the latter instance, a heat plenum or warm-air chamber is built into the ceiling; supply grilles are installed around the plenum to provide central heating for the entire dwelling unit. The principal manufacturers of these units are building material suppliers.

Plastics. The use of plastics in construction was rising in the early 1970s. Plastic drain waste and vent piping is well established in many areas of the world. Plastic water piping may prove to be successful. The ease of joining, coupled with more complex shapes possible with plastic piping, can reduce plumbing labour by 75 percent. Bonded vinyl exterior finishes not only replace paint in some instances but also provide the protection necessary for the use of metals as siding materials. The durable beauty of vinyl has been demonstrated through its use on autos. Inexpensive thermosetting plastic bonding agents under development, to be used like mortar, may fill a well-defined gap in the economic spectrum of building materials.

Site preparation. Modular design normally associated with prefabrication can help organize and standardize site work, but there will always be site preparation. A site-constructed structural system is necessary to compensate for the natural soil irregularities of the earth and the finished shapes of the building. Soil factor irregularities may include surface contours, load-bearing capability, or dimensional instability due to moisture-content changes and freezing.

Connections to water, sewer, gas, electricity, and telephone remain as projects at the building site. These items are no more expensive for prefabricated construction than for conventional buildings, but any unnecessary work at the site cuts down on the savings to be expected from factory assembly.

BIBLIOGRAPHY. There are no good English language texts published on this subject. Information may be found in periodical literature, such as the following: K.J. SEYMOUR-WALKER, "Czechoslovakia: Concrete Panel Production," *Interbuild,* 14:16–19 (1967), a description of sophisticated types of equipment for concrete panel wall production used in Czechoslovakia, where a large majority of dwellings are constructed by industrialized methods; "Building a City with King Kong Blocks," *Progressive Architecture,* 47:226–236 (1966), a detailed, illustrated description of Habitat '67, the housing project in Montreal, Canada, that was the first structure making practical use of precast concrete in house-size units; "System Building, Switzerland," *Interbuild,* 14: 24–27 (1967), a description of the "Norm-Module" precast concrete constructional system developed in Switzerland, primarily for industrial, scientific, and office building application; L.O. ANDERSON, "Development of an Improved System of Wood-Frame House Construction," *U.S. Forest Serv. Res. Pap. No. FPL47* (1965), reports on a new system of wood-frame residential construction using factory assembled, prefinished components for rapid assembly on site with minimum labour; H.B. DICKENS, "Housing Trends in Europe: A Seminar in Prague," *Nat. Res. Coun., Ottawa, Tech. Pap. No. 196* (1965), a discussion of system building activities and trends in industrialized housing in various European countries; H.L. SCOGGIN, "Prefabricated Building Systems in Europe," *Building Research,* 3:58–61 (1966), reports on a

survey of prefabricated building systems throughout the world, particularly in Europe, with emphasis on the degree of mechanization, range of application, and method of construction; R.M.E. DIAMANT, *Industrialised Building: 50 International Methods* (1964), selected articles from the *Architect and Building News* that describe systems representing the most advanced international thinking in the field of factory made buildings; R.T. LASKO, "Economics of Precast Concrete Construction," *Building Research*, 3:46–48 (1966), a discussion of the advantages and limitations of precast construction.

(E.B.F.)

Pregnancy

Pregnancy, as discussed in this article, is the physiologic process and the series of changes that normally take place in a woman's organs and tissues as a result of her having a developing offspring within her body. The terms "pregnancy" and "gestation" will be used synonymously. Most of the changes that take place during pregnancy serve to safeguard the mother and the fetus. Besides changes of this type, the present article will discuss some of the problems that can arise during the course of a pregnancy. (Gestation in animals is treated in the articles REPRODUCTION; REPRODUCTIVE SYSTEMS, ANIMAL.)

This article is outlined as follows:

I. The normal events of pregnancy
  Initiation of pregnancy
  Diagnosis of pregnancy
  Anatomic and physiologic changes that attend the progress of normal pregnancy
II. Abnormal changes in pregnancy
  Ectopic pregnancy
  Abortion
  Effects of systemic diseases on pregnancy and or pregnancy on systemic diseases
  Diseases of pregnancy

I. The normal events of pregnancy

INITIATION OF PREGNANCY

A new individual is created when the elements of a potent sperm merge with those of a fertile ovum, or egg. Before this union both the spermatozoon (sperm) and the ovum have migrated for considerable distances in order to achieve their union. A number of actively motile spermatozoa are deposited in the vagina, pass through the uterus, and invade the uterine (fallopian) tube, where they surround the ovum. The ovum has arrived there after extrusion from its follicle, or capsule, in the ovary. After it enters the tube, the ovum loses its outer layer of cells as a result of action by substances in the spermatozoa and from the lining of the tubal wall. Loss of the outer layer of the ovum allows a number of spermatozoa to penetrate the egg's surface. Only one spermatozoon, however, normally becomes the fertilizing organism. Once it has entered the substance of the ovum the nuclear head of this spermatozoon separates from its tail. The tail gradually disappears, but the head with its nucleus survives. As it travels toward the nucleus of the ovum (at this stage called the female pronucleus) the head enlarges and becomes the male pronucleus. The two pronuclei meet in the centre of the ovum, where their threadlike chromatin material organizes into chromosomes.

Union of the ovum and the sperm

Originally the female nucleus has 44 autosomes (chromosomes other than sex chromosomes) and two (X, X) sex chromosomes. Before fertilization a type of cell division called a reduction division brings the number of chromosomes in the female pronucleus down to 23, including one X-chromosome. The male gamete, or sex cell, also has 44 autosomes and two (X, Y) sex chromosomes. As a result of a reducing division occurring before fertilization, it, too, has 23 chromosomes, including either an X or a Y sex chromosome at the time that it merges with the female pronucleus.

After the chromosomes merge and divide in a process termed mitosis, the fertilized ovum, or zygote, as it is now called, divides into two equal-sized daughter cells. The mitotic division gives each daughter cell 44 autosomes, half of which are of maternal and half of paternal origin. Each daughter cell also has either two X-chromosomes, making the new individual a female, or an X- and a Y-chromosome, making it a male. The sex of the daughter cells is determined, therefore, by the sex chromosome from the male parent.

Fertilization occurs in the uterine tube. How long the zygote remains in the tube is unknown, but it probably reaches the uterine cavity about 72 hours after fertilization. It is nourished during its passage by the secretions from the mucous membrane lining the tube. By the time it reaches the uterus it has become a mulberry-like solid mass called a morula. A morula is composed of 60 or more cells. As the number of cells in a morula increases, the zygote forms a hollow bubble-like structure, the blastocyst. The blastocyst, nurtured by the uterine secretions, floats free in the uterine cavity for a short time and then is implanted in the uterine lining. Normally, the implantation of the blastocyst occurs in the upper portion of the uterine lining. (The mechanism of implantation is described below.)

The morula and the blastocyst

DIAGNOSIS OF PREGNANCY

Symptoms and signs; biological tests. There is usually little doubt about the diagnosis when a young woman has missed several menstrual periods, has morning nausea, fullness and tenderness of the breasts, and has been having a normal sexual life; but the positive and certain signs of gestation are the sounds of the fetal heartbeat, which are audible with a stethoscope between the 16th and the 20th week of pregnancy; X-ray views of the fetal skeleton, which can be observed by the 14th or the 16th week; and fetal movements, which usually occur by the 18th to the 20th week of pregnancy. X-ray examinations generally are avoided, however, because of their potential hazard to the fetus.

Persons who note their body temperature upon awakening, as many women do who wish to know when they are ovulating, may observe continued elevation of the temperature curve well beyond the time of the missed period; this is strongly suggestive of pregnancy. During the early months of pregnancy, women may notice that they urinate frequently, because of pressure of the enlarging uterus on the bladder; that they feel tired and drowsy; dislike foods that were previously palatable; have a sense of pelvic heaviness; are subject to vomiting (which can be severe) and to pulling pains in the sides of the abdomen, as the growing uterus stretches the round ligaments that help support it, singly or together. Most of these symptoms subside as pregnancy progresses. The signs and symptoms of pregnancy are so definite by the 12th week that the diagnosis is seldom a problem.

Biological tests for pregnancy depend upon the production by the placenta (the temporary organ that develops in the womb for the nourishing of the embryo and the elimination of its wastes) of chorionic gonadotropin, an ovary-stimulating hormone. In practice, the tests have an accuracy of about 95 percent, although false negative tests may run as high as 20 percent in a series of cases. False negative reports are frequently obtained during late pregnancy when the secretion of chorionic gonadotropin normally decreases. The possibility not only of false negative but also of false positive tests makes the tests, at best, probable rather than absolute evidence of the presence or absence of pregnancy. Chorionic gonadotropin in a woman's blood or urine indicates only that she is harbouring living placental tissue. It does not tell anything about the condition of the fetus. In fact, the greatest production of chorionic gonadotropin occurs in certain placental abnormalities and disorders that can develop in the absence of a fetus.

Biological tests for pregnancy

Tests using immature mice (the Aschheim-Zondek test) and immature rats have been found to be extremely accurate. Tests using rabbits (the Friedman test) have been largely replaced by the more rapid and less expensive frog and toad tests.

The use of the female South African claw-toed tree toad, *Xenopus laevis*, is based on the discovery that this animal will ovulate and extrude visible eggs within a few hours after it has received an injection of a few millilitres of urine from a pregnant woman. The male common

frog, *Rana pipiens,* will extrude spermatozoa when treated in the same way. Both of these tests are considered somewhat unsatisfactory because false positive reactions are not uncommon.

Several immunological reaction tests in common use are based upon the inhibition of hemagglutination (clotting of red cells). A positive test is obtained when human chorionic gonadotropin (HCG) in the woman's urine or blood is added to human chorionic gonadotropin antiserum (rabbit blood serum containing antibodies to HCG) in the presence of particles (or red blood cells) coated with human chorionic gonadotropin. The hormone from the woman will inhibit the combination of coated particles and antibody, and agglutination does not occur. If there is no chorionic gonadotropin in her urine, agglutination will occur and the test is negative.

Several "signs" noted by the physician during an examination will suggest that a patient may be in the early months of pregnancy. Darkening of the areola of the breast (the small, coloured ring around the nipple) and prominence of the sebaceous glands around the nipple (Montgomery's glands); purplish-red discoloration of the vulvar, vaginal, and cervical tissues; softening of the cervix and of the lower part of the uterus and, of course, enlargement and softening of the uterus itself are suggestive, but not necessarily proof of pregnancy.

Conditions that may be mistaken for pregnancy. Other conditions may confuse the diagnosis of pregnancy. Absence of menstruation can be caused by chronic illness, by emotional or endocrine disturbances, by fear of pregnancy or by a desire to be pregnant. Nausea and vomiting may be of gastrointestinal or psychic origin. Tenderness of the breasts can be due to a hormonal disturbance.

Absence of menstruation; false positive tests

Any condition that causes pelvic congestion, such as pelvic tumour, may cause duskiness of the genital tissues. At times a soft tumour of the uterus may simulate a pregnancy. The question of pregnancy may be raised if the woman does not menstruate regularly; the absence of other symptoms and signs of gestation indicates that she is not pregnant. There are rare ovarian and uterine tumours that produce false positive pregnancy tests. It may be difficult for the physician to exclude pregnancy on the basis of his examination if the uterus is tipped back and difficult to feel, or if it is enlarged by a tumour within it. He is helped by the fact that the other signs of pregnancy are absent and that the tests for pregnancy are usually negative.

Childless women who greatly desire a baby sometimes suffer from false or spurious pregnancy (pseudocyesis). They stop menstruating, have morning nausea, "feel life," and have abdominal enlargement caused by fat and intestinal gas. At "term" they may have "labour pains." Signs of pregnancy are absent. Treatment is by psychotherapy.

Menopausal women often fear pregnancy when their periods stop; information that they show no signs of pregnancy usually reassures them. Retained uterine secretions of bloody or watery fluid, caught above a blocked mouth of the uterus (cervix), prevent menstruation, cause softening and enlargement of the uterus, and may cause the patient to wonder whether she is pregnant. There are no other signs of pregnancy, and the hard cervix, closed by scar tissue, explains the problem.

Duration of pregnancy. There are, as a rule, 266–270 days between ovulation and childbirth, with extremes of 250 and 285 days. Physicians usually determine the date of the estimated time for delivery by adding seven days to the first day of the last menstrual period and counting forward nine calendar months; *i.e.*, if the last period began on January 10, the date of delivery is October 17. This is approximately 280 days or ten lunar months. Courts of law, in determining the legitimacy of a child, may accept much shorter or much longer periods of gestation as being within the periods of possible duration of a pregnancy. One court in the state of New York has accepted a pregnancy of 355 days as legitimate. British courts have recognized 331 and 346 days as legitimate with the approval of medical consultants. Fully developed infants have been born as early as 221 days after the first day of the mother's last menstrual period.

Since the exact date of ovulation is usually not known, it is seldom possible to make an accurate estimate of the date of delivery. There is a 5 percent chance that a baby will be born on the exact date estimated from the above rule. There is a 25 percent chance that it will be born within four days before or after the estimated date. There is a 50 percent chance that delivery will occur on the estimated date plus or minus seven days. There is a 95 percent chance that the baby will be born within plus or minus 14 days of the estimated date of delivery.

ANATOMIC AND PHYSIOLOGIC CHANGES THAT ATTEND THE PROGRESS OF NORMAL PREGNANCY

Changes in organs and tissues directly associated with childbearing. *Ovaries.* The ovaries of a nonpregnant young woman who is in good health go through cyclic changes each month. These changes centre about a follicle, or "egg sac." A new follicle develops after each menstrual period, casts off an egg (ovulation), and, after ovulation, forms a new structure (the corpus luteum).

Ovulation and formation of corpus luteum; progesterone and estrogen

If the egg is fertilized, it is sustained for a short time by the hormones produced by the corpus luteum. Progesterone and estrogen, secreted by the corpus luteum, are essential for the preservation of the pregnancy during its early months. If pregnancy does not occur, the egg disintegrates and the corpus luteum shrinks. As it shrinks, the stimulating effect of its hormones, progesterone and estrogen, is withdrawn from the endometrium (the lining of the uterus), and menstruation occurs. The cycle then begins again.

Pregnancy, if it occurs, maintains the corpus luteum by means of the hormones produced by the young placenta. The corpus luteum is not essential in human pregnancy after the first few weeks because of the takeover of its functions by the placenta. In fact, human pregnancies have gone on undisturbed when the corpus luteum has been removed as early as the 41st day after conception. Gradually the placenta, or afterbirth, begins to elaborate progesterone and estrogen itself. By the 70th day of pregnancy the placenta is unquestionably able to replace the corpus luteum without endangering the pregnancy during the transfer of function. At the end of pregnancy the corpus luteum has usually regressed until it is no longer a prominent feature of the ovary.

During the first few months of pregnancy the ovary that contains the functioning corpus luteum is considerably larger than the other ovary. During pregnancy, both ovaries usually are studded with fluid-filled egg sacs as a result of chorionic gonadotropin stimulation; by the end of pregnancy, most of these follicles have gradually regressed and disappeared.

The blood supply to both ovaries is increased during pregnancy. Both glands frequently show plaques of bright red fleshy material on their surfaces, which, if examined microscopically, show the typical cellular change of pregnancy, called a decidual reaction. In this reaction, cells develop that look like the cells in the lining of the pregnant uterus. They result from the high hormone levels of pregnancy and disappear after the pregnancy terminates.

The uterus and the development of the placenta. The uterus is a thick-walled, pear-shaped organ measuring seven centimetres (about 2.75 inches) in length and weighing 30 grams (about one ounce) before pregnancy. It has a buttonlike lower end, the cervix, that merges with the bulbous larger portion, called the corpus. The corpus comprises approximately three-fourths of the uterus. There is a flat, triangular-shaped cavity within the uterus. At term, the uterus is a large, thin-walled, hollow, elastic, fluid-filled cylinder measuring approximately 30 centimetres (about one foot) in length, weighing approximately 1,200 grams (2.6 pounds), and having a capacity of 4,000 to 5,000 millilitres (4.2–5.3 quarts).

The uterus before and during pregnancy

The greater size of the uterus as a result of pregnancy is due to a marked increase in the number of muscle fibres, blood vessels, nerves, and lymphatic vessels in the uterine wall. There is also a five- to tenfold increase in the size of the individual muscle fibre and marked enlargement in the diameters of the blood and lymph vessels.

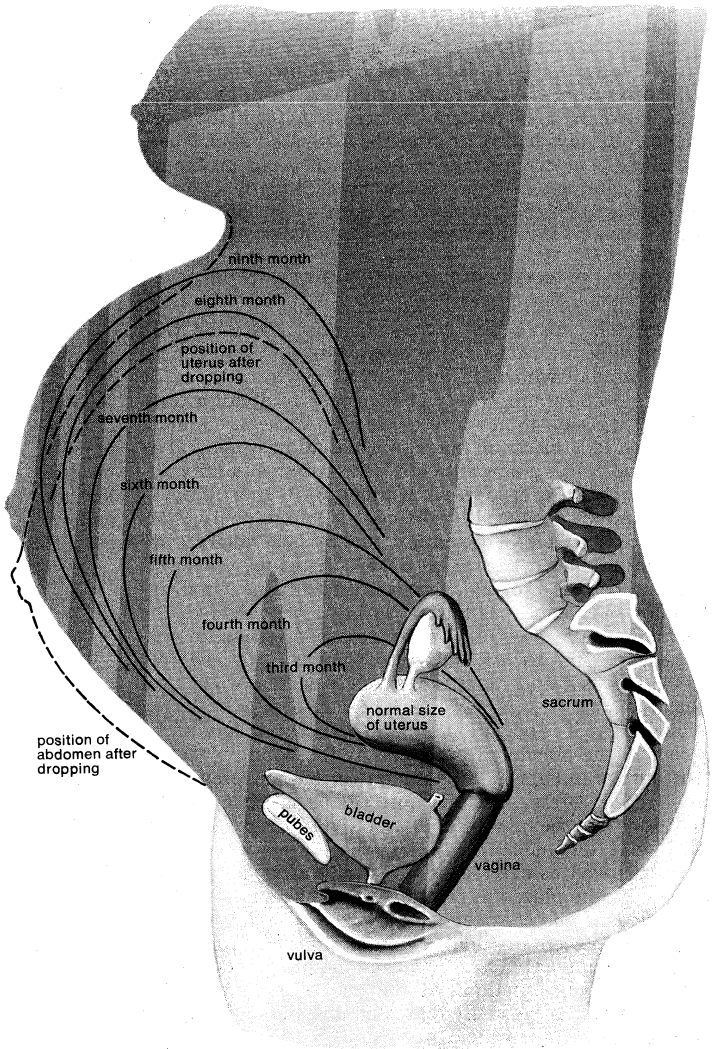


Figure 1: Levels of the uterus with advancing months of pregnancy.
From *Dorland's Illustrated Medical Dictionary*, 24th ed. (1965); W.B. Saunders Co., Philadelphia

During the first few weeks of pregnancy, the shape of the uterus is unchanged, but the organ becomes gradually softer. By the 14th week it forms a flattened or oblate spheroid. The fibrous cervix becomes remarkably softer and acquires a protective mucus plug within its cavity, but otherwise it changes little before labour. The lower part of the corpus, the isthmus, first becomes elongated and then, as the uterine contents demand more space, stretches and unfolds to form a bowl-shaped formation called the lower uterine segment. The fibrous nature of the cervix causes it to resist this unfolding action.

The uterine wall is stretched and thinned during pregnancy by the growing conceptus, as the whole product of conception is called, and by the fluid that surrounds it. By term, this process converts the uterus into an elastic, fluid-filled cylinder. It is only late in pregnancy that the cervix gradually thins out and softens; during labour it dilates for passage of the infant.

As pregnancy progresses, the uterus rises out of the pelvis and fills the abdominal cavity. It is top-heavy near term so that it falls forward and, because of the large bowel on the left side, rotates to the right. It presses on the diaphragm and pushes the other organs aside. The uterus may sink downward in the pelvis several weeks before term in a process that is known as lightening or dropping. This occurs as the fetal head descends into the pelvis. In some women, particularly those who have borne children, lightening does not occur until the onset of labour. Lightening may be impossible in women who have an abnormally small pelvis, an oversized fetus, or a fetus lying in an abnormal position.

For a short time after fertilization, the conceptus, a minute bubblelike structure called a blastocyst, lies unattached in the uterine cavity. The cells that will become the embryo (the embryonic disk) form a thickened layer on one side of the bubble. Elsewhere, the walls of the bubble consist of a single layer of cells; these cells are the trophoblast, which has a special ability to attach to and invade the uterine wall. The trophoblast plays an important role later in the development of the placenta or afterbirth. The conceptus makes contact with the uterine lining about the fifth or sixth day after conception. After contact the blastocyst collapses to form a rounded disk with the embryonic mass on the surface and the trophoblast against the endometrium (uterine lining). The part of the trophoblast that is in contact with the endometrium grows into and invades the maternal tissue. Concomitant disintegration of the endometrium allows the conceptus to sink into the uterine lining.

Implantation

Soon the entire blastocyst is buried in the endometrium. Proliferation of the trophoblast over the part of the collapsed bubble that is opposite the embryo is part of the implantation procedure that helps to cover the blastocyst. After a few days, a cavity forms that bears the same relation to the embryonic disk that the blastocyst cavity did before; this cavity will become the fluid-filled chorionic cavity containing the embryo. Ultimately it will contain the amniotic fluid that surrounds the fetus, the fetus itself, and the umbilical cord.

The chorionic cavity

The body stalk, which will become the umbilical cord, then begins to separate the embryo from the syncytiotrophoblast, the outer layer of the trophoblast lying against the endometrium; the inner lining of the trophoblast is called cytotrophoblast. As the syncytiotrophoblast advances into the endometrium, it surrounds minute branches of the uterine arteries that contain maternal blood. Erosion of the endometrium about these blood sinuses allows them to open into the small cavities in the trophoblast. The cytotrophoblast, which lines the cavity,

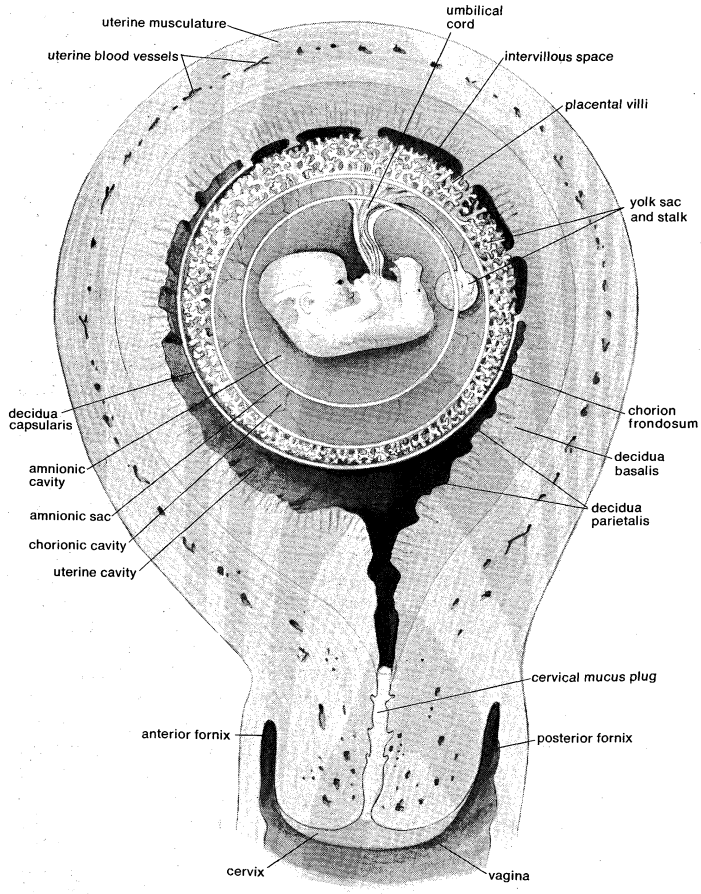


Figure 2: Sagittal section of the pregnant uterus at the seventh week of pregnancy showing the placenta and fetal membranes.

forms fingers of proliferating cells extending into the syncytiotrophoblast. After the placenta is developed, these fingers will be the cores of the rootlike placental villi, structures that will draw nutrients and oxygen from the maternal blood that bathes them. This is the first step in uteroplacental circulation, which supplies the fetus with all of the sustenance necessary for life and growth and removes waste products from it. During the third week of pregnancy, the syncytiotrophoblast forms a single layer of cells covering the growing villi and lining the syncytial lacunae or small cavities between the villi. The conceptus is buried in the endometrium, and its whole surface is covered at this time by developing villi. The greater part of the chorionic wall is now cytotrophoblast. Fingers of cytotrophoblast in the form of cell masses extend into the syncytial layer. Soon thereafter, a layer of connective tissue, or mesoderm, grows into the villi, which now form branches as they spread out into the blood-filled spaces in the endometrium adjacent to the conceptus.

By the end of the third week, the chorionic villi that form the outer surface of the chorionic sac are covered by a thick layer of cytotrophoblast and have a connective tissue core within which embryonic blood vessels are beginning to develop. The vessels, which arise from the yolk sac, connect with the primitive vascular system in the embryo. As growth progresses the layer of cytotrophoblast begins to regress. It disappears by the fifth month of pregnancy.

The layer of endometrium closest to the encroaching conceptus forms, with remnants of the invading syncytiotrophoblast, a thin plate of cells known as the decidua basalis, the maternal component of the mature placenta; it is cast off when the placenta is expelled. The fetal part of the placenta—the villi and their contained blood vessels—is separated from the decidua basalis by a lakelike body of fluid blood. This pool was created by coalescence of the intervillous spaces. The intervillous spaces in turn were formed from the syncytial lacunae in the young conceptus. Maternal blood enters this blood mass from the branches of the uterine arteries. The pool is drained by the uterine veins. It is so choked by intermingling villi and their branches that its continuity is lost on gross inspection.

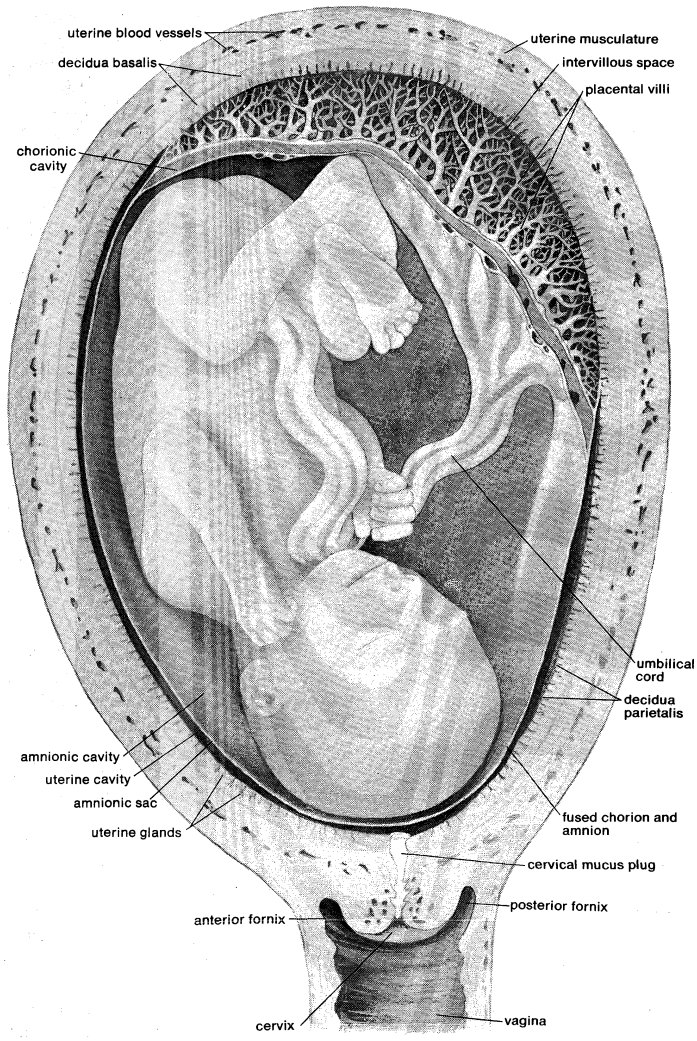


Figure 3: Sagittal section of the uterus at the fourth month of pregnancy.

The decidua capsularis, the chorion laeve, and the chorion frondosum

The chorionic cavity contains the fluid in which the embryo floats. As its shell or outer surface becomes larger, the decidua capsularis, which is that part of the endometrium that has grown over the side of the conceptus away from the embryo (*i.e.*, the abembryonic side) after implantation, becomes thinner. After 12 weeks or so, the villi on this side, which is the side directed toward the uterine cavity, disappear, leaving the smooth chorion, now called the chorion laeve. The chorion frondosum is that part of the conceptus that forms as the villi grow larger on the side of the chorionic shell next to the uterine wall. The discus-shaped placenta develops from the chorion frondosum and the decidua basalis.

At term, the normal placenta is a disk-shaped structure approximately 16 to 20 centimetres (about six to seven inches) in diameter, three or four centimetres (about 1.2–1.6 inches) in thickness at its thickest part, and weighing between 500 and 1,000 grams (1.1 and 2.2 pounds). It is thinner at its margins, where it is joined to the membrane-like chorion which spreads out over the whole inner surface of the uterus and contains the fetus and the amniotic fluid. The amnion, a thinner membrane, is adherent to and covers the inner surface of the chorion. The inner or fetal surface of the placenta is shiny, smooth, and traversed by a number of branching fetal blood vessels that come together at the point—usually the centre of the placenta—where the umbilical cord attaches. The maternal or uterine side of the placenta, covered by the thin, flaky decidua basalis, a cast-off part of the uterine lining, is rough and purplish-red, and has a raw appearance. When the placenta is cut across, its interior is seen to be made of a soft, crepelike or spongy matrix from which semisolid or clotted blood, caught when it is separated from the uterine wall to which it was attached, can be squeezed. Detailed examination shows that the villi and their branches form an arborescent (treelike) mass within the huge blood lake of the intervillous space. Anchoring villi extend outward from the fetal side and fuse with the decidua basalis to hold the organ's shape. Others, algae-like, float freely in the blood lake. Dividing partitions, formed from the trophoblast shell, project into the intervillous space from the decidual side. They divide the placenta into 15 or 20 compartments, which are called cotyledons.

Maternal blood flows from the uterine vessels into the trophoblast-lined intervillous blood lake. Within each villus is a blood vessel network that is part of the fetal circulatory system. Blood within the villous vessel is circulated by the fetal heart. The blood vessel wall, the connective tissue of the villous core, and the syncytiotrophoblast covering the villus lie between the fetal and the maternal bloodstreams. This is known as the placental barrier. As pregnancy progresses, the fetal blood vessels become larger, the connective tissue stretches over them, and the syncytiotrophoblastic layer becomes fragmentary. As a result, the placental barrier becomes much thinner. Normally, blood cells and bacteria do not pass through it, but nutrients, water, salt, viruses, hormones, and many other substances, including many drugs, can filter across it.

Uterine tube as site of fertilization

Uterine tubes. One of the two uterine tubes is the pathway down which the ripe ovum travels on its way to the uterus or womb. The spermatozoa from the male migrate up the tube, and it is there that they meet the ovum and fertilization occurs. During the first few days after fertilization the zygote, or fertilized egg, moves downward in the tube toward the uterus. While it is lying free in the tubal canal the young conceptus is nourished by secre-

tions from the tube. After the fertilized egg (or conceptus) passes into the uterus the tube ceases to play any part in the pregnancy; in fact, the only function the tube has is carried out during those few days before, during, and after conception. As pregnancy goes on the tube gradually enlarges, however, and contains more blood, as do all the pelvic organs; some of its cells may show a reaction, called a decidual reaction, to the hormones of pregnancy. As the uterus increases in size, the tubes stretch upward with it until they become two greatly enlarged elongated strands, one on each side of the uterus.

Vagina. The pinkish-tan colour of the lining of the vagina gradually takes on a bluish cast during the early months of pregnancy as a result of the dilation of the blood vessels in the vaginal wall; later the vaginal wall tends to become a purplish-red colour as the blood vessels become further engorged. The cells of the vaginal mucosa increase in size. Added numbers of these cells peel off the surface of the mucosa and mix with the increased vaginal fluid. This produces a profuse vaginal secretion. Thickening, softening, and relaxation of the loosely folded, succulent lining of the vagina and the sodden tissues beneath it greatly increase distensibility and capacity of the vaginal cavity; this is a process that partially prepares the birth canal for the passage through it of the large fetal mass.

External genital structures. Changes in the external genitalia are similar to those in the vagina. The tissues become first softened and more succulent and later extremely fragile, as an increasing amount of blood and fluid collects in them. They take on a purplish-red colour because of increased blood supply. Darkening of the vulvar skin, frequently seen during pregnancy, is particularly common among women of Mediterranean ethnic groups.

The blood vessels, muscles, and ligaments

Other pelvic tissues. The pelvic blood vessels and lymph channels become larger and longer. They develop new branches adequate to transport the greatly increased amounts of blood and tissue fluid that accumulate in the uterus and the other pelvic organs during pregnancy. Congestion and engorgement of blood in the pelvis, both within and without the uterus, are characteristic of pregnancy.

Changes in the muscles, ligaments, and other supporting tissues of the pelvis begin early in pregnancy and become progressively more pronounced as pregnancy continues. These changes are induced by the greatly increased hormonal levels in the mother's blood that characterize pregnancy. Before labour starts, the pelvic supporting tissues must have sufficient elasticity and strength to permit the uterus to grow out of the pelvis and yet support it. The muscles must be soft and elastic enough during delivery so that they can stretch apart and not obstruct the baby's birth. Softening and greater elasticity is brought about not only by the growth of new tissue but also by congestion and retained fluid within the tissues themselves.

The bones forming the mother's pelvis show relatively few changes during pregnancy. Loosening of the joint between the pubic bones in front and of the joints between the sacrum and the pelvis in back occurs as a response to the hormone called relaxin, which is produced by the ovary. Although relaxin, which causes marked separation of the pelvic joints in some animals, usually has too slight an effect in human beings to be noticed, softening of the attachments between the bones may be sufficient to cause a few women considerable distress. The strain on the joint between the sacrum and the spine becomes greater near term when the woman tilts her pelvis forward and bends the upper part of her body backward to compensate for the weight of the heavy uterus. When relaxation is excessive the woman suffers from backache and difficulty in walking. If it is extreme, she may have a waddling gait. Relaxation of the pelvic joints does not disappear quickly after delivery; it accounts for much of the backache complained of by women with new babies.

The mother's bones show no structural change if her calcium reserve and intake are normal. If her reserve and intake are not adequate, the fetus may draw so much calcium from her bones that the bones become soft and deformed. This condition is rarely seen, except in areas of the world where extreme poverty and serious calcium deficiency are major problems.

Changes in the breasts before and after delivery

Breasts. The earliest changes in the breasts in pregnancy are an exaggeration of the frequently experienced premenstrual discomfort and fullness. The sensation is so specific for pregnancy that many women who have been pregnant before are made aware of their condition by the feeling that they have in their breasts. As pregnancy progresses the breasts become larger, the lightly pigmented area (areola) around each nipple becomes first florid or dusky in colour and then appreciably darker; during the later months the areola takes on a hue that is deep bronze or brownish-black, depending on the woman's natural pigmentation. The veins beneath the skin over the breast become enlarged and more prominent. The small oily or sebaceous glands (glands of Montgomery) about the nipple become prominent.

These changes are due to the greatly increased levels of estrogen and progesterone in the woman's blood. These ovarian hormones also prepare the breast tissue for the action of the lactogenic (milk-causing) hormone, prolactin, produced by the pituitary gland. During the later part of pregnancy a milky fluid, colostrum, exudes from the ducts or can be expressed from them.

After delivery, the decrease in the estrogen and progesterone levels presumably permits the pituitary gland to release prolactin, which causes the breast to secrete milk. It is thought that the high hormonal levels inhibit the action or secretion of prolactin before delivery. Prolactin continues to be produced, and lactation usually continues, as long as the mother feeds her baby at the breast.

Anatomic and physiologic changes in other organs and tissues. *Cardiovascular and lymphatic systems.* During pregnancy the increasing needs of the growing fetus and of her own tissues throw an added burden on the mother's heart. The work that the heart does is measured by the amount of blood it expels per minute (the cardiac output). Rapid increase in the cardiac output occurs between the 9th and the 14th week of gestation. During the period from the 28th to the 30th week, when the load is heaviest, the heart of a pregnant woman is doing 25 to 30 percent more work than it was doing before pregnancy. As the time of delivery approaches, the heart's work load diminishes to some extent; when the baby is born the load is approximately equal to what it was when the mother was in the nonpregnant state. This decrease in cardiac output and cardiac work, which occurs in spite of the continued needs of the fetus and of the maternal tissues for blood-borne oxygen and nutriments, is explained by the more efficient way that the tissues draw on the mother's blood for oxygen and nourishment during the terminal weeks of pregnancy.

The position of the heart

The position of the heart is changed to a greater or lesser degree during pregnancy. As the uterus enlarges it elevates the diaphragm. This in turn pushes the heart upward, to the left, and somewhat forward, so that it is nearer the chest wall beneath the breast. Near the end of gestation the large uterus may raise the heart until the latter lies almost at a right angle to the long axis of the woman's body. These changes, which also bring some rotation of the heart, vary considerably in different individuals. When present to a marked degree, they may give an examining physician the erroneous impression that a normal heart is considerably enlarged. Actually, in spite of its greater work load, a healthy heart enlarges little or not at all even during the midportion of pregnancy, when the load is greatest.

Changes in the position of the heart, the greater work load, the increased volume of blood that the heart expels per beat, the decreased viscosity of the blood, and the larger amount of blood in the woman's blood vessels (discussed below), will, in many women, cause some distortion of the sounds that the physician hears when he listens to a patient's heart with his stethoscope. Such distorted sounds, called "functional" murmurs (as distinguished from "organic" murmurs, which may be present when the heart is diseased), do not indicate that anything is amiss, although they may be sufficiently atypical to

cause the obstetrician to refer the patient to a cardiologist for evaluation. Pregnancy sometimes produces minor changes in the electrocardiogram, but these changes are within normal limits.

Such is the ability of the heart to respond to an increased work load that even the pregnant woman with serious heart disease, given proper care and without an unexpected complication, will usually go through her pregnancy and delivery without a catastrophe. She is likely, however, to encounter disaster when she tries to cope with the stress of caring for her family after the baby is born.

Blood pressure and pulse rate

Normal pregnancy does not increase the mother's blood pressure. Indeed, a slight lowering of the blood pressure is commonly noted during the course of the pregnancy. Any notable rise in a pregnant woman's blood pressure is reason for alertness on the part of her physician, and if it continues to rise, for concern; it usually foretells the onset of toxemia (see below).

The pulse rate is a trifle more rapid during pregnancy, reflecting the more rapid heartbeat that is necessary in order to move the larger volume of blood present. The rate at which blood flows through the myriads of small blood vessels in the skin (the peripheral circulation) is accelerated during pregnancy, leading to the elevated skin temperature, the tendency to perspire, and, in part, to the redness of the palms and the tiny dilated blood vessels in some women as their pregnancies progress.

The most notable change in the circulatory system during pregnancy, other than those described in the heart, is a slowing of the blood flow in the lower extremities. With this decrease in the rate of flow there is an increase in the pressure within the veins and some stasis—stagnation—of the blood in the legs. These changes, which are believed to be caused primarily by the pressure of the uterus on the large blood vessels in the pelvis, are progressive during pregnancy and disappear after delivery. They also are thought to be caused in part by the marked increase in the amounts of the hormones estrogen and progesterone in the circulating blood. Increased venous pressure, slowing of the rate of venous flow, and partial stasis of the blood in the veins are major factors in causing the swelling of the legs and the varicose (abnormally dilated) veins of the lower legs that are commonly present near the end of pregnancy.

Edema in legs

The lymphatic vessels of the pregnant woman's pelvis become enlarged in response to the increased amount of tissue fluid in the engorged pelvic organs. As the uterus grows in size it presses on these channels, causing impairment of the lymphatic drainage from the woman's legs, with resultant swelling and distention of her feet and legs.

Although some fluid almost invariably collects in the feet, ankles, and legs near the time of delivery, sudden swelling of the feet and legs or a notable increase in swelling may be an early signal of impending toxemia, a serious disorder of pregnancy that is discussed below. Generalized swelling—*i.e.*, swelling of the hands, face, and other parts of the body—is a cause for serious concern.

Respiratory tract. One would expect that, as the uterus grows larger and pushes the diaphragm up, it would interfere with breathing, but the lungs actually work as efficiently as they do in the nonpregnant state. This is due to a change in the shape of the chest cavity during pregnancy; the chest diameter increases as its height decreases, so that there is actually a slight increase in the space that the lungs occupy.

The amount of air drawn in and expelled per minute by the lungs increases progressively during pregnancy. Immediately before delivery the number of breaths per minute is approximately twice what it is after the baby is born. This, like so many of the other changes in the mother's body, is an adaptation of one of her vital functions that is necessary to supply her tissues and those of the growing fetus with increasing amounts of oxygen.

Gastrointestinal tract. A number of alterations, often causing more or less distress, occur in the physical condition and functions of the gastrointestinal tract during pregnancy.

Taste and smell

Disturbances of the sensations of taste and smell, relatively common during early months of gestation, are often accompanied by a dislike of odours and a distaste of foods formerly found to be not disagreeable. The inflammation of the mouth and gums that some pregnant women complain of is more often caused by poor oral hygiene, by vitamin deficiencies, or by anemia than by the pregnancy itself.

Hydrochloric acid and pepsin, adequate amounts of which are necessary for satisfactory digestion, are produced by the stomach in decreased amounts during pregnancy. This decrease in the amount of acid in the stomach may explain some of the otherwise inexplicable anemias that occasionally occur during the course of an otherwise seemingly normal pregnancy.

During pregnancy the stomach muscles lose some of their tone and become more flabby, and the contractility of the stomach is reduced. As a result, the time it takes for the stomach to empty its contents into the intestinal tract is prolonged. As pregnancy progresses, the stomach is pushed upward; near term it lies like a flabby pouch across the top of the uterus instead of hanging downward, as it normally does, in a semivertical position. The loss of tone of the stomach muscles, the decrease in stomach acidity, and the change in position of the stomach are conducive to the flow of intestinal contents back into the stomach.

These disturbances in gastric function are responsible, in part at least, for the intolerance for fatty foods, the indigestion, the discomfort felt in the upper part of the abdomen, and the heartburn experienced by most pregnant women at some time during their pregnancies.

The musculature not only of the stomach but also of the entire intestinal tract loses much of its tonicity. As a result, peristalsis, the series of wave-like movements of the intestines, is slowed, the length of time it takes food to pass through the intestinal tract is prolonged, and there is more or less stagnation of the intestinal contents.

Intestinal problems

Constipation and hemorrhoids that cause rectal pain and bleeding are common complaints during pregnancy. The constipation is caused by lack of tone of the intestinal tract and stagnation of the bowel contents. Pregnant women may also lose the urge to defecate because of the pressure of the uterus on the lower bowel and inhibition of a reflex stimulus, known as the gastrocolic reflex, from the stomach to the rectum. The latter mechanism, which depends on normal stomach function, is responsible for the increased activity of the lower bowel that follows increased stomach activity, such as that induced by eating. It is this reflex that causes many persons to feel a desire to defecate within an hour or so after eating a full meal. Hemorrhoids—greatly enlarged or varicose veins in the lower rectum—that appear during pregnancy are due to constipation, to stasis of blood in the pelvic veins, and to pressure by the enlarging uterus on the blood vessels in the pelvis.

Liver. The liver, which plays an essential role in many of the vital processes—processes as diverse as participating in the metabolism of nutriments and vitamins and the elimination of the waste products of metabolism—changes anatomically and functionally during pregnancy to meet the added load placed on it by the maternal organism, the enlarging uterus, and, to a lesser extent, the growing fetus.

The liver's ability to synthesize proteins and to supply minerals and nutriments is augmented in response to the increased requirements of the mother's tissue and the fetus. The liver adjusts to the greatly augmented amounts of hormones circulating in the mother's blood during pregnancy. It helps to dispose of or detoxify the larger amounts of waste material produced by the metabolic processes in the growing fetus, the enlarging uterus, and the mother's tissues. Furthermore, the blood vessels in the liver enlarge to accommodate the larger amount of blood in the mother's blood vessels. At the same time, the liver must compensate for the larger number of circulating red blood cells.

In response to these demands the liver increases in size and weight, and its blood vessels become larger, but

otherwise its anatomical structure changes relatively little during pregnancy.

The hormones produced by the placenta and the metabolic changes in the maternal organism, rather than the fetus, are the factors responsible not only for the increased work the liver does but also for many of the physical and functional alterations that appear during gestation.

Changes in bladder function

Urinary tract. Changes that take place in the bladder and the urethra during pregnancy are attributable to relaxation of the muscles supporting these structures, to change in position, and to pressure.

The uterus lies over the bladder and presses upon it during early pregnancy. Later the uterus rises out of the pelvis. As the uterus grows larger and moves upward, the bladder is pushed forward and pulled upward. The urethra, the tube through which urine is discharged from the bladder, is stretched and distorted. As these distortions take place, the wall of the bladder becomes thickened, the blood vessels become enlarged, and fluid collects in the tissues forming the wall of the bladder. The results are swelling, stasis of blood in the blood vessels, and some mechanical inflammation of the bladder wall.

The woman is likely to urinate frequently during the early months of pregnancy when the heavy uterus presses on the bladder. Frequent urination is less common during midpregnancy, but it recurs after the baby descends into the pelvis near the time of delivery. As the bladder and urethra are pulled upward and distorted by the growing uterus, the stretched muscles that control urination are less efficient, and the woman may lose some urine involuntarily when she coughs, sneezes, or laughs; this is known as stress incontinence.

The swelling, mechanical inflammation, and stasis of blood in the blood vessels of the bladder near the end of pregnancy are conducive to bladder infection, a symptom of which is pain on urination. A microscopic examination of the urine is necessary to differentiate between the effect of pregnancy on bladder function and the symptoms caused by a bladder infection. An untreated bladder infection may lead to serious urinary tract troubles later.

Structure and functioning of ureters

Changes in the structure and function of the ureters, the two rubbery, spaghetti-like tubes that carry urine from the kidneys to the bladder, are present in 80 percent of all pregnancies. As pregnancy progresses each ureter becomes larger, so that it lies in multiple broad curves rather than forming an almost straight line downward from the kidney. In addition, both ureters, but particularly the right one, become greatly dilated, so that the urine flows very slowly or collects in them.

The funnel-like part of the kidney, called the kidney pelvis, also becomes dilated. With this dilation of the kidney pelvis and the ureters there is also a loss of tonicity or contractility in the pelvis of the kidney and the ureters. This loss of tonicity during pregnancy is similar to that mentioned in the description of the changes in the intestinal tract. Since it is the contractility of peristalsis within the ureter that propels urine downward from the kidney into the bladder, stasis of urine in the ureter is accentuated during the pregnancy. In the nonpregnant state the hydrostatic pressure in the kidney is greater than that in the bladder; during pregnancy the situation is reversed. This change of pressure further increases the stasis of urine in the ureter and kidney pelvis. As a result, bladder infections are more serious during pregnancy, because they are more likely to involve the kidney.

After delivery the ureters rapidly return to their normal condition.

Ordinary functioning of kidney; changes in pregnancy

The kidney of a healthy person selectively filters and secretes water, sodium, potassium, chlorides, protein, and other substances from the blood. It then reabsorbs water and essential elements in amounts that are needed to maintain the fluid, electrolytic, and other chemical balances in the body. It also filters waste products of metabolism from the blood and excretes them in the urine. During pregnancy the kidney continues to carry on these functions. The work load placed on it, however, is greater because of the increase in the amount of water and blood and in the rate of metabolism during gestation.

In early pregnancy, secretion of large amounts of dilute urine of decreased acidity, together with pressure of the uterus on the bladder, causes frequency of urination and nocturnal voiding. Less urine is excreted toward the end of pregnancy. The storage of large amounts of nitrogen, as part of the metabolism of proteins, causes a decrease in the urinary excretion of urea and of total nitrogen during gestation.

Although many healthy pregnant women occasionally show a trace of protein (albumin) in their urine, the detection of even small amounts of protein in the urine is a cause for alertness on the part of a physician, because anything more than an extremely small amount may be the first signal of impending toxemia or kidney disease, both of which are serious complications.

The kidney's ability to reabsorb sugar (glucose) is lower during pregnancy, and for this reason many pregnant women have transient periods during which their urine contains small amounts of glucose; such women have unimpaired ability to metabolize carbohydrates and have normal sugar levels in the blood. Glucose in the urine also may be the first sign that a person has diabetes mellitus, however; consequently, a pregnant woman whose urine contains traces of glucose is tested to make sure that she can metabolize sugar normally.

The preceding discussion of kidney function illustrates the need for a pregnant woman to be under a physician's care, an essential part of which is periodic examination of her urine for protein, sugar, pus, bacteria, and other abnormal constituents.

Increase in blood volume

Blood. The total amount of blood in a pregnant woman's body has increased by approximately 25 percent by the time of delivery. The increase is accounted for by the augmented volume of blood plasma (the liquid part of the blood), which is caused by fluid retention, plus an increase in the total number of red blood cells. Additional blood is needed to fill the large vessels of the uterus. Also, more blood is required to carry the oxygen and nutriments needed by the fetus and the maternal tissues and to carry away waste products. Furthermore, it is a protective reserve in case of hemorrhage during delivery.

During pregnancy the blood-forming organs, such as the bone marrow, make more erythrocytes, or red blood cells, which carry iron and oxygen. Despite this, there is usually a decrease in a pregnant woman's blood cell count—the number of red cells per cubic millimetre of blood—because the amount of blood plasma increases approximately 30 percent, while the total number of red blood cells increases by only about 20 percent. This results in apparent anemia. With these changes, the viscosity of the blood decreases and the hematocrit, which measures the relative amounts of liquid and solid constituents in the blood, is lower. Usually there is a moderate increase in the number of white blood cells per cubic millimetre during early pregnancy; this increase disappears during the latter part of pregnancy.

If a pregnant woman is otherwise healthy and receives adequate available iron for the production of hemoglobin, her red blood cell count does not ordinarily fall below 3,750,000 cells per cubic millimetre, her hemoglobin below 13.5 grams per 100 cubic millimetres of blood, and her hematocrit below 35. (Normal values for nonpregnant women are 4,200,000–5,400,000 cells, 13.8–14.2 grams hemoglobin, and 37–47 hematocrit.) Physicians usually make blood counts for their pregnant patients every two months because of the need for repeated evaluation.

Endocrine system. Most of the endocrine glands become larger, and some display alterations in function, during pregnancy; they all revert to a normal state after delivery.

Pituitary, thyroid, pancreas, and adrenals

The anterior lobe of the pituitary gland increases in size during pregnancy, but the production of pituitary gonadotropins, the gonad-stimulating hormones, ceases soon after the placenta begins to produce chorionic gonadotropins. The pituitary continues to secrete the hormones that stimulate the other endocrine glands. Near term, as the mother's estrogen level drops, a milk-stimulating hormone, prolactin, is produced by the pituitary. The poste-

rior lobe of the pituitary gland does not change in size or weight during pregnancy.

The thyroid gland enlarges moderately, but there is no true increase in thyroid function during gestation. The parathyroid glands also increase in size during pregnancy but presumably are not otherwise affected by it.

The part of the pancreas that secretes insulin, the islets of Langerhans, becomes larger. Whatever increase in function is displayed may be assumed to be a balanced response to the body's demand for the products of carbohydrate metabolism. The level of plasma insulin or of insulin-like substances in the plasma is higher during pregnancy, and the destruction of insulin is also more rapid.

The blood and urinary levels of 17-hydroxycorticosteroids, hormones that affect protein, fat, and carbohydrate metabolism and that are produced by the adrenal glands, rise during pregnancy; but there is no increased effect from the hormones, because their higher level is more than offset by the increased levels of transcortin, a protein that inactivates them.

As gestation progresses, there is an elevation in the secretion of aldosterone, an adrenal hormone that plays a role in the retention of salt and water in the body. It has been suggested that this is a protective mechanism to counterbalance the tendency for progesterone to cause the excretion of sodium ions in the urine.

Changes in sweat glands and skin colour

Skin. Pregnancy usually causes an increase in the secretion of the oil and sweat glands in the skin. Body odours may become more pronounced. Many women notice that their hair becomes thinner and drier and their nails more brittle. Others may develop an increased amount of facial and body hair. The "mask of pregnancy" seen particularly in brunettes is a deposit of brownish pigment in the skin of the forehead, the cheeks, and the nose. Puffiness and thickening of her skin may cause the pregnant woman's face to appear coarse and almost masculine. Increased pigmentation, particularly of the smooth skin about the nipples (the areolas of the breasts) and the vulva, is almost universal.

Bright red discoloration of the palms of the hands and tiny spiderweb-like red blood vessels in the skin of the arms or face are not unusual during pregnancy. Many of these changes are thought to be associated with the greatly increased levels of estrogen in the mother's bloodstream. Most of the changes disappear after delivery.

"Stretch marks," which appear on the breasts and abdomen during pregnancy, are due to the tearing of the elastic tissues in the skin that accompanies enlargement of the breasts, distention of the abdomen, and the deposition of subcutaneous fat. They are pink or purplish-red lines during pregnancy. The lines become permanent grayish-white scarlike marks after delivery. Some women never develop stretch marks despite bearing several children; others lose most of the tone in their skin after one pregnancy. Stretch marks cannot be considered evidence that a woman has borne a child, however, because they sometimes are seen in women who have not been pregnant.

Metabolic changes. Metabolic changes during pregnancy are among the many adjustments that the mother's organs make to meet the requirements created by the increase in her own breast and genital tissues and the growth of the conceptus (the fetus and afterbirth). In addition, reserves must be established to meet the demands that will be put on her body during pregnancy, delivery, and the lying-in period.

The basal metabolic rate. The amount of oxygen consumed is an index of the pregnant woman's metabolism when she is at rest—her basal metabolism. The rate begins to rise during the third month of pregnancy and may double the normal rate (+10 percent) by the time of delivery. The rate rises in specific proportion to the size of the fetus and represents the effects of the mother's activities plus those of the fetus and the uterine structures. An elevation of the basal metabolic rate (BMR) to 20 or 25 percent during pregnancy is not an indication of an overly active thyroid gland.

Weight. The early part of pregnancy usually is accompanied by moderate weight loss caused by the woman's lack of appetite and her nausea and vomiting. Between the third and the ninth month of pregnancy most women gain about 20 pounds or more. Ideally, during pregnancy body weight is gained at the rate of about one pound per week for a total of not more than 20 to 25 pounds. In an average pregnancy the infant, the afterbirth, and the fluid in the uterus weigh about ten pounds. The uterus and the breasts together weigh approximately five pounds. The remaining five pounds consist of stored fluids and fat. Weight gain exceeding 25 pounds usually represents fat and fluids that are in excess of the reserve requirements for a normal pregnancy. A woman loses approximately 15 pounds at delivery, and another five pounds of stored fluid are eliminated as the uterus shrinks. She does not lose many additional pounds during the weeks following the delivery of the baby unless she limits her caloric intake. Fat stored during pregnancy is lost more slowly than stored fluids, proteins, and carbohydrates.

Excessive weight gain during pregnancy is a matter of concern for both the patient and the doctor. Although it may be only the result of overeating, it may be caused by a disturbance in metabolism and by an abnormal retention of fluids and salts. In the latter instance it may be the first sign of toxemia.

Changes in protein, carbohydrate, and fat metabolism

Protein. During pregnancy, nitrogen, derived from the metabolism of ingested protein, is needed for growth of the fetus, the placenta, the uterus, and the mother's breasts and other tissues. A considerable amount of nitrogen also is required for the increase in the mother's red cell volume and blood plasma. The fetus's demand for nitrogen is slight at first, but during the last month of pregnancy it acquires almost half of its total protein. In the process of accumulating this store and of building a reserve for the period after delivery, the woman who is on an adequate diet retains between two and three grams of nitrogen daily during her pregnancy; by term she and the fetus will have acquired approximately 500 grams (about 1.1 pounds) of nitrogen.

Carbohydrates. During pregnancy greater quantities of blood are being processed through the kidneys, but the kidneys are incapable of reabsorbing increased amounts of sugar. Consequently, a lower level of sugar in the blood is tolerated, and slight amounts of sugar are excreted in the urine. During pregnancy the level of sugar in the blood after fasting is slightly lower, probably because there is less utilizable insulin in the blood to regulate the sugar metabolism. Oral glucose tolerance tests show a prolonged elevation of blood sugar after ingestion of glucose; this may be an indication that carbohydrate utilization is less rapid or that the absorption of glucose from the gastrointestinal tract is slower. Glucose tolerance tests that depend on injection of the sugar solution into the veins show no difference between nonpregnant and pregnant nondiabetic women. A few women who are potential or mild diabetics demonstrate diabetes for the first time when they are pregnant. This is because pregnancy taxes insulin productivity in women with a marginal pancreatic islet reserve, so that diabetes may first become evident during gestation.

Fat. The total blood lipids average 600 to 700 milligrams per hundred millilitres of blood in the nonpregnant woman. They increase to approximately 900 to 1,000 milligrams per hundred millilitres of blood during the latter part of pregnancy. This increase, which involves all the lipid fractions, has not been explained, but it is worthy of notice that the gain in fat reaches its acme during the period that the fetus acquires most of its adipose (fatty) tissue.

Increases in body fluids

Water. Pregnancy is characterized by increases in the amount of body water and in the total volume of body fluid. During pregnancy between 3,500 and 4,000 millilitres of fluid (about 3.2 to 3.6 quarts) will be added to that already present in the tissues of a healthy woman. The uterus, the placenta, the amniotic fluid, and the fetus each account for approximately equal amounts. In addition to the water that makes the increased blood volume, there is also added fluid in the mother's muscles, her pelvic soft tissues, her breasts, and her other tissues.

Toward the end of pregnancy a considerable amount of

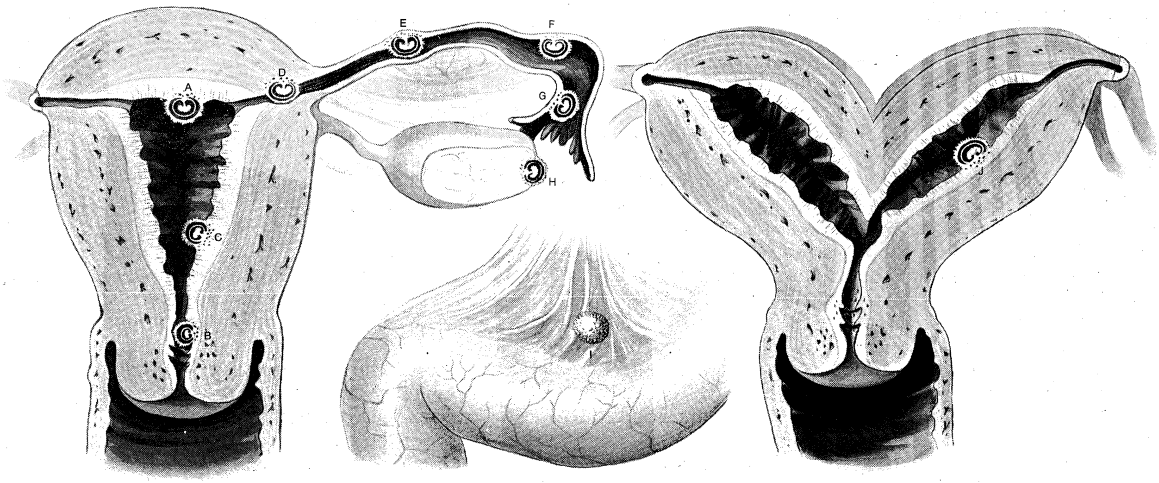


Figure 4: *Sites of implantation.*
(A) Normally in the upper part of the corpus. (B) In the lower corpus with later development of a placenta praevia. (C) Low implantation. (D) In the interstitial portion of the uterine tube. (E) In the isthmic part of the tube. (F) In the ampulla of the tube. (G) In the infundibulum of the tube. (H) In the ovary. (I) On the peritoneum with development of an abdominal pregnancy. (J) In the cornu of a double or rudimentary horn of a uterus.

retained fluid accumulates in the woman's lower extremities. It is this fluid that produces the pitting and swelling of the legs that many normally pregnant women display during the month or two before delivery.

Retention of large amounts of electrolytes, particularly sodium, accompanies the increase in the amount of body fluids. Approximately 12 grams of sodium are retained monthly. In addition to a positive sodium balance, there is a positive chloride and potassium balance during pregnancy. As a result, additional water is required to maintain the balance of the solution of sodium, chloride, and potassium in the blood, in the fluid of the spaces between the tissue cells, and within the cells themselves. Not all of the sodium, however, goes into fluid. Some of it is stored, and some replaces potassium in the cells.

A number of factors contribute to a positive sodium balance, which in turn leads to retention of fluid; these include alterations in the kidneys' excretion of sodium and water; increased retention of water in the pregnant woman's legs; the large amounts of hormones, particularly estrogen, that the placenta secretes; and the secretion of adrenal hormones, especially aldosterone. The latter, in particular, reduces the kidneys' secretion of sodium. Since sodium and water interact with each other, whatever contributes to the retention of one leads to the retention of the other. Generalized swelling appears when the accumulation of sodium and water becomes too great. If the excess continues the pregnant woman eventually develops toxemia.

Minerals. The pregnant woman's reserves and intake of iron and calcium must be enough not only for her own needs but also for those of the fetus. An increase in serum copper levels occurs during pregnancy. The mother has some phosphorus reserve but must acquire enough from her diet to supply her own tissues and those of the fetus. The utilization of phosphorus and that of calcium are interdependent, so that the utilization of phosphorus depends on the calcium intake.

II. Abnormal changes in pregnancy

ECTOPIC PREGNANCY

An ectopic pregnancy is one in which the conceptus (the products of conception; *i.e.*, the placenta, the membranes, and the embryo) implants or attaches itself in a place other than the normal location in the lining of the upper uterine cavity. The site of implantation may be either at an abnormal location within the uterus itself or in an area outside the uterus. Ectopic pregnancies outside the uterine cavity occur about once in every 300 pregnancies. They are one of the major causes of maternal deaths.

Normally an ovum or egg passes from the ovary into the tube, is fertilized in the tube, and moves downward into the uterus. It buries itself in the lining of the upper part of the uterine cavity. It may pass farther down and attach itself to the lining of the mouth of the uterus (the cervix), creating a cervical pregnancy. These are rare and cause severe vaginal bleeding; the conceptus is expelled or discovered within a few months after implantation.

Types of ectopic pregnancy

If a conceptus attaches itself to the lower part of the uterine cavity it is a low implantation. When a low implantation occurs the placenta grows over the cervical opening, in a formation called a placenta praevia. This causes the woman to bleed, often profusely, through the vagina, because the placenta tears as the cervix begins to open during the latter part of pregnancy.

When the fertilized egg implants in the narrow space or angle of the uterine cavity near the connection of the uterus with the fallopian tube, it is called an angular pregnancy; many angular pregnancies terminate in abortions; others go to term but are complicated because the placenta does not separate properly from the uterine wall after birth of the baby. An angular pregnancy differs from a cornual pregnancy, which develops in the side of a bilobed or bicornate uterus.

Implantation in the narrow part of the fallopian, or uterine, tube, which lies within the uterine wall, produces what is called an interstitial pregnancy. This occurs in approximately 4 percent of ectopic pregnancies. An interstitial pregnancy gradually stretches the wall of the uterus until, usually between the 16th and 20th week of gestation, the wall ruptures in an explosive manner and the woman bleeds profusely into her abdomen.

Most persons associate ectopic pregnancies with tubal pregnancies, because most ectopic pregnancies occur in the uterine tubes. The tube beyond the uterus has three parts: the isthmus, a narrow section near the uterus; the ampulla, which is wider and more dilatable; and the infundibulum, the flaring, trumpet-like portion of the tube nearest the ovary. A tubal ectopic pregnancy is designated by the area of the tube in which it is implanted. An isthmic pregnancy differs from one in the ampulla or infundibulum because the narrow tube cannot expand. Rupture of the affected tube with profuse intra-abdominal hemorrhage occurs early, usually within eight weeks after conception.

Ampullar pregnancies, which are by far the most common, usually terminate either in a tubal abortion, in which the embryo and the developing afterbirth are expelled through the open end of the tube into the abdomen; by a tubal rupture; or, less commonly, by absorption of the conceptus.

Sometimes the tube ruptures into the tissues attaching it to the wall of the pelvis, producing an intraligamentous pregnancy. Rarely, the embryo is expelled into the abdo-

men and the afterbirth remains attached to the tube; the embryo lives and grows. Such a condition is referred to as a secondary abdominal pregnancy. Primary abdominal pregnancies, in which the fertilized egg attaches to an abdominal organ, and ovarian pregnancies are rarer still.

Causes of tubal pregnancies

It is generally believed, but not proved, that most tubal pregnancies are caused by scars, pockets, kinks, or adhesions in the tubal lining resulting from tubal infections. The infection may have been gonorrhea; it may have occurred after an abortion, after the delivery of a baby, or after a pelvic surgical operation; or it may have been caused by appendicitis. Kinking, scarring, and partial adhesions of the outside of the tube may be the result of inflammation following a pelvic operation or of peritonitis or other abdominal inflammation. Tubes, defective from birth, may be too small for the passage of the conceptus or may be pocketed or doubled with one tubal half forming a blind pocket. There may be areas in the tubal lining that behave like the lining of the uterus (they show a decidual reaction that is conducive to implantation) so that they offer a favourable spot for the fertilized egg to implant. Pelvic tumours may distort the tube or press on it and obstruct it so that the conceptus cannot move downward. Theoretically, endocrine disturbances may delay tubal motility.

Whatever the cause, when a tubal implantation occurs it may be assumed that either migration of the fertilized egg within the tube was delayed by an extrinsic factor so that the egg grew to the point where it should implant or that the mechanism for implantation within the egg itself was prematurely activated in the tube. One or the other of these causative factors can sometimes be seen when a woman is operated upon for an ectopic pregnancy. There are a great number of cases, however, in which no abnormality of the tube can be found. There is no satisfactory explanation for most abnormal implantations in the uterus, although defective uterine structure has been noted in some cases.

Primary abdominal and ovarian pregnancies can best be explained by a mechanism in which the fertilized ovum is swept out of the tube by a reverse peristalsis of the tube, but it is quite possible that, in rare instances, the ovum and spermatozoa meet and fertilization and implantation take place within the abdomen.

Signs of ectopic pregnancy

Ectopic pregnancy is frequently mistaken for other disorders, and other disorders are frequently mistaken for it. Typically, but not invariably, the woman who has an ectopic pregnancy in the ampullar part of the tube will have missed one or two menstrual periods. She need not have other symptoms of pregnancy. She has felt enough discomfort in the lower part of her abdomen to lead her to consult a physician. She has had recurrent episodes of rather light, irregular bleeding from the vagina. She has felt weak or faint at times. The signs of pregnancy are not likely to be present, and results of a pregnancy test are more often negative than positive. The physician, on pelvic examination, feels a tender, soft mass in one side of the pelvis. At this stage he has to differentiate between an ectopic pregnancy and an intra-uterine pregnancy with abortion; acute appendicitis; intestinal colic; inflammation of a fallopian tube; and a twisted ovarian tumour. Unless the diagnosis can be made, the patient continues to complain for several more days and then has a sudden severe pain and collapses from brisk bleeding within the abdomen.

Only rarely are sudden acute abdominal pain and collapse due to severe hemorrhage the first signs that something is amiss. If this does happen it is usually because implantation has occurred in the isthmic portion of the tube and hemorrhage and tubal rupture occur simultaneously. More frequently a woman has missed one menstrual period, has a sensation of pelvic pressure, feels that she must urinate and collapses in the bathroom. She may be unconscious and pulseless from loss of blood when she arrives at the hospital.

Interstitial pregnancies are often mistaken for intra-uterine ones, but the patient has more pain than she would with a normal pregnancy and may have intermittent vaginal bleeding. After several months she has sudden severe pain, collapses from a massive intra-abdominal hemorrhage, and may die before surgical help can reach her. Most of the women who die from ectopic pregnancies do so from interstitial ones.

Combined pregnancies, in which there is an ectopic pregnancy and a normal one in the uterus, or a fetus in each tube, have occurred and have compounded the difficulty in making a diagnosis. In a number of instances the ectopic conceptus has been removed without complications, and the uterine fetus has progressed to term.

Not all ectopic pregnancies end with a catastrophic hemorrhage and collapse. In a few instances tubal, abdominal, and broad ligament pregnancies have gone on until a living baby was obtained at the time of operation. In other cases the fetus died and, if very young, was resorbed; in others, when the fetus was larger, death was followed by absorption of the fluid in the sac and the fetus was gradually converted into a more or less mummified mass. Some ectopic pregnancies of this type have caused no symptoms and have been carried by women for years. Undoubtedly many ectopic pregnancies that are in an early stage when they are expelled emerge through the open end of the uterine tube, are resorbed, and are never recognized.

Once diagnosed, the treatment of ectopic pregnancies outside the uterine cavity is almost always a matter of prompt surgical intervention with proper attention to replacement of blood and fluid.

ABORTION

What abortion is

Abortion is the termination of a pregnancy before the infant can survive outside the uterus. The age at which a fetus is considered viable has not been completely agreed upon. Many obstetricians use either 21 weeks or 400–500 grams (0.9–1.1 pounds) birth weight as the base line between abortion and premature delivery, because few infants have survived when they weighed less than 500 grams at birth or when the pregnancy was of less than 21 weeks' duration. Generally speaking, the fetus has almost no chance of living if it weighs less than 1,000 grams (2.2 pounds) and if the pregnancy is of less than 28 weeks' duration. In one effort to resolve the matter, the American College of Obstetricians and Gynecologists has defined abortion as the expulsion or extraction of all (complete) or any part (incomplete) of the placenta or membranes, with or without an abortus, before the 20th week (before 134 days) of gestation. Early abortion is an abortion that occurs before the 12th completed week of gestation (84 days); late abortion is an abortion that occurs after the 12th completed week but before the beginning of the 20th week of gestation (85–134 days).

In the past the word abortion has usually meant to nonmedical persons the criminal interruption of a pregnancy, whereas "miscarriage" indicated that there had been a spontaneous expulsion of the uterine contents. The term miscarriage is seldom used medically.

Spontaneous abortion is the expulsion of the products of conception before the 20th week of gestation without deliberate interference. As a general rule, natural causes are responsible for loss of the pregnancy. An induced abortion is the deliberate interruption of a pregnancy by any means before the 20th week of gestation. It may be therapeutic or nontherapeutic. A therapeutic abortion is the interruption of a pregnancy before the 20th week of gestation for legally acceptable, medically approved indications. This term may be used when a pregnancy is terminated because it endangers the mother's life or health or because the baby presumably would not be normal. A nontherapeutic abortion is the interruption of a pregnancy for nonmedical reasons by any means before the 20th week of gestation. The terms criminal abortion and illegal abortion are legal rather than medical terms.

Stages of spontaneous abortion

A spontaneous abortion usually passes through several progressive stages. The first stage is a threatened abortion in which a woman, known to be less than 20 weeks pregnant, notices a small amount of bloody discharge from her vagina and, perhaps, a few cramping pains in her uterus. By pelvic examination it is determined that her cervix has not started to open or dilate. Either the symp-

toms subside or the matter progresses to an inevitable abortion, in which there is increased bleeding, the uterine cramps become more severe, and the cervix, or mouth of the uterus, opens for the expulsion of the uterine contents. An inevitable abortion terminates either as a complete or an incomplete abortion, depending on whether or not all of the products of gestation are expelled. The process may start abruptly with pain and profuse bleeding and be over in a few hours, or it may go on for days with only a modest loss of blood. Spontaneous abortions early in pregnancy tend to be complete. When the pregnancy is further advanced it is more likely to be incomplete. Usually the physician removes the retained tissue in the uterus surgically when there is an incomplete abortion.

If the fetus dies and is retained in the uterus for eight weeks or longer the condition is referred to as a missed abortion. Women who lose three or more consecutive pregnancies of less than 20 weeks' duration are said to suffer from habitual abortion. An infected abortion is an abortion associated with infection of the genital organs.

Approximately 10 percent of all pregnancies terminate in spontaneous abortion. Some are lost so early that the woman and her physician are not sure whether she aborted or had a menstrual period that was slightly delayed, particularly heavy, and more painful than usual. The majority occur between the 6th and the 12th week after conception. The number of nontherapeutic abortions is unknown, but it probably equals the number of spontaneous ones. Modifications in the abortion laws in several countries, including the United States, have greatly increased the number of requested abortions; it is believed that in some areas the number of abortions exceeds that of babies delivered alive.

At least half of all early abortions are caused by defective development of the structures that grow from the fertilized egg; *i.e.*, the baby, the afterbirth, and the membranes that surround the baby and hold the water that it floats in. Some of these abortions are genetic accidents caused by abnormal characteristics carried in the egg or sperm or by the failure of normal development of the chromosomes after the egg and sperm unite. It has been shown in animals that disturbances in the transportation of the fertilized egg to the uterus may cause premature or delayed implantation of the conceptus; too young or too old fertilized eggs tend to abort. Inadequate secretion of the ovarian hormones estrogen and progesterone, needed for the development of the newly fertilized egg, may cause failure of the lining of the uterus and its secretions to sustain the young embryo. Later, failure of the placenta to take over the hormone-producing function of the ovary may adversely affect the growth of the uterus and its contractility. X-rays in large doses, radium, and certain drugs may cause abortion because they damage embryonic tissues. Abnormal development of the mother's uterus may make it impossible for it to retain the pregnancy. Late abortion is sometimes caused by the weakness of the mouth of the uterus or by fetal death following knotting of the umbilical cord. Uterine tumours may cause abortion because they increase uterine irritability or create an unfavourable environment for embryonic growth. In most instances in which psychic factors allegedly caused an abortion, examination of the baby and of the afterbirth have shown defects in one or both that had occurred before the mother had suffered her emotional disturbance. Physical injury to the mother is a causative factor in only one in a thousand abortions. Abortions thought to be caused by automobile accidents, falls, kicks, and so forth are often the result of deleterious changes in the fetus and sac that occurred before the injury. Systemic diseases may play a role in causing an abortion. This is particularly true of acute infectious diseases with high fever and bacteria in the bloodstream, or of diseases such as pneumonia, in which there is a marked reduction in the supply of oxygen to the fetus. Heart disease, kidney disease, diabetes, high blood pressure, and other chronic diseases may be associated with premature birth and fetal death after the 21st week but do not ordinarily cause abortions.

Causes of abortion

Perhaps 3 percent of threatened abortions are prevented by rest and hormonal therapy. Most abortions are inevitable because the fertilized egg is abnormal; these cannot be controlled medically. Many women who suffer from habitual abortion respond well to treatment; in some of these cases corrective surgery is necessary. An early spontaneous abortion without infection is rarely followed by ill health when the affected person is properly treated; but no abortion, regardless of where it is performed or by whom, is without hazard to the woman's health and life. Infected abortions, many the result of nontherapeutic interruptions of pregnancy, have caused sterility, prolonged suffering, and chronic pelvic distress; an appreciable loss of life has resulted from such infections.

Avoidance of abortion; effects of abortion

EFFECTS OF SYSTEMIC DISEASES ON PREGNANCY AND OF PREGNANCY ON SYSTEMIC DISEASES

Heart disease. Heart disease occurs in approximately 1 percent of pregnant women. It is first as a cause of maternal deaths among the disorders incidental to pregnancy and fourth, after hemorrhage, toxemia, and infection, as a cause of all maternal deaths. Rheumatic heart disease is the most prevalent type. Congenital heart disease accounts for approximately one-fourth of the cases.

A number of factors, including her response to physical activity, a history of heart failure, the type of heart disease that she has, and her age, are used in estimating how a particular woman will fare during pregnancy and labour. A person whose heart disease causes no limitation of normal physical activity will usually go through a normal pregnancy and delivery without notable difficulty, provided that she avoids undue physical activity, has sufficient rest, avoids infections, and is under the care of an obstetrician and a cardiologist who are on the alert for signs of early heart failure. Such a person will not face an appreciably increased risk, and her heart disease will not be affected by the pregnancy.

The woman whose physical activity is limited to some extent because it causes undue fatigue, shortness of breath, heart palpitation, or heart pain, but who has never experienced heart failure, will seldom suffer heart failure if she follows a strict regimen outlined by her physicians throughout the pregnancy, the labour, and the puerperium (the period immediately after childbirth) and if she does not experience a complication of pregnancy or of her heart disease. A diseased heart, although able to carry the load put on it by pregnancy, may not be able to stand up under an additional burden. This is particularly true if the pregnant woman gains an excessive amount of weight; if she develops toxemia, kidney disease, pulmonary disease, or an infection; or if she overworks physically, is subjected to sudden severe emotional stress, or becomes anemic. The possibility that a woman with serious heart disease will have heart failure is greater if she is over 35 years of age. A number of persons with heart disease who get along well during pregnancy with good care do suffer heart failure when they take on the added labours of caring for their family, home, and the new baby.

Over half of the women who have suffered from heart failure before they become pregnant do so again during pregnancy, usually between the fifth and the ninth month, when pregnancy throws the greatest work load on the heart. Because so many women with a history of previous heart failure have difficulties during pregnancy, many obstetricians and cardiologists restrict the physical activity of such women and try to keep them in the hospital and under close medical surveillance. Some women with serious heart disease are kept in bed in the hospital throughout the course of the pregnancy and thus avoid heart failure. Cardiac surgery during the first few months of pregnancy, although a hazardous procedure, has lessened the necessity for prolonged bed rest in some cases and materially improved the prognosis in others.

Management of pregnancy preceded by heart failure

Women with serious heart disease often deliver prematurely, and their labours are often short and their deliveries easy. There is an increase in fetal mortality because many pregnancies are interrupted and because many of the babies of women with heart disease are born prematurely. Babies who are not born prematurely are not notably different from those of normal mothers.

Endocrine diseases. *Diabetes.* Before insulin was available, most diabetic women were sterile, or, if they became pregnant, aborted. Half of the babies and one-fourth of the mothers died if they went to term. Today, if they are adequately supervised, less than 1 percent of pregnant diabetic women die of diabetes during pregnancy or the puerperium. Diabetic women do suffer from an increased incidence of toxemia, infections, and polyhydramnios (excessive amniotic fluid). Abnormalities of labour are increased because the babies tend to be unusually large, and congenital abnormalities of the fetus are more common, as is polyhydramnios; polyhydramnios is a problem in 25 percent or more of diabetic women.

Complications from diabetes

Untreated diabetes is associated with a high incidence of fetal defects, abortion, stillbirths, premature labour, and excessively large babies. Even with diet and insulin, over 50 percent of the babies delivered by diabetic women weigh over eight pounds at birth. Even though they appear healthy at birth, many of them are not as strong as smaller babies whose mothers are not diabetic. Fetal loss is greater if the mother became diabetic in childhood, if she has been diabetic for a long time, or if she has vascular or kidney disease. Toxemia, which occurs with approximately 30 percent of diabetic pregnancies, would be more common if it were not for the practice of terminating such pregnancies several weeks before term.

Pregnancy frequently has an adverse effect on diabetes, and diabetes may first become evident during pregnancy. There is a tendency for the carbohydrate metabolism of the diabetic patient to be upset. Most diabetics need more insulin during gestation; a few, for reasons not understood, need less. The changing condition from day to day makes some diabetics, who have no problem maintaining a balance when they are not pregnant, difficult to manage. Even so, adequate medical supervision can bring most diabetics and their babies safely through pregnancy.

Goitres; pituitary insufficiency; adrenal hypo- and hyperfunction

Thyroid disease. Simple goitres that are not associated with a change in the amount of thyroid hormone in the mother's blood do not affect pregnancy, nor does pregnancy affect the thyroid in such a case. An inactive or too active thyroid gland, if not adequately treated during pregnancy, may be associated with an increased incidence of abortion. In the few cases in which persons with untreated myxedema, a severe form of hypothyroidism (deficiency of thyroid hormone), have conceived and gone to term there has been an increased incidence of congenital anomalies of the fetus. Pregnancy and hyperthyroidism (overabundance of thyroid hormone) seem to have no adverse effects on each other.

Pituitary disorders. Most persons with pituitary hypofunction fail to ovulate because their pituitary glands do not produce the gonadotropic hormones necessary for stimulation of the ovaries. Most of these persons also suffer from a lack of hormones from their other endocrine glands because these, too, lack stimulation by the pituitary. A few persons with hypopituitarism have, nevertheless, become pregnant. Their condition is better when they are pregnant because their placentas produce many of the hormones that their endocrine glands, lacking pituitary stimulation, do not ordinarily secrete.

Adrenal glands. Women suffering from adrenal gland insufficiency are not likely to become pregnant. If they do so, they have more tendency to suffer from circulatory disturbances and carbohydrate, electrolyte, and fluid imbalance because of the important role the adrenal glands play in the metabolism of water, sodium, potassium, chlorides, and glucose. Such patients and their babies do well if they receive hormonal therapy during gestation.

The increased secretion of adrenal hormones that occurs with hyperplasia of the adrenal cortex (enlargement of the outer layer of the adrenal gland, also called Cushing's disease) usually inhibits ovulation. A number of women with this disorder, after treatment with cortisone, have conceived, gone to term, and delivered normal children. Cushing's disease complicated by pregnancy is rare; the few cases reported have been associated with a high incidence of severe high blood pressure.

The maternal death rate is approximately 50 percent and the death rate of the child immediately before or after birth is approximately 40 percent when pheochromocytoma (a type of adrenal tumour associated with, among other things, a high blood pressure) complicates pregnancy.

Kidney and bladder infections

Urinary tract diseases. Infections of the urinary tract are more frequent during pregnancy, and women who have acute infections of the bladder and kidneys while pregnant have a higher incidence of premature labour. This is in accord with the known fact that pregnant women with any type of acute infection tend to deliver prematurely. Many women with pyelonephritis (infection of the kidney) in one pregnancy will enter a second pregnancy with bacteria already in the urinary tract but causing no symptoms. These women have a greatly increased chance of developing acute urinary tract infections during their prepartum course and have some risk of eventually developing serious kidney disease. Glomerulonephritis, a kidney disease that affects the clusters of capillaries in the nephrons, the functioning kidney units, usually is preceded by infection with streptococcus organisms. The incidence of abortion and of premature delivery is increased among women in whom the condition develops during pregnancy. If the glomerulonephritis has become chronic the fetus may not survive and the mother's life may be in danger of kidney failure.

Healed tuberculosis of the kidney is not a contraindication to pregnancy if the disease has been quiescent for three years or longer and kidney function is normal. If tuberculosis of the kidneys is present but without symptoms, pregnancy may cause it to become active. If this happens, and if the infection is limited to one kidney, there is an increased danger that the opposite kidney will become infected in some way. The interference with the flow of urine that is characteristic of pregnancy is an important factor in the development of such infections. The accepted treatment when tuberculosis was present in one kidney during pregnancy formerly was therapeutic abortion followed by removal of the tuberculous kidney. This procedure is now avoided in some instances because of the effectiveness of the antituberculotic drugs that have been developed.

Pregnancy after removal of a kidney. It is sometimes necessary to remove a person's kidney because of an infection, a stone, a tumour, or tuberculosis. The remaining normal kidney has a reserve that is greatly in excess of the demands that will be made by gestation, provided that it does not become infected. Infections, impaired kidney function, congenital defects, and toxemia, however, are more serious for a woman with a solitary kidney than they are for the patient with a normal urinary tract.

Pulmonary disease. Pulmonary disorders have a bad effect on pregnancy if they seriously decrease the amount of oxygen supplied to the fetus, if they make the mother desperately sick, or if they create a blood infection that is transmitted to the placenta.

Respiratory infections

An infection of the upper respiratory tract—the nose and throat—does not ordinarily disturb the course of gestation. It may be serious when it occurs in late pregnancy because of the danger that the mother will transmit disease-causing bacteria to her own genitalia or will carry virulent bacteria from her own nose and throat into the labour room and develop a blood infection after the delivery.

Epidemic influenza is associated with an increased incidence of maternal deaths. Many women who suffer from it abort or deliver prematurely. The infection may pass through the placenta and cause infection in the fetus. Pregnant women who acquire epidemic influenza are more likely to develop pneumonia than are persons who are not pregnant.

Pregnancy may increase or decrease the severity of asthma or may fail to affect it. A severe attack of asthma may be followed by abortion, but otherwise asthma does not affect pregnancy.

Pneumonia occurring during pregnancy is associated with a high rate of maternal and fetal death unless the pulmonary infection is susceptible to antibiotics or chemotherapy. The mother's cardiovascular system, already carrying the load placed on it by pregnancy, cannot sustain the added stress produced by pneumonia. The fetus

often dies from oxygen starvation or from intrauterine infection.

Severe bronchitis and bronchiectasis—abnormal dilation of bronchi with some destruction of bronchial walls—may so interfere with the mother's respiration that the extra strain put on her cardiorespiratory system by pregnancy may put her life in jeopardy. If the disorders are severe enough to cause impaired pulmonary ventilation the fetus may suffer from a lack of oxygen and may be either stillborn or delivered prematurely. Pregnancy does not adversely affect the course of these pulmonary diseases.

Pulmonary tuberculosis is not, as a rule, affected by pregnancy. This is particularly true if the patient's infection has been quiescent for several years before she becomes pregnant. Even women with active tuberculosis, if given adequate care, usually go through pregnancy without any deterioration in their pulmonary condition. This is not universally true, however, because there is a small group with active disease whose disease becomes worse during pregnancy. For that reason individual evaluation of each person is necessary.

Although there have been a few cases of infection transmitted to the fetus prenatally, the great majority of babies born of tuberculous mothers are healthy at birth.

Thoracic surgery. Pregnant women who have had portions of their lungs removed for tuberculosis, tumours, or other reasons do well provided that, before becoming pregnant, they are not short of breath with ordinary exertion. The added load of an additional pulmonary infection may not leave such persons with enough pulmonary reserve for the added burden of pregnancy; they may therefore experience difficulties if they contract pneumonia, severe influenza, or acute bronchitis during pregnancy.

Peptic ulcer; colitis; appendicitis

Gastrointestinal diseases. Women may already suffer from a gastrointestinal disease such as gastric or peptic ulcer, gallbladder disease, or ulcerative colitis when they become pregnant; or they may develop some type of gastrointestinal disturbance during the course of the pregnancy. In either event, pregnancy complicates their problems because the gastrointestinal disturbances that often accompany pregnancy may confuse the diagnosis in an individual case.

Gastrointestinal diseases have little or no effect on pregnancy. Pregnancy, on the other hand, tends to aggravate gastrointestinal disorders; the exception is gastric ulcer, which often improves because the concentration of acid in the stomach is decreased with pregnancy. Women with chronic ulcerative colitis are generally advised to avoid pregnancy until their bowel disease has been quiescent for two years; actually, since the woman's psychological reaction to pregnancy is what affects the bowel, the colitis may be made either better or worse by gestation.

Acute appendicitis, occurring during pregnancy, is often confused with other gastrointestinal complaints, and many patients' lives have been jeopardized either because they ignored the symptoms or because the diagnosis was confused by pregnancy. A diagnosis of acute appendicitis calls for an immediate operation regardless of the duration of the pregnancy or the hazard to the fetus.

Nervous system disorders. Neurological disorders and pregnancy most often are coincidental and have no effect on each other, but there are a few neurological diseases that develop during pregnancy, have a deleterious effect on it, or are adversely affected by it.

Epilepsy; poliomyelitis; polyneuritis; neuralgia

Epilepsy of unknown cause does not affect the course of pregnancy but may occur for the first time during gestation. An epileptic person may find her condition improved, aggravated, or unchanged by pregnancy; the effect of gestation cannot be foretold. There is some evidence that excessive fluid and salt retention induces epileptic seizures.

Pregnant women are more susceptible to poliomyelitis (infantile paralysis), but pregnancy does not affect the severity or the course of the disease, nor does poliomyelitis affect the course of pregnancy. If the muscles of respiration are paralyzed the patient will have difficulty during the latter part of pregnancy, when the uterus presses upward on the diaphragm. There have been a few instances in which babies have acquired infections from the mother before birth.

Polyneuritis, a disorder of the nerves usually resulting from vitamin B deficiency, may complicate pregnancy; this is particularly likely if the patient has suffered from severe and prolonged vomiting. Polyneuritis does not affect the gestation.

Neuralgia (pain that radiates along the nerve) occurs frequently near term. It affects especially the sciatic nerve, which is compressed between the pelvic wall and the head of the fetus.

Brain injury; psychoses

Brain injury, including hemorrhage into the substance of the brain, sometimes occurs as part of the clinical picture of severe toxemia. Some types of brain tumours appear to be adversely affected by pregnancy, but, for the most part, brain tumours are not altered by pregnancy and do not disturb gestation.

Latent psychiatric disorders in unstable persons may be aggravated by pregnancy, but major psychiatric problems seldom appear for the first time during the period before delivery. There are a number of mild emotional disturbances, such as increased anxiety, emotional irritability, and fear of labour or for the normality of the fetus, that are likely to be most intense during the early months of gestation. Such disturbances seem to be most prevalent in women who did not anticipate becoming pregnant or who are unduly worried about the baby. Psychiatric disorders rarely influence pregnancy. Emotional disturbances have been said to be a factor in some spontaneous abortions, but satisfactory proof of the relationship is lacking. An intensely unwanted pregnancy may so seriously disturb an emotionally unstable woman that therapeutic abortion may be recommended after psychiatric evaluation.

DISEASES OF PREGNANCY

Acute toxemia. The toxemias are disorders occurring spontaneously only in human females and only during pregnancy or the puerperium. They are characterized by one or more of the following signs: high blood pressure; an accumulation of fluid in the tissues, often with marked swelling (edema); weight gain and proteinuria (albumin in the urine). If the toxemias are severe the affected women may suffer convulsions and loss of consciousness (coma). In addition there may be eye, brain, kidney, heart, pulmonary, or gastrointestinal disturbances. Severe toxemia may cause death. The factors producing toxemia are unknown, but it has been shown that there is a generalized spasm of the arterioles (minute arteries) with associated severe circulatory effects.

The term toxemia of pregnancy is a misnomer, although it has long been in common use; there is no evidence to link the patient's symptoms and behaviour with the effects of toxic substances. The terms pre-eclampsia and eclampsia, used of subtypes of toxemia, are not particularly descriptive, although they are acceptable. Generally speaking, pre-eclampsia refers to the less severe, nonconvulsive type of toxemia. The term eclampsia is reserved for those instances in which there are convulsions or coma.

Toxemias classified by severity

Toxemias of pregnancy are classified according to the severity of their symptoms. The following classification, now in general use in the United States, was adopted by the American Committee on Maternal Welfare:

1. Acute toxemia of pregnancy (onset after 24 weeks)
   - *a.* Pre-eclampsia
     - (1) Mild
     - (2) Severe
   - *b.* Eclampsia (convulsions or coma, usually both, when associated with high blood pressure, edema or proteinuria)
2. Chronic hypertensive (vascular) disease with pregnancy
   - *a.* Without superimposed acute toxemia (no exacerbation of hypertension, no proteinuria)
     - (1) Hypertension known to antedate pregnancy
     - (2) Hypertension appearing during or before the 24th week and persisting after pregnancy
   - *b.* With superimposed acute toxemia
3. Unclassified toxemias
4. Recurrent toxemias

A woman is said to have mild pre-eclampsia when she is 24 or more weeks pregnant and exhibits one or more of the following signs: a definite trace or more of protein in her urine on two successive days; persistent edema of her face or hands; a systolic blood pressure of 140 millimetres of mercury (mm Hg) or more; a diastolic blood pressure of 90 mm Hg or more; or an elevation of 30 mm Hg or more in the systolic or 15 mm Hg in the diastolic pressures. (The systolic is the highest blood pressure after the heart has contracted; the diastolic, the lowest after the heart has expanded.) Her blood pressure is elevated on two occasions six hours or more apart.

Severe pre-eclampsia is characterized by the presence, after the 24th week of pregnancy, of any of the following: a systolic blood pressure of 160 mm Hg or above, or a diastolic pressure of 110 mm Hg or above, on two or more occasions at least six hours apart, with the woman resting in bed; the excretion of five grams (0.165 ounce) or more of protein in the urine in 24 hours; the excretion of 400 millilitres (about four-fifths of a pint) or less of urine in 24 hours; cerebral or visual disturbances; pulmonary edema or cyanosis (bluish or purplish colour of the skin).

Cerebral and visual disturbances and pulmonary edema and cyanosis are indications of a rapidly deteriorating situation in which the toxemia is seriously affecting, in the first instance, the patient's central nervous system or, in the second instance, her cardiorespiratory apparatus. Cyanosis and pulmonary edema (fluid in the lungs) indicate that the heart is not able to carry out its task and that heart failure is in progress.

The division of pre-eclampsia into mild and severe types by these rigid standards is not truly satisfactory because the physician, when he is attending a patient with toxemia, must evaluate many different things: her age and general appearance, her blood pressure, the amount of swelling, the severity of the proteinuria, her urinary output, and the presence of neurological symptoms. Any condition that has so many facets requires individualization in diagnosis and treatment for each patient.

The woman who develops pre-eclampsia often goes through a characteristic course: she has an inexplicable sudden weight gain of several pounds in a few days, moderate elevation of her blood pressure, and a small amount of albumin in her urine. She is unaware of the latter two symptoms. She is in serious jeopardy by the time swelling of her face and hands, headaches, blurred vision or even blindness, and reduced urinary output occur.

Characteristics of eclampsia

Eclampsia is the presence of convulsions and coma superimposed on the symptoms of pre-eclampsia; *i.e.*, high blood pressure, protein in the urine, and fluid retention.

During pregnancy essential hypertension (chronic hypertensive vascular disease, which many people identify more commonly as "high blood pressure") may become extremely severe, with the symptoms identical to those of pre-eclampsia/eclampsia. Also, pre-eclampsia/eclampsia may be superimposed on chronic hypertensive vascular disease, leading rapidly to an extremely dangerous situation and often ending in eclampsia. Although the reason is not known, chronic hypertensive vascular disease appears to be more common in nonwhite than in white populations.

Some degree of toxemia occurs in from 5 to 10 percent of pregnant women. It is most common in women having their first baby and is common in women with twins, diabetes, or hydatidiform mole (an abnormality of the conceptus that is discussed below); it is one of the three major causes of maternal deaths, along with infection and hemorrhage. It is thought that at least 30,000 babies in the United States are lost annually through stillbirths or prematurity as a result of the toxemias. Eclampsia occurs in less than 1 in 1,000 pregnant women when they receive adequate antenatal care. It is interesting to note that the toxemias occur more frequently in young women or in those who are having their first babies, and in women who are diabetic or have excessive amounts of amniotic fluid.

Eclamptic convulsion

An eclamptic convulsion may seize a toxemic patient at any time. It usually occurs near the end of pregnancy. It may happen when the woman is asleep. Characteristically it begins with facial twitching, a fixed stare, and body rigidity. The eyes then protrude, the face becomes a deep purplish-red, and all the muscles of the body go through rapid, strong, and agitated contractions and relaxations so that the woman throws herself about violently. She usually bites her tongue, foams at the mouth, and falls to the floor. The seizure lasts approximately 60 seconds and is followed by a period of unconsciousness, in which the woman breathes again, her face assumes a normal colour, and her breathing gradually becomes more rapid and deeper. There may be only one attack or many. Death may occur during or after the first seizure or may not occur even though the patient is unconscious for many days. A woman may be in labour when a seizure occurs or may go into labour after one. As a rule, delivery is followed by cessation of the convulsions. Eclampsia after delivery is rare.

When eclampsia occurs, it is usually, but not necessarily, in a patient who has toxemia. The pathological changes characteristic of eclampsia have been discovered in women dying with pre-eclampsia without convulsions, and even in a few rare instances in which the pregnant woman did not have hypertension, edema, or proteinuria. Death may be due to clots or hemorrhages into the brain, liver, kidneys, or heart muscles. The changes noted are characteristically in the blood vessels and are essentially due to generalized spasm of small arteries in many of the vital organs.

The cause of toxemia of pregnancy is unknown. A toxemia-like state has been produced experimentally in primates by partially obstructing the blood supply to the uterus. It is generally agreed that there is no evidence that toxic substances are responsible for the toxemias. Whatever the cause, it would seem that it produces an increase in tonicity or contractility of the walls of the finest capillaries throughout the body; this explains the proteinuria and the rise in blood pressure that are part of the clinical picture of toxemia.

Treatment of eclampsia

The earliest possible recognition of mild pre-eclampsia and its energetic treatment from the first appearance of the symptoms that herald its onset are effective in warding off severe pre-eclampsia and eclampsia. The effect of such care was dramatically demonstrated in Sydney, Australia, in the 1950s; when prevention was based on the earliest possible discovery of excessive weight gain and increased blood pressure, there were only ten cases of eclampsia among 37,850 patients delivered between 1948 and 1957. In order to achieve such a record it was necessary at times for the medical authorities in Sydney to send the constabulary to inform women who did not keep their appointments for prenatal care.

The treatment of mild pre-eclampsia is directed toward avoiding severe pre-eclampsia and eclampsia. Preferably the affected woman is hospitalized; retained salt and water are eliminated; and an adequate protein diet, rest in bed, and close observation are essential.

The treatment of severe pre-eclampsia attempts to overcome central nervous system hyperexcitability, to reduce vasospasm, and to lower the blood pressure through the use of suitable drugs. As a result, cerebral and renal circulation are increased, urinary output is promoted, and retained sodium and water are excreted. A patient with severe pre-eclampsia belongs in the hospital. Usually, but not invariably, it is safest to perform a therapeutic abortion if a woman has severe pre-eclampsia.

The treatment of eclampsia is directed toward prevention of further convulsions; decreasing vasoconstriction; reducing blood pressure; increasing urinary output; and maintaining fluid balance.

Termination of pregnancy by inducing labour or by caesarean section is the only definitive treatment for toxemia. If waiting is safe for the mother, it is desirable not to deliver the fetus until it can survive. Delivery is accomplished, after eclampsia is controlled, by whatever means seem best, induction of labour being preferable to caesarean section.

Diseases of the placenta. *Placenta praevia.* Implantation takes place in the lower half of the uterus in ap-

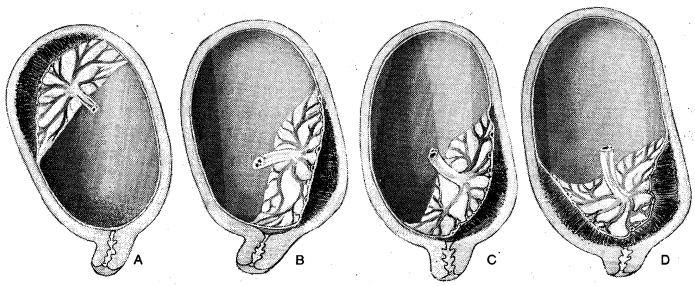


Figure 5: (A) Normal implantation of the placenta in the top or fundus of the uterus. (B) Low implantation. (C) Partial placenta praevia. (D) Complete placenta praevia.
From J. Huffmann, *Gynecology and Obstetrics* (1962); W.B. Saunders Co.

proximately 1 in 500 pregnant patients. The condition is known as placenta praevia when the placenta lies over all or a portion of the internal opening of the cervix. A total placenta praevia is present when the cervical opening is completely covered. When there is a low implantation of the placenta, the latter lies close to but not over any part of the cervical opening.

Recurrent painless bleeding from the vagina without other symptoms after the sixth month of pregnancy is the typical manifestation of placenta praevia. It is caused by disruption of the placenta as the cervix and lower uterine segment are pulled upward. Each bleeding episode tends to become heavier. Without proper treatment the baby is likely to die and the mother may do so. Unremitting watchfulness of the woman with placenta praevia until the fetus has a chance of survival, with preparation for immediate delivery if hemorrhage becomes brisk, a practice accepted in many clinics, has resulted in a decreased infant mortality without an increase in maternal deaths.

Characteristics, course, and outcome of abruptio placentae

Abruptio placentae. Abruptio placentae is separation, during the latter half of pregnancy, of the normally implanted placenta from its attachment to the uterus before birth of the baby. It also is correctly referred to as "premature separation of the normally implanted placenta" and is called "accidental hemorrhage" in Great Britain. It occurs in approximately 1 in 100 pregnancies. The cause is unknown. It is more common in women who have borne several children.

When a small portion of the placenta separates from the uterus, a condition called partial abruptio placentae, blood either collects in a pool between the uterus and the placenta (concealed hemorrhage) or seeps out of the uterus into the vagina (external hemorrhage). When the entire placenta separates from the uterus, there is massive hemorrhage into the uterine cavity and sometimes into the wall of the uterus. Massive hemorrhage is associated with uterine tenderness, abdominal pain, shock, and loss of fetal movement and fetal heart tones. The baby usually dies. If hemorrhage is severe the mother's life is in danger. Defective blood clotting occurs in at least 35 percent of patients with abruptio placentae. Kidney failure develops in approximately 1 percent of the cases; it is seen most often in those instances in which treatment has been delayed. Blood replacement, the treatment of shock, the administration of fibrinogen if the patient's clotting mechanism is defective, the administration of oxytocics, and early delivery are the basic essentials of the treatment of abruptio placentae. Delivery is usually by caesarean section.

Placental infarction. Infarction is degeneration and death of a tissue and its replacement with scar tissue. Small areas of infarction, caused by interference with the maternal circulation, occur normally in the placenta as pregnancy progresses. The fetus usually is not affected by infarction of the placenta unless the process is extensive.

Placenta accreta. Placenta accreta is an abnormal adherence of the placenta to the uterine wall. The chorionic villi attach themselves directly to the uterine muscle in areas where the decidua is poorly developed or absent. All or part of the placenta may be affected. As a result of this abnormality of implantation the placenta does not separate normally at the time of delivery. Attempts to remove it manually by the physician are frequently followed by severe hemorrhage. Removal of the uterus may be required to save the mother's life.

Placental cysts and tumours

Placental cysts and benign tumours. Placental cysts and benign tumours are relatively rare. Chorionic cysts of small size are disk-shaped, grayish-white structures filled with a yellowish fluid and located on the fetal side of the placenta. Decidual cysts are smoothly lined small cavities in the centre of the placenta; they are the result of decidual degeneration and are not true tumours. Angiomas, hemangiomas, fibromas, myxofibromas, and the like are benign growths arising from the placental blood vessels and connective tissue. Solid or semisolid tumours, usually creating small nodular elevations on the fetal side of the placenta, are rarely of clinical significance.

Inflammation of the placenta. Inflammation of the placenta is usually secondary to infection of the membranes. Most often such infections follow the introduction of pus-forming bacteria into the uterus by instrumentation through the vagina; they are the aftermath of prolonged labour or of prolonged rupture of the membranes. If labour is prolonged, bacteria penetrate the fetal side of the placenta, enter the fetal circulation, and often cause death of the infant after delivery.

The placenta may become infected from organisms in the maternal blood. Maternal syphilis, toxoplasmosis, tuberculosis, and malaria may affect the placenta. The viruses of chickenpox and smallpox may cause placental lesions. A number of pathogenic bacteria and viruses cross the placenta and sometimes kill the fetus without causing any specific changes that have been noted in the placenta.

Placental anomalies. Abnormalities in the structure of the placenta are relatively common. It may be partially divided into two or more lobes; there may be extra lobes; or the placenta may be divided into two or more separate structures. Abnormal placentas result from shallow and from deep implantation. The former type, called placenta circumvallata, is associated with several maternal and fetal complications; the latter type, called placenta membranacea, may cause problems at delivery; *e.g.*, bleeding, failure of the membrane to separate.

Anomalies of the umbilical cord. "False knots," which are simply enlarged blood vessels in the cord, are not significant. Actual knots in the cord may become tightened and kill the fetus by cutting off the blood to it. Twisting of the cord also may kill the fetus in the same manner. Spontaneous rupture of the cord interferes with the fetal blood supply and causes fetal death. Extreme shortness of the umbilical cord may interfere with delivery, cause premature separation of the placenta, or tear and cause fetal death from hemorrhage. Another abnormality, called velamentous insertion of the cord, in which multiple blood vessels spread out over the membranes and cervix rather than forming one single cord, is dangerous for the baby because the vessels may tear or be compressed during labour and delivery.

Excessive amniotic fluid

Abnormalities of the amniotic fluid. *Polyhydramnios.* Polyhydramnios is the presence of an excessive amount of amniotic fluid. Normally the uterus contains approximately 1,000 millilitres (slightly more than one quart) of amniotic fluid; anything over 2,000 millilitres is abnormal. Accumulations of more than 3,000 millilitres occur in approximately one pregnancy in a thousand. Lesser degress of polyhydramnios probably occur in about 1 in 150 deliveries. The appearance of large amounts of fluid within the space of a few days is rare; such a condition is met with in less than 1 in 4,000.

Polyhydramnios occurs most often in association with fetal abnormalities, particularly those of the nervous, digestive, and renal systems; when the fetus has erythroblastosis, a disease resulting from incompatibility between the infant's and the mother's blood; when there is more than one fetus; or when the mother has diabetes or one of the toxemias. Almost all pregnancies in which the fetus suffers from obstruction of the esophagus and half of those in which there are severe brain anomalies are accompanied by excessive amniotic fluid.

Acute polyhydramnios causes rapid overdistention and enlargement of the uterus. The woman experiences ab-

dominal pain, nausea and vomiting, and difficulty in breathing. Her heart and blood vessels are put under severe stress; she may show signs of heart failure. Swelling of the feet and legs develops. These manifestations are all caused by the pressure of the rapidly enlarging uterus upon the other viscera.

Chronic polyhydramnios usually causes enough pressure from the abnormally enlarged uterus to make the affected person uncomfortable.

The cause of polyhydramnios is unknown. The most tenable theory is that there is a reduction in the amount of fluid that passes from the fetus to the mother and an increase in the amount that passes from the fetus to the amniotic sac. This would explain the relationship between fetal anomalies and polyhydramnios.

The baby's prospects are poor when there is polyhydramnios. Many pregnancies complicated by an abnormal amount of amniotic fluid terminate prematurely. The fetus has a greatly increased chance of suffering from congenital anomalies. Roughly half of the babies in this group have been lost in the series of cases that have been reported. The greater the amount of fluid, the higher the fetal mortality. Women with polyhydramnios are also faced with a somewhat higher risk. Premature separation of the placenta and postpartum hemorrhage are the two most significant maternal complications associated with it.

Minor degrees of polyhydramnios require no treatment. Removal of the excess fluid is the only effective management if symptoms from uterine distention become too distressing. This may be done either by perforating the membranes through the cervix or, preferably, by inserting a needle through the abdominal wall and the wall of the uterus; care is taken to avoid injury to the woman's bowel or the placenta. Either procedure is likely to start labour.

Too little amniotic fluid

Oligohydramnios. True oligohydramnios, a deficiency in amniotic fluid, is a rare condition of unknown cause. If it occurs early in pregnancy there are usually firm adhesions between the membranes and the embryo, with distortion of the fetus. A decrease in the amount of fluid later in pregnancy allows the membranes and uterine wall to press on the baby. The baby's position is distorted, and as a result it may be born with a clubfoot or wryneck. Its skin is dry and thickened. Defective development of the kidneys is common with oligohydramnios. As a rule, the condition causes the mother no distress, but the infant has a greatly increased chance of being born with major anomalies.

Trophoblastic disease. *Hydatidiform mole.* A hydatidiform mole is an abnormality of the conceptus in which changes that began early in embryonic life convert the placental villi into a mass of thin-walled, grapelike, translucent vesicles, or blisters, filled with a gelatinous or watery fluid. In a typical case the uterus is distended by a spongy mass of these vesicles. The primary cause of molar changes is unknown; however, it has been correctly described as a "temporary missed abortion of a blighted ovum." The embryo is either absent or dead. The immediate condition that causes hydatidiform swelling is disappearance of the blood vessels in the villi, with continued growth and often overgrowth of the trophoblast. Distention of the villi by fluid is due to continued activity of the trophoblast in the absence of a functioning villous circulation.

In the ova there are many degrees of hydatidiform change; many of the changes, usually in younger specimens, are not marked enough to warrant being called hydatidiform moles. True moles—characterized by hyperplasia, or overgrowth, of the trophoblast, edema of the villous connective tissue framework, and defective growth of the villous blood vessels—occur perhaps once in 2,000 pregnancies. They are not tumours and are not the aftermath of a former pregnancy. They are themselves an abnormality of a current pregnancy. Occasionally in a twin pregnancy one fetus is normal and the other a mole. Eighty percent of the moles are expelled about the 20th week of pregnancy and bring the patient no more trouble. Approximately 16 percent of hydatidiform moles invade the uterine muscle, causing bleeding. This type of mole, referred to as an invasive mole or chorioadenoma destruens, may in rare instances perforate the uterus and cause death from hemorrhage. Rarely molar villi are carried to the lung or brain. When they are, the patient may suffer from hemorrhage into the lung or die from hemorrhage within the brain.

Signs of hydatidiform mole; complications

The woman who develops a hydatidiform mole has the symptoms of pregnancy; her uterus usually enlarges more rapidly than it should, she is more likely to suffer from toxemia, and she begins to bleed vaginally, usually by the 20th week of gestation. The molar pregnancy is expelled vaginally, or, if hemorrhage is severe, the obstetrician may remove it by surgery.

In approximately 2.5 percent of patients, hydatidiform moles change into choriocarcinoma, a highly malignant tumour of the trophoblast. For that reason, patients who have hydatidiform moles are observed carefully. Continued bleeding or a rising quantity of chorionic gonadotropin in her urine or blood after passage of a mole suggests that a patient has either an invasive mole or a choriocarcinoma. Chemotherapy has been effective treatment for many cases of this type. Removal of the uterus may be necessary. The complexities of diagnosis and the differences in situations require that therapy be keyed to the individual.

Choriocarcinoma. Choriocarcinoma is a rare, extremely malignant type of tumour arising from the trophoblast. The reasons that normal chorionic cells undergo cancerous change, with exaggeration of their natural and potent tendency to invade the uterine muscle and break down blood vessels, are unknown. Choriocarcinoma occurs approximately once in 160,000 normal pregnancies. In approximately 50 percent of the cases the tumour develops from a hydatidiform mole, in another 25 percent after an abortion, and in 25 percent after a normal pregnancy. Occasionally it appears after a tubal pregnancy. It has been known to coexist with pregnancy. It is, for some unknown reason, more common in the Orient. Choriocarcinoma developing as a teratoid tumour of the ovary (a tumour made up of a number of different tissues) is a rare entity not related to pregnancy and is not to be confused with the tumour being discussed here.

Signs of choriocarcinoma

As a rule, in the development of a choriocarcinoma there has been a normal pregnancy, an abortion, or the delivery of a mole, and the uterus has not returned to its normal size. The woman begins to bleed from the vagina. Blood loss may be modest or excessive in amount. Tissues obtained by a curettage (scraping) may be, but are not always, indicative of choriocarcinoma.

The tumour begins in the uterus, where it forms a spongy, bleeding mass of easily torn tissue or a shaggy ulcer. When examined microscopically, it is found to consist of both cytotrophoblast and syncytiotrophoblast. The cells spread rapidly by way of the bloodstream, producing secondary tumours in the lung, the brain, the liver, or elsewhere.

Choriocarcinoma formerly was almost invariably fatal. Today, an impressive (two out of three in some case series) number of patients have survived for many months after the administration of chemotherapeutic agents. Most workers in this field at this time are using methotrexate. The rapidly growing embryonic cells of the trophoblast need nucleic acids for growth and division; for the synthesis of nucleic acids, folinic acid is essential, and methotrexate, by preventing the conversion of folic acid to folinic acid, cuts off the supply of the latter. A number of other cytotoxic drugs (drugs destructive to cells) are also being used in the treatment of choriocarcinoma, and other chemotherapeutic agents are being tested for effect on this type of tumour. Removal of the uterus is frequently, but not always, a part of the treatment of choriocarcinoma.

BIBLIOGRAPHY. Additional information on this subject may be found in the following texts, where the contents in most instances are indicated by their titles: L.B. AREY, *Developmental Anatomy*, 7th ed. (1965); W.J. HAMILTON, J.D. BOYD, and H.W. MOSSMAN, *Human Embryology*, 3rd ed. (1962); C.A. VILLEE (ed.), *The Placenta and Fetal Membranes* (1960); A.T. HERTIG, *Human Trophoblast* (1968);

J.W. WILLIAMS, *Obstetrics*, 13th ed. by N.J. EASTMAN and L.M. HELLMAN (1966); W.J. DIECKEMANN, *The Toxemias of Pregnancy* (1952); S.R.M. REYNOLDS, *Physiology of the Uterus* (1949); E.L. HOLLAND and A.W. BOURNE, *British Obstetric and Gynecological Practice*, 2 vol. (1955); C.S. BURWELL and J. METCALFE, *Heart Disease and Pregnancy* (1958); C.L. MENDELSON, *Cardiac Disease in Pregnancy* (1960); F.J. and J.C. MCCLURE BROWNE, *Antenatal Care and Postnatal Care*, 9th ed. (1960); F.E. HYTTEN and I. LEITCH, *The Physiology of Human Pregnancy* (1964); J.D. BOYD and W.J. HAMILTON, *The Human Placenta* (1970); and J.W. HUFFMAN, *Gynecology and Obstetrics* (1962).

(J.W.Hu.)

Prehistoric Religion

Prehistoric religion comprises the religious beliefs and practices of early man. In the absence of any written records, statements about the nature of prehistoric religion are necessarily dependent on a complex body of theory, concepts, and data, and on inferences and reconstructions from archaeological remains.

BASIC PROBLEMS AND METHODS OF STUDY

Problems: nature and scope of prehistory. The term prehistory refers to the course of human events and to the study of these events, in so far as they were decisive in shaping contemporary and later conditions, before the invention of writing. As a subject of inquiry, prehistory comprises both the course of events in preliterate times and man as a cultural being in a state of illiteracy. Similarly, the study of prehistoric religion comprises two subjects: the history of early man's religion and the study of religion as a part of human culture or human nature or both.

Reconstruction of the past from material remains. As the British prehistorian G.E. Daniel has noted, "prehistory is written from many sources—the material remains of the past, deductions from language, physical anthropology, place names, and comparative ethnology. All these are sources that in historical times are auxiliary to the written sources. But in prehistory they are the main sources," and, he emphasizes, material remains are "by far the most important of them." The "reading" of such material remains requires, in the main, a functional interpretation; *i.e.*, discernment of their cultural and spiritual function in the life of prehistoric man. Above all, this functional interpetation makes use of comparisons with objects, circumstances, and processes that are directly known or that have been transmitted in written, pictorial, or oral form. Since the major part of prehistory leads back into the realm of so-called primitive cultures, comparative studies must take into account the conditions and behaviour patterns that research on contemporary "primitive" people has revealed. Such studies also serve to demonstrate the fragmentary character of archaeological finds. In addition, they show that for many archaeological objects, various functional interpretations are possible.

The significance of material remains

Problem of discerning cultural and religious phenomena. Religion is a mental or spiritual phenomenon in which the sacred or supernatural "word" plays an important part. Obviously, this essential expression of religion cannot be investigated archaeologically—the remains are wordless. Religious meanings can only be inferred indirectly and in exceptional circumstances from the material finds. The investigation of prehistoric religion faces the problem of determining which objects and finds are to be taken as signs of religious intention, experience, and activity. Not all unusual and puzzling finds necessarily reflect something religious; moreover, there are some remains that may have had a profane as well as a religious or magical function.

Conjectural nature of scholarly statements. In all probability, only a small number of the material remains of prehistoric religious practices have been preserved and still fewer of them discovered and adequately documented. This intensifies the necessarily fragmentary nature of all statements about prehistoric religion. Uncertainties about methodology further aggravate disputes among scholars as to the reliability or probability of their interpretations. As with historians in general, there is a positivistic position, the proponents of which desire little more than to classify, describe, and define their specific finds and findings. That stance is opposed by another position that, in addition, considers it important to relate these findings to common human experiences and to seek out analogies with situations in more recent times; that is, to attempt a reconstruction of prehistoric spiritual life.

Methods. *Material remains.* Although religion is largely a mental phenomenon, it involves numerous material accessories: artifacts and places of cultic and ceremonial significance, pictures and symbols, sacrifices, and votive offerings. In many cases religion makes use of art and, to some degree, certain inferences about religious conceptions can be drawn from burial sites. The interpretation of such sources from analogies with present-day religious practices means, in effect, that a more or less complete congruence or similarity is being inferred from a partial congruence that is observed. Not infrequently, particular findings can be explained in various ways. It is often unclear, for example, to which religious category a find belongs, since a sacrifice and a burial, cannibalism and human sacrifice, or animal sacrifice and an animal cult are frequently not distinguishable archaeologically. Thus, it is not enough to select some particular present-day "primitive" religious phenomenon and apply it automatically to archaeological material; rather, it is first necessary to carry out extensive comparative studies in order to ascertain connections of sufficient scope and to establish a basic congruence of meanings. Usually only general trends, rather than concrete particulars, can be comprehended in this fashion. Nevertheless, the find may be interpreted in various ways and, hence, it becomes necessary to select the most workable, or the most probable, interpretation.

Archaeological concepts. The first step in interpreting prehistoric finds consists in relating them only to those recent phenomena that are truly comparable, that is, that have basically similar or congruent structures. A historical religious phenomenon that is connected with the cultivation of domesticated plants, for example, cannot be employed, without further inquiry, to explain an object from a Paleolithic (Old Stone Age) hunting and gathering culture. Such an explanation, if it is to be valid, presupposes that the chronological sequence of the archaeological phenomena has been determined and that certain stages of the cultural development can be worked out. Today this task may be considered as solved in its main outlines, although the reconstruction as a whole still needs some refinement.

Archaeological finds and cultural stages

Less clear, and still disputed, is the reconstruction of corresponding stages on the basis of modern "primitive" phenomena, as well as the connection of the latter to the prehistoric stages; *i.e.*, the answer to the question of whether and how phenomena that still exist may be correlated with those of the past. Even where this can be done only tentatively, constructing comparable cultural stages or structures in the past is important for the interpretation of the phenomena. This is even more the case if the phenomena can be attributed to certain smaller units within the stages, such as "culture circles." The best way to establish such connections is a direct demonstration of continuity; *e.g.*, connecting the content of North African rock-pictures with recent religious conceptions within the "culture circle" of "Eurafrican hunter-collectors" of past and present times. Even where continuity or diffusion cannot be demonstrated, conformity or analogy in cultural structure or style furnishes valuable clues.

Methods of interpreting archaeological materials. Criticism of the use of historical-cultural analogies—particularly that of prehistoric phenomena to modern "primitive" religions—points especially to the large time span between prehistoric and contemporary "primitives," the numerous possibilities for changes in meaning, and the present proliferation of phenomena. At the same time, numerous examples show that a long elapsed time does not necessarily mean that fundamental changes have occurred, but rather that there are very strong tendencies toward stability. During all of prehistory, for example, the first truly epochal transformation, in terms of the

total culture (*i.e.*, the new form of developed hunter-collector cultures of the Upper Paleolithic Age), took place around 30,000 BC; thus, in comparison to the hundreds of thousands of years of early humanity, the time elapsed since then is rather insignificant. Consequently it is less unimaginable than one might think that primary elements of very early conditions have endured to the present time. In addition, in contrast to the present multitude and diversity of phenomena, it can be shown that the number of possibilities that can be realized under simple conditions is really rather limited. Further, a mental phenomenon, such as religion, does not develop entirely freely but is also governed by functional interdependencies and limitations, *e.g.*, even by economic or ecological conditions (see below). Temporal duration, the interaction of the tendencies of stability and change, the multiplicity of possibilities and limitations through general conditions, independent development, and functional interdependence are all to be weighed carefully one against the other. The different conceptions of the value of each of these criteria are to a large extent the basis for disputes among scholars over the meaning and persuasiveness of the inferences they draw from them.

Assumption of the psychic unity of mankind

A fundamental prerequisite for the use of historical-cultural analogies is the assumption of the psychic unity of human nature (*i.e.*, that human beings in all times and places are essentially the same mentally); hence, the basic question arises as to whether or not early prehistoric humans can be included in that unity. The particular answer one gives to this question corresponds to a great extent to his particular approach, whether from the point of view of biology or the humanities, of biological evolutionism or history. The biological sciences tend to view early man—even psychically—as half animal, while human studies emphasize more strongly the fundamental similarities that connect even the earliest humans and cause them to stand apart from the animals. Whatever the case, it can no longer be doubted that earliest man stands closer to present-day man from the evolutionary and biological aspect than to any of the animals known today (including the "highest" apes), and that the quality of early man's mental accomplishments separates him from other animals and binds him fundamentally to present-day humans. For that reason it is generally sound—at least as a guiding principle in empirical research—to view prehistoric man as essentially human, to understand him in terms of human nature, and to proceed from the assumption of the psychic unity of mankind without thereby excluding further evolution and development.

INFERRED PRACTICES AND BELIEFS

Burial customs and cults of the dead. The oldest known burials can be attributed to the Middle Paleolithic Period. The corpses, accompanied by stone tools and parts of animals, were laid in holes in the ground and sometimes the corpses were especially protected. In some cases, the findings give the impression that the dead were to be "held onto." Whether or not that meant that the dead were to be cared for lovingly or that their return was to be feared, it implies, in any case, a belief in life after death in some form. But it is not necessary to infer a belief in separate souls; rather, it could also indicate the concept of a "living corpse."

Implication of life after death

From the Upper Paleolithic Period on, the burials manifest richer grave goods; however, it is not possible to conclude from this that religious concepts had changed. The same holds for the adoption of other burial practices, as, for example, secondary burials, in which the bodies were first allowed to decompose fully and then the bones were buried, or in the burning of bodies (evident from the Neolithic Period). From these facts it is not possible to infer the existence of a definite belief in souls; it is also not possible to determine the advent of such concepts from archaeological evidence. Even the increase in the discoveries of grave goods, occasionally also including other human remains, is evidence not for a change of religious concepts but for increased needs of the dead in the beyond—*i.e.*, needs after death that are dependent on economic and social status in life. Analogies to recent (primitive) phenomena demonstrate that it is not possible to connect particular burial customs with particular notions of the beyond, or to any other religious conceptions. Other than the burial of the whole body, the disposition of the individual parts of the body, and especially the skull, is important. Ritual deposition of skulls is confirmed for the Middle Paleolithic Period. From even earlier periods, however, individual or multiple human skulls and long bones have been found within a single site (for example, with Peking man). It is not necessary to interpret these findings as remains of headhunting or developed skull cults; for even today some simple hunting and gathering societies have the custom of preserving such parts of corpses for long periods of time and even of carrying them around on their bodies. The same practice is observed also in the Upper Paleolithic and even later periods; but it is not possible to infer an elaborated ancestor cult directly from such prolonged connections of the living with the dead.

Burial of skulls and skull cults

The situation is different with findings from permanent settlements of agrarian people, in contrast to constantly shifting hunter-collectors. Evidences for ancestor cult practices dating to the 7th century BC were first discovered at Jericho in Palestine, where several skulls were deposited in a separate room, some of them covered with a plastic modelling of faces similar to that found on the ancestral skulls of present-day agrarian peoples of South Asia and Oceania. An elaborated skull cult is usually connected with the veneration of ancestors. An important theme of ancestral cults is the belief in a connection between the dead and the fertility of the land of their descendants.

Megalithic graves

An especially noteworthy kind of burial is that of the megalithic (huge stone) graves that appear in various areas from the Neolithic Period on. It is probable that in this practice there was also a vital believed link between the living and the dead, and that occasionally sacred areas and gathering places were connected with such graves. The practices of the megalith builders were probably rooted, to a considerable extent, in ideas about the dead and in ancestor cults to which their stones gave a particular durability and a monumental form. It is more difficult to explain the individual erect stones (menhirs), which, of course, could be the symbol or seat of ancestors, especially where they show indications of being sculpted in human form. It surely would be a mistake, however, to look for a uniform intepretation of all megalithic monuments or even to speak of a distinct megalithic religion. The megalithic monuments are rather to be understood as a complex of grandiose manifestations of ideas that could well have been diverse, but among which the cult of the dead, nevertheless, played an important role.

Cannibalism. In finds belonging to the Paleolithic Period, pieces of human bodies as well as the bones of other animals are found scattered throughout the archaeological layers and are sometimes broken or charred. This is often taken as evidence for cannibalism, but other interpretations are just as likely (*e.g.*, the action of carrion-eating animals [such as hyenas] turning up the bones to the surface and thus causing their burning by later fires at the same place). To be sure, the finds allow the interpretation of cannibalism; however, they do not necessarily or intrinsically require it but rather permit that explanation if one proceeds from the prior conviction that cannibalism already existed at that time. This obsolete conception, still held by some scholars today—*i.e.*, that cannibalism is an especially "primitive" phenomenon and therefore very ancient—must be abandoned. Ethnological studies show clearly that cannibalism appears almost exclusively in the practices of agrarian peoples, that is, in a later cultural stage, and evidently is essentially bound up with religious or magical conceptions in which cultivated plants play a large role. Even if a Paleolithic cannibalism existed on a large scale, it could not be explained by means of concepts that originated in a cultural stage so differently structured.

The situation in later periods, especially in the Neolithic, is different. Here, rather than isolated parts of human

skeletons scattered about a settlement, human remains occasionally are found in association with remains of foodstuffs in waste pits or in holes and tunnels that served as sacrificial sites. Especially where human skulls have been broken open and the hollow bones split, the interpretation of cannibalism is unavoidable. Since this inferred practice occurred in the realm of agrarian cultures, it is more feasible to make comparisons with present-day cannibalism, where the meaning is generally the acquisition of the powers and other qualities of the victim.

Sacrifices. Sacrifices (*i.e.*, the presentation of offerings to higher beings or to the dead) appear as early as the Middle Paleolithic Period. Pits with some animal bones have been found in the vicinity of burial sites; thus, it is a likely possibility that they represent offerings to the dead. There is a dispute over the interpretation of the arrangement of the skulls and long bones of bears, since they are deposited in such a manner that it is hardly possible to discern a profane explanation. It is assumed that they had a cultic or magical significance. Most likely, certain parts of the prey, such as the head and the meaty shanks, or at least the bones with brain and marrow, were sacrificed. Even if it cannot be definitely stated who the recipient of these sacrifices was, analogies with present-day "primitive" phenomena make it likely that a part of the prey was offered to a higher being who was believed to dispense nourishment. It could also, however, have been a matter of preserving parts of animals in order to resurrect the entire animal and preserve the species. Furthermore, finds of bones and drawings show that the preservation of skulls with still attached vertebrae, ribs, and front legs of oxen and reindeer played a certain religious or magical role. The sinking of whole reindeer into lakes is hard to explain other than as a sacrifice. This might be traced to the idea that what occupies the centre of attention is not the individual hunted animal but the whole herd; no longer only a part of an animal but a whole animal as part of a herd is sacrificed. The custom also existed in recent times among hunters and herders of central and north Asia. As such finds become more numerous, it seems evident that certain specific animals and parts of their bodies are selected for sacrifice.

Cultic and magical significance of sacrifices

It is difficult to differentiate between animal sacrifices and the immediate cultic veneration of an animal at the burial sites of animals. In the Neolithic Period, the sites become especially profuse and are usually found in connection with human burials; nevertheless, there are such burial sites of animals that are not related in this manner and that occur with pronounced frequency, characteristically in particular groups of cultures. In these cases, domestic animals almost exclusively are involved, and among them the dog and the ox predominate.

The question of human sacrifice is of special significance here. Human sacrifices often were related to cannibalism and to the sacrifice of animals. With conspicuous frequency victims discerned in ceremonial remains are females and children, sometimes along with young pigs. This practice is similar to fertility and agricultural rites that are known to have been practiced in the early Mediterranean civilizations. It is also similar to beliefs and practices observed among present-day "primitive" agrarian peoples (in which pigs are often substituted for humans), such as in ceremonies of secret societies, initiation rites, sacrifices, celebrations of feasts of the dead, and notions about fertility, especially in connection with the growing and ripening of cultivated plants.

In comparison, the inclusion of servants or women in the burial sites of highly placed persons can hardly be called sacrifice in a strict sense—that is, an offering to a higher power or deity. Such inclusions most likely reflect the social status of the deceased leader and his need for servants in the afterlife, rather than an offering. It is a sacrifice in the wider sense of respect and awe for the person and status—and all that this conveyed—of the deceased leader. This practice becomes more important only where correspondingly differentiated social conditions are found (such as in the royal graves at Ur in Mesopotamia and in those of the Shang dynasty in China). Sometimes it took on almost unbelievable forms, especially in terms of the numbers of persons and animals interred with the deceased leader.

The ritual preservation of objects also must be included in the realm of sacrifice (in a wider sense). This can be demonstrated for the first time in the Neolithic Period (for instance, the ritual depositing of axes); in later periods, it plays a large role. In finds from the Bronze Age on, weapons and jewelry frequently are found in wells and springs. In Iron Age finds, such objects are found in almost unbelievable quantities in a number of swamps and other bodies of water. It seems probable that they represent the sacrifice of war booty.

Ritual preservation of objects

Hunting rites and animal cults. In the oldest known examples of graphic art, the representations of animals play a large part; humans appear rarely and then frequently with animal attributes or as mixed human–animal figures. In the context of the whole situation, the view that these representations were merely ornamentations or served a purely artistic need may be dismissed; they are found without boundaries and background on rock walls and are not part of an interrelated scene. It is evident that animals played a predominant role in the mental world of the Upper Paleolithic Period insofar as this role is reflected in the art of the period. What is represented is, first of all, that which is essential to the animal, partly in its relation to the hunt, but also in relation to anthropomorphic figures showing the intermixing of human and animal forms. This indicates a special and intimate relationship between humans and animals that transcends and overcomes the boundaries between different realms of being that modern concepts and understanding require.

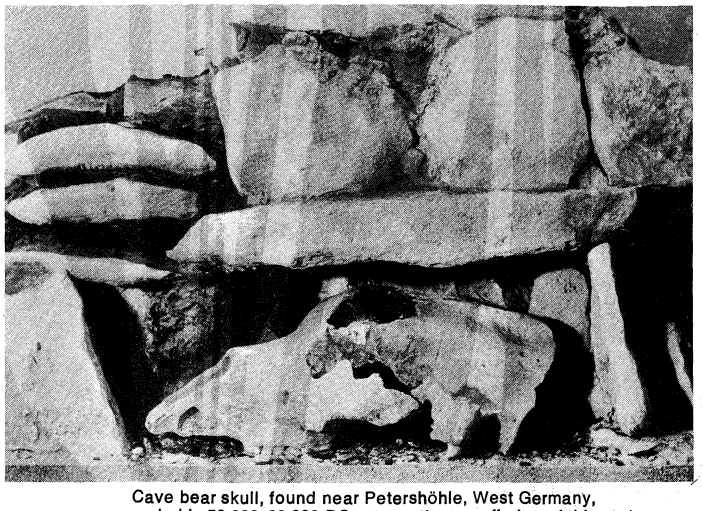
This phenomenon is similar to what is still known today as animalism (or nagualism or theriocentrism). It is characterized by close magical and religious ties of humans with animals, especially with wild animals. It is also characterized in terms of otherworldly and superworldly realms and practices, such as placating and begging for forgiveness of the game killed, performing oracles with animal bones, and performing mimic animal dances and fertility rites for animals. Animals were thought to be manlike, to have souls, or to be equipped with magical powers. Animalism thus expresses itself in various conceptions of how animals are regarded as guardian spirits and "alter egos," of the facile and frequent interchangeability between human and animal forms, and also of a theriomorphically (animal-formed) envisioned higher being—one who changes between human and animal forms and unifies them. Higher, often theriomorphic, beings are gods who rule over the animals, the hunters, and the hunting territory, or spirits in the bushland and with the animals. It is obviously not possible to identify special occurrences or forms of such higher beings during the Paleolithic Period, but their general features may be safely assumed.

Totemism

Animalism is, to a large extent, a basis for totemism, which involves various permanent relationships of individuals or groups to certain animals or other natural objects; hence animalism is occasionally called "protototemism." Individual and cultic totemism, as opposed to group totemism of an almost solely social function, are particularly close to animalism, whereas religious and cultic meanings in group or clan totemism are usually poorly developed. It is not possible to determine to what extent animalism had already assumed the character of true totemism in the Paleolithic Period; the early existence of clan totemism is improbable because it occurs primarily among peoples who are to some extent agrarian, and possibly a certain kind of sedentary life was prerequisite to its development.

Also, special sacrificial traditions were closely connected to game, particularly the custom of preserving the animal skeleton or a part of a skeleton in order to placate the ruler of the animals (see above) and to provide for continuation of the species.

A certain kind of bear ceremonialism is rooted in this conception and is to be recognized in several finds and pictures from the Upper Paleolithic Period on. A skin with attached head was evidently draped over the body of



Cave bear skull, found near Petershöhle, West Germany, probably 50,000–30,000 BC, suggesting an offering of this part of the prey.

By courtesy of the Naturhistorische Gesellschaft, Nurnberg, West Germany; photograph, Georg Pandura

a bear made out of clay; the skull and long bones of the bear were buried separately (a practice begun in the Middle Paleolithic Period); the bear was shot with arrows and killed by a shot or a thrust into the lungs; the animal or a bearlike figure was surrounded by dancers. Similar phenomena are documented for more recent periods, above all for the hunting cultures of Neolithic Siberia. These observations can be effortlessly fitted into the practice of bear ceremonialism that is still widely distributed in northern Eurasia and North America.

The question of whether animals were the immediate objects of a cult is extremely difficult to judge in each particular case. Nevertheless, with the beginning of the Neolithic Period, animal phenomena appear that probably go beyond functioning merely as a sacrifice and symbol. This applies especially to representations of oxen and bulls and to the symbolism of bull heads and bull horns.

Fertility statuettes

Female fertility deities. Small female figures, the so-called Venus statuettes, appear for the first time in the Upper Paleolithic Period. In some cases they are very schematically formed, and it is often difficult or impossible to recognize female attributes. In other cases, however, they are naturalistic representations of corpulent women whose secondary sexual characteristics (their breasts and buttocks) were given special prominence, though their faces, feet, and arms were almost completely neglected. Such strong emphasis on the anatomical zones that are related to the bearing of children and nourishing them easily conveys to one the idea of female fertility. Nevertheless, it is not necessarily true of all these small figures.

Dortes—B. Arthaud



Female holding a bison horn, stone bas-relief from Laussel, Dordogne, France. Gravettian period (25,000/20,000–18,000/15,000 BC). In the Musée d'Aquitaine, Bordeaux, France. Height of figure 37.72 cm.

Ethnological analogies with present-day primitive phenomena offer the equally plausible view that such figures were regarded as the representations of the abodes of spirits whose function was to help and protect, and especially during hunting. They also may have been conceived, among other things, as mothers or rulers of the animals, goddesses of the underworld, helpers during hunting and donators of game, and as sovereigns of the land and other regions and of natural forces, including that of fertility.

No known direct continuum connects these earlier Paleolithic figures to similar ones of the early Neolithic and later periods. In settlements and shrines of these later periods are found large numbers of female figurines of widely differing types. They may have been representations of deities and symbols or, perhaps, votive offerings, somehow connected with female fertility. This can be safely assumed for figurines that show an obvious indication of fertility or are connected with children, and even more for shrines containing figures with sculptured pairs of breasts, and figures on the walls of women in childbirth. Not all female figures can, however, be understood merely as fertility symbols; rather, in many cases they are assumed to be house gods or representations of ancestors, and, especially when appearing in graves, as substitutes for the bodies of maids, wives, and concubines. An appearance of a large number of smaller figures suggests a votive or magical usage.

Shamanism, sorcery, and magic. Shamanism is a rather variable and highly stratified complex of practices and conceptions; characteristic among these are the use of ecstasy, the belief in guardian spirits (who are often in animal form, with the function of helping and guiding the dead on their voyage to the beyond), and beliefs concerning metamorphosis (change of form) and travelling to the beyond (see also SHAMANISM). Pictures from the Upper Paleolithic Period indicate the existence at that time of ecstatic practices and of beliefs in protective and helping spirits, which assume the form of birds and other animals. On the other hand, it is doubtful whether shamanism existed in fully developed form at that time. Also, in the course of prehistory, objects appear that may well have belonged to the paraphernalia of shamanism. Noise-making objects (to drive away evil spirits) are often found in the material remains of the Iron Age and probably are connected with shamanism.

Belief in guardian spirits

Recent studies stress the religious character of shamanism, though in practice it is related to sorcery and magic. Shamanism is not to be identified with sorcery and magic if they are understood as attempts to manipulate the supernatural through certain human techniques, in contrast with religion, in which man approaches higher beings (gods) in an attitude of supplication. Magic or sorcery thus appears as the opponent of true religion and gains importance when religion declines or is overwhelmed. In fact, magic and sorcery may take over cultic forms and rob them of their religious meaning when this occurs. For these reasons, it is often difficult to decide whether prehistoric phenomena were of religious or magical character.

Magic also can be practiced to a large degree without the use of material objects, and it is, therefore, as hard to grasp archaeologically as true religion. In the interpretation of the art of the Upper Paleolithic, scholars have given great importance to magic because, for example, missiles (spears and arrows) were drawn on the pictures of animals. This has been interpreted to mean that an effort was made to insure and compel the success of hunters through magical action. But this interpretation is

highly speculative and it remains uncertain what these drawings mean. It is just as difficult to decide whether or not other pictures, sculptures, abstract symbols, amulets, and similar objects were used to make magic in this and later periods.

HISTORICAL DEVELOPMENT

Religious patterns and economic stages. Religion is always closely related to other realms of life, such as economic activities. These relations are partly direct and partly mediated by social forms. The latter are, on the one hand, at least partially dependent on economic conditions; on the other hand, social structures influence the formation of religious phenomena and often serve as models for their elaboration. In a negative sense, then, it is often possible to eliminate certain religious phenomena as inappropriate to a particular society. It is inconceivable, for example, that the religious conception of simple hunters and gatherers included an elaborately organized hierarchy of gods with detailed division of labour between the individual figures. Similarly, it is a mistake to attribute to hunters and gatherers conceptions that are bound up with agriculture and the fertility of fields. In a positive sense, however, certain economic and social conditions will encourage the development of certain corresponding religious conceptions. Animalistic notions will be especially effective in situations where animals play a large role as partners of humans. Nevertheless, the spiritual ties to animals will be considerably different among hunters or agrarian peoples who still find it necessary to rely heavily on hunting for their meat supply as compared with pastoral peoples. In fully agrarian cultures, on the other hand, ideas about the fertility of fields and cultivated plants play an important part; they are connected with other notions about fertility and influence other spheres of life.

Burial sites that indicate a belief in life after death

Religious patterns in the various periods. *Lower or Early and Middle Paleolithic.* The oldest burials that attest to a belief in life after death can be placed in the period between about 50,000 and 30,000 BC. The earliest evidence of human activity in any form, on the other hand, goes back more than 1,000,000 years. Yet, since religious conceptions are not always bound to material objects, and since there is evidence that truly human beings existed even during early Paleolithic times, it is inadmissible to infer that earliest man had no religion from the mere fact that no identifiable religious objects have been found.

A study of very simple hunters and gatherers of recent times shows that several religious conceptions generally considered to be especially "primitive" (*e.g.*, fetishism) hardly play an important part, but rather that, among other things, the supposedly "advanced" conception of a personal creator and preserver of the world does play an important part. Such a belief could never be discovered by examining archaeological sources—the material remains—and hence cannot be ruled out for the Early Paleolithic Period. Whether or not the sacrifices in that era involved divine creators or preservers or other beings can only be a matter of conjecture. Features of animalism, magic, and various other views and practices may have played a role, but probably less so than in later epochs.

Upper Paleolithic and Mesolithic. The animalistic features encountered in the art of the Upper Paleolithic Period were most likely only a part of the religion that existed at that time. Among present-day "primitives" the animalistic realm often occupies only a lower sphere of what can be considered religious, and beyond and above that sphere are still other notions about gods. Practices concerning the resurrection of animals and the preservation of species evidently also played an important part and were closely tied to animalistic conceptions. The corresponding rituals clearly took on a special significance in relation to bears and became the basis for the bear ceremonies that were later widely diffused. Although shamanism may have been initiated somewhat earlier, it was now evident, at least in some of its aspects.

The realm of hunting was primarily a masculine sphere; nevertheless, it also includes in religious phenomena the feminine aspect, as symbols of female fertility (and probably also of female deities) demonstrate.

Proto-Neolithic and Neolithic. The characteristics of early religion were continued but transformed in the proto-Neolithic and Neolithic periods. Shamanism developed, especially among the pastoralists of central and north Asia. Animals, viewed as the hypostases (essences) of higher beings, especially the eagle or falcon and the raven, became highly significant in shamanism. Animalistic conceptions continued and often assumed the proportions of a true animal cult. Hoofed animals, especially sheep and oxen, played an important part as sacrifices and bulls particularly assumed a leading role; they seem to have been relegated to the masculine sphere. Horses appear as domesticated animals and as sacrifices only toward the end of the Neolithic Period. They may have been connected with a heavenly divinity, as later evidence suggests.

In the early period of agriculture, before the full development of the Neolithic Period, deposits of human skulls appear that suggest the presence of ancestor cults. A spiritual identification between humans and plants apparently played a predominant part in conceptions connected with headhunting and cannibalism. The death of a god was often considered a prerequisite to the appearance and prospering of the plants, and this mythical event was repeated through human sacrifice that was either accompanied by or replaced by animal sacrifice.

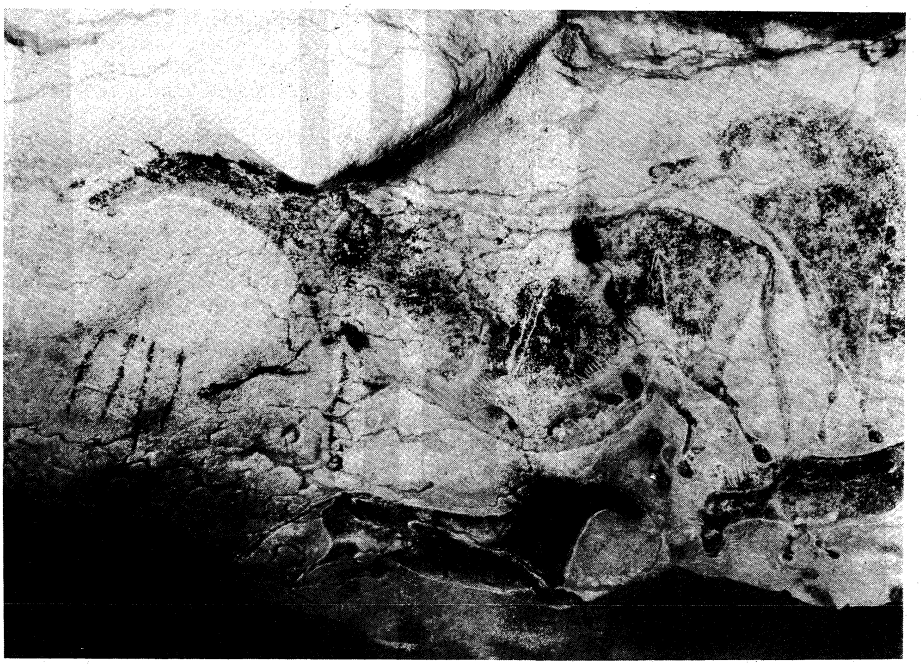
Dualism of masculine and feminine elements

At an early stage, in addition to an agricultural connection with the earlier feminine aspects, the masculine aspect appears in the form of portrayals of sexual union and, perhaps, of the "holy wedding," or sacred coupling, as well as in portrayals of couples and families. Among the material remains, however, the direct representation of the male element recedes sharply, yet perhaps the symbol of the axe and probably also that of the bull may indicate the male element. This dualism of the masculine and feminine aspects can possibly be interpreted in terms of father sky and mother earth, and in their union as a couple by which they become parents of the world. In the early civilizations, the conception of a supreme being or a heavenly god (which cannot clearly be recognized either in pictures or in other material objects) plays a minor role. That does not mean, however, that such a conception is necessarily of recent origin but rather that it probably existed at an early period in places where there was no literate tradition (predominantly among pastoral cultures).

Prehistoric religions and religions of the early civilizations. The decisive factors that brought about the early civilizations were the new kinds of economic and social organization, the large-scale exploitation of human energy, the formation of ruling classes, hierarchical organization, and the administrative division of labour. Under such conditions polytheism, which had undoubtedly been nascent before, could develop fully. The social order is mirrored in the conception of city and state gods and of a hierarchically organized "state of gods" with a division of labour. The concentration of power and people in one place, in contrast with the wandering of earlier nomadic cultures, enabled fixed central shrines to become influential. Yet the old traditions continued, and not least among them, that of animalism, in the form of conceptions about a ruler of the animals, animal cults, and similar phenomena. Female fertility figures remain generally prominent, such as the Great Mother and the Earth Mother.

PRESENT STATUS AND PROSPECTS FOR THE STUDY OF PREHISTORIC RELIGION

The material remains of prehistoric times naturally furnish more information about economics, technology, and society than about spiritual meanings. The poverty of the sources and the conjectural nature of what can be said point to the fact that the study of prehistoric religion remains predominantly at the periphery of scientific research. Its future expansion will depend not so much on an increase of source material and of research, but on whether or not it progresses methodologically—in its



Cave painting of a pregnant mare struck by arrows. Design at left is variously interpreted as a trap or a magical symbol meant to ensnare. Upper Paleolithic Period (20,000–10,000/8000 BC). Lascaux, Dordogne, France. Height 109.86 cm.
Jean Vertut—Ziolo

methods of inquiry and interpretation and in its basic theoretical concepts.

Knowledge of prehistoric religion will always remain fragmentary. Every attempt to sketch broad outlines of the area contains its own particular dangers. On the one side are the enthusiasts who try to make too much out of meagre sources; and on the other, the positivists, who set unnecessarily narrow limits to their interpretations. The use of preconceived evolutionary concepts is just as dangerous as undertaking a reconstruction that does not take into account the fact that the chronological sequence of material remains that reflect religious phenomena is not the same thing as a developmental sequence of great religious ideas. Some day perhaps the proper combination of factual, theoretical, and reconstructive capacities will provide a more satisfactory picture of prehistoric religion.

BIBLIOGRAPHY. Comprehensive presentations of prehistoric religion are rare and for the most part not written in English. A useful collection of materials on the Paleolithic, Mesolithic, and Neolithic-Chalcolithic periods is E.O. JAMES, *Prehistoric Religion: A Study in Prehistoric Archaeology* (1957). Even though it is antiquated and restricted to the Paleolithic Period, a classical work that is still very useful is T.H. MAINAGE, *Les Religions de la préhistoire: L'âge paléolithique* (1921). The method of historical-cultural comparison, which is used in this book, is consistently worked out after the "Vienna school" in J. MARINGER, *Vorgeschichtliche Religion* (1956; Eng. trans., *The Gods of Prehistoric Man*, 1960). The same basic concept, restricted to the most simple forms of culture, is shown by F. KERN, *Der Beginn der Weltgeschichte* (1953; Eng. trans., *The Wildbooters*, 1960). K.J. NARR, "Approaches to the Religion of Early Paleolithic Man," in *History of Religions*, 4:1–22 (1964), evaluates the problem in the perspective of the psychic status of early man. Further articles by the same author on specific problems of the religion of the Upper Paleolithic are: "Weibliche Symbolplastik der Älteren Steinzeit," in *Antaios*, 2:132–157 (1961) and "Bärenzeremoniell und Schamanismus in der Älteren Steinzeit Europas," in *Saeculum*, 10:233–272 (1959). No comprehensive presentation of the Neolithic Period has yet been written. The following book offers fundamental material for one of its important problems: G.E. DANIEL, *The Megalith Builders of Western Europe* (1958). The conception of a Megalithic mother goddess in the above book is criticized by A. FLEMING in "The Myth of the Mother Goddess," *World Archaeology*, 1:247–261 (1969). Also dealing with the mother goddess motif is E.O. JAMES, *The Cult of the Mother Goddess* (1959). Justified hesitations against a hasty or premature interpretation of female figurines portrayed as goddesses are brought forward by P.J. UCKO, *Anthropomorphic Figurines of Predynastic Egypt and Neolithic Crete* (1968). Another, probably exaggerated, interpretation of the Megalithic Period in its emphasis on the heavenly god is D.J. WOLFEL, "Die Religion des vorindogermanischen Europa," in F. KONIG (ed.), *Christus und die Religionen der Erde*, vol. 1, pp. 161–537 (1951).

(K.J.N.)

Preschool Education

Preschool education involves the earliest phases of childhood, beginning in infancy and ending upon entry into primary school at about five, six, or seven years of age (the age varying from country to country.) Although preschool education is ordinarily not compulsory, there exists in most countries some variety of preprimary schools, usually both public and private (most Socialist countries have only public facilities). Preschool education differs from the services of day-care and play centres in that the latter are usually merely custodial, providing care, safety, and health services for the children of working mothers or for children who cannot otherwise be cared for at home. The aim of preschool education is to foster the physical, emotional, social, and intellectual development of children—the relative emphasis depending on the prevailing philosophy and goals of the society.

Variety of preschool institutions

The institutional arrangements for preschool education vary widely around the world, as do the names applied to the institutions. The terms usually given to centres for the care of infants—those in the first phase of childhood (about three months to three years of age)—are infant school, nursery school, day nursery, and crèche—the term crèche being used not only in French-speaking countries but also in such places as Scandinavia, the United Kingdom, Czechoslovakia, Poland, the Soviet Union, and Israel. Nursery schools, or crèches, however, may admit older children as well; in Sweden, for instance, crèches welcome children from the age of three months on, until they enter kindergarten at four or five; in English-speaking countries, nursery schools include children of all ages under five or six, because education beginning in the fifth or sixth year is customarily considered a part of primary education.

For the second phase of early childhood, other institutional names and arrangements exist, the most common being the kindergarten and the "maternal school" (*école maternelle*). Typically, the maternal school (for ages

three to four or five) precedes kindergarten (for ages four or five to six), but in some countries—Italy, for instance—a child goes from the maternal directly to the primary school. In Germany, in addition to the *Kindergarten,* there is the *Schulkindergarten* (school kindergarten), which is for children of school age who are not considered sufficiently mature and which therefore serves as a kind of preparatory school for primary school. In the United States, kindergarten is considered a part of primary education.

Many countries have also instituted special schools: in the Soviet Union there are sanatorium kindergartens for children who are either physically handicapped or recuperating from illness; in the United States there are kindergartens for exceptional children, who are mentally retarded or otherwise physically or mentally handicapped.

BACKGROUND OF PRESCHOOL EDUCATION

The earliest institutions for the very young appear to have been philanthropic, established mainly as substitutes for home care rather than as educational supplements to the home. Most of these early institutions were for destitute or homeless children and were frequently in the hands of religious orders.

Early schools. The name usually associated with the initiation of early childhood education in modern times is Johann Friedrich Oberlin, an Alsatian Lutheran pastor in Waldersbach, who took it upon himself to educate his townspeople and, among other things, founded in 1767 the first *salle d'asile* (literally, "hall of refuge"), or infant school, for the care and instruction of very small children while their parents worked in the fields. Young women were employed to teach the young to play and the slightly older to spin, sew, and engage in handicrafts; all were engaged in nature study, with short field trips undertaken to gather flowers and other specimens. Local opposition to Oberlin's innovations were strong, but the French government eventually awarded him a money prize, and Germans as well as French began imitating his infant school—in Lippe-Detmold, Berlin, Kaiserswerth, Paris, and elsewhere. In France, the *salles d'asile* changed from private to state-supported institutions in 1833 when they were made part of the national educational system. Later, their name was officially changed to *écoles maternelles.*

Salles d'asile, or early maternal schools

Seemingly independently of the Continental infant-school movement, the Scottish reformer Robert Owen in 1816 founded in his model community New Lanark an "Institute for the Formation of Character." It served about 100 children of the workers in his cotton mills, mostly from 18 months to 10 years of age; and there were separate infant classes for two- to five-year-olds, who spent half their time in instruction and half in recreation. There they were "trained to associate in kind and friendly manner with their little companions; thus practically learning the pleasure to be derived from such conduct in opposition to envious bickerings or ill-natured disputes." No punishment was employed in the process of education. In bad weather, instruction was conducted indoors; in good weather, out-of-doors for the "benefit of sufficient exercise in the open air."

First infant school

The New Lanark school led to the establishment of England's first infant school in London in 1818. This school was set up by the man who had directed Owen's institute, James Buchanan, and it cared for children aged one to six years. According to contemporary accounts, Buchanan brought to London the methods that he had evolved at New Lanark:

> He began with simple gymnastic movements, arm exercises, clapping the hands, and counting the movements. *Viva voce* lessons followed, arithmetical tables, etc. . . . Watt's Divine and Moral Songs and similar hymns soon followed, and the children never tired of singing them to the accompaniment of his flute. He also gave the little people simple object lessons in which they did most of the talking, and learned to observe and describe.

Buchanan's school was imitated by others, notably by Samuel Wilderspin, who wrote some of the earliest and most widely disseminated monographs on infant education. It may be noted that the founders of the *salles d'asile* in Paris in the 1820s drew upon the work not only of Oberlin but of Buchanan and Wilderspin as well and seem even to have derived their chief inspiration from the English experience.

Meanwhile, in the late 1820s in Italy a Roman Catholic father, Ferrante Aporti, while working to establish elementary schools, became dissatisfied with the progress made by his pupils and was led to investigate the conditions under which their preschool years had been passed. Among other things, he read a translated work of Wilderspin's and, as a result, established Italy's first infant school in Cremona in 1829 and devised an educational plan that aimed at a harmonious combination of moral, intellectual, and physical education. Manual work, at all educational ages, was to give education a certain concreteness and rationality, making it a process of pupil involvement; the very young were to start off becoming accustomed to discipline, friendly cooperation, and piety.

The chief drawback of all these early efforts—at least the chief drawback as seen from a modern viewpoint—is that the schools for infants were largely miniature copies of schools for older children, or they became so as time passed and as public authorities assumed a larger role in the direction of infant schools or *écoles maternelles.* Young children were subjected to books and lessons and training in "good habits." Three-, four-, and five-year-olds were required to sit still in rows in large classrooms, recite lessons in which pictures took the place of living things or activities, and spend hours on reading, writing, and arithmetic.

Froebel and later reformers. With Friedrich Froebel, the German founder of the kindergarten, there arose the first systematic theory of early childhood pedagogy: instead of considering early schooling a form of babysitting or social philanthropy or considering it merely a period of preparation for adult roles, Froebel saw early childhood development as a special phase during which the child expresses himself through play. Child's play was a spontaneous expression of the "divine within him"; it was a "free manifestation of the inner world," a process of discovery and recognition that educated the child to the unity, as well as the diversity, of things in nature. These educational premises guided Froebel's pedagogical institute at Keilhau (founded in 1816), but it was not until 1837 at nearby Bad Blankenburg that he opened his first infant school, which he later called a *Kindergarten,* or "garden for children." He called his school a garden for children for the simple reason that he wished his children to grow and develop as do things in nature. He pointed out that "young animals and plants are given rest, and arbitrary interference with their growth is avoided, because it is known that the opposite practice would disturb their pure unfolding and sound development," but regrettably this has not been the practice with the "young human being," who has been "looked upon as a piece of wax, a lump of clay, which man can mold into what he pleases."

At his kindergarten Froebel sought to put his theories into practice. He devised a collection of geometric playthings (or "gifts," as he called them) that were somehow to symbolize, for the child, nature's concurrent unity and diversity—the sphere representing unity, the cube diversity, and the cylinder a blending of the two. Other figures—solids, planes, lines, and dots—were also included, and various exercises or occupations, such as folding, cutting, and weaving, were offered to make the symbolic forms real or dynamic for the child. Froebel believed that the young child learned best not through formal instruction but through play and imitation, "self-activity." He felt, "All true education in training and instruction should . . . at every moment, in every demand and regulation, be simultaneously double-sided—giving and taking, uniting and dividing, prescribing and following, active and passive, positive yet giving scope, firm and yielding." Since Froebel's day, educators have viewed the preschool years as a period of great educational possibilities. Indeed, his educational plan gained widespread acceptance. Within 25 years after Froebel's death in 1852, kindergartens were

founded in leading cities in Austria, Belgium, Canada, Germany, Great Britain, Hungary, Japan, The Netherlands, Switzerland, and the United States.

In 1892 in Italy, the Agazzi sisters, Rosa and Carolina, initiated a blending of Aporti's infant school and Froebel's kindergarten and produced a prototypical Italian maternal school (*scuola materna*). Rosa Agazzi agreed with Froebel's notion that within the child himself were motivations for self-improvement and discovery and "humanity." In the school the children were thus induced to become collaborators in the search for the instruments of their own education—seeking *realia* (objects from real life) as well as Froebelian symbolic objects to examine. The point of such education was to allow each young human being to find his own singular life-style through natural processes of socialization and acts of responsibility.

The Montessori method

Similarly concerned with nurturing or favourably exploiting the young child's natural impulses—in a safeguarded, constructive way—was one of the most famous figures in preschool education, Maria Montessori, a graduate in medicine and a researcher in experimental psychology and pedagogic anthropology. She began her studies of educational problems with culturally deprived and mentally deficient children in 1899, when she became director of the Orthophrenic School, in Rome. There she achieved remarkable results, for children under her tutelage passed the state examination in reading and writing for normal children. Because her methods worked with defective children, she felt that they might yield even better results with normal children. Thus, in 1907 she took under her care 60 children, aged three to six, from the slums of the San Lorenzo quarter of Rome and thus inaugurated her Casa dei Bambini ("Children's House"). This and later Case dei Bambini justified her hopes, and the schools became world famous.

Individual initiative and self-direction characterized the Montessori philosophy, and thus the teacher was to withdraw to the background and merely supervise the use of "didactic materials," which the children used individually unless helping or being helped by other children. The aim was to train children to be inquirers—so sensitive to the specific characteristics of things about them and so expert in classifying them that everything would possess interest and value for them. The didactic materials were a large complex of educational tools that Dr. Montessori herself developed—such as lacing frames, number rods to develop concepts of numbers, map puzzles, sandpaper letters that children were to trace with their fingers. Also, although usually each child worked alone, group or social activity was not ignored, for there were group gymnastics, games, and religious exercises; and social manners were taught in serving meals, waiting on tables, and the like. The children learned to read, write, and count and to express themselves artistically.

Montessori's method was based on the theory that children go through a series of "sensitive periods," during which they are most susceptible to learning certain skills, such as writing (starting in sand and then progressing to paper), basic arithmetic, and so on. When these periods are missed, or when the child is held back by the teacher, difficulties in learning reading or number concepts may result. Thus, they must be given a measure of freedom (though not license), each child working at his own pace without the tensions of competitiveness. Each child must achieve a sense of order, however; Montessori strongly believed that real freedom demanded both self-discipline and respect for authority. As one of her major achievements, Montessori put to practice the idea that early childhood represents a distinct phase of maturation.

Across Europe, in Belgium, meanwhile, another doctor of medicine, Ovide Decroly, was pioneering in the education of the very young, also proceeding from the psychological study of abnormal or exceptional children. In 1907 he opened his École de l'Ermitage ("School of the Hermitage"), near Brussels. Unlike Montessori's children, however, Decroly's worked in groups; and like the Agazzis' children, they worked with real things drawn from everyday life. His educational system was based upon three processes: observation, expression (oral, written, manual, artistic), and association of space and time. He stated that the young child started to explore his immediate environment and then proceeded to a broader one. He felt the universal needs of the child to be food, protection against danger, endurance for the frustrations of life, work, play, self-evaluation, and self-discipline.

Across the channel in Great Britain were two pioneers in the movement to improve the health and environment of the very young: Grace Owen and Margaret McMillan. Both saw the nursery school as a place fostering health and physical development (prerequisites to any other kind of development) and as a place that should be an extension of the home. Owen wanted every housing development to have a nursery school, where children of various ages would comprise a group resembling a large family and where play would facilitate socialization. McMillan wanted the nursery school to make up for any neglect that the child experienced at home (she worked chiefly with underprivileged children in the London slums) and to provide a kind of compensatory education and care; she had regular meetings with the mothers and required her staff of teachers to live in the neighborhood of the nursery school. She outlined a plan for a three-year course for training teachers for the nursery schools, maintaining that only trained personnel should work with children from three to six years of age. Training centres at Manchester (under Owen), at Deptford (under McMillan), and at London supplied nursery teachers for the entire British Commonwealth as well as for the early nursery schools in the United States. In short, the training of teacher-nurses was important in the theoretical systems of these women; moreover, McMillan felt that one of the main tasks of the teachers should be to help to develop imagination in the young children because "reason is a later growth, depending very much . . . on the previous upspringing of the creative power."

Nursery schools as supplements to the home

The first decade of the 20th century saw the start of what might be called "collective" upbringing. In what was then Palestine the new settlers established kibbutzim, in which were established separate homes for the children in order to free the mothers to work in the commune. As the system has now evolved, all children of a kibbutz from birth to one year remain in an "infant house," cared for by a *metapelet* (upbringer) in charge of four or five babies. During the nursing stage mothers feed their babies in the infant house. The "toddler house," containing about eight children one to three or four years old, emphasizes socialization. All children visit home daily for a few hours. In the next stage, kindergarten, the child three or four to seven years old is under the care of a teacher and her three assistants (*metapelet*). The aim of this period is readiness for the first grade.

Another variety of collective preschool education is found in the Soviet Union, where crèches and kindergartens (*detskiye sady* and *yasli*) were inaugurated about 1919, partly through the persuasions of N.K. Krupskaya (the wife of Lenin), who viewed preschool education as the first step in creating a new Soviet man. Today, children are placed in the crèches (voluntarily) from two months until three years of age; these crèches are under the jurisdiction of the Ministry of Health. The kindergarten, under the jurisdiction of the Ministry of Education, accepts children three to seven years old. All teaching materials used are didactic. The socialization process, respect for authority, and the subordination of individual needs to those of the collective are stressed. Self-discipline and self-reliance are key teaching objectives.

MODERN VIEWS OF PRESCHOOL EDUCATION

The proliferation of nursery schools and other institutions of preschool education in the 20th century can be traced to a number of developments: (1) a new scientific interest in early childhood, resulting from applications in the fields of psychology, medicine, psychiatry, and education; (2) recognition of the importance of child guidance and parent education; and (3) the efforts of individuals and agencies to improve the educational programs of day nurseries already established for the care of children of working mothers. Because the nursery school movement

has sprung from such a variety of social forces, no one type of school may be described as representative of the movement. Nevertheless, it is profitable to consider a few modern views of early childhood education.

Psychological theories. Psychoanalytic theory—with its concepts of unconscious motivation, early stages of libido development, the concept of id, ego, and superego as integral parts of the personality—has clear relevance to early childhood education; and indeed Anna Freud, together with her followers, applied many of her father's and her own concepts to her work with very young children. In the 1950s the American psychologist Erik Erikson enriched the traditional ideas of child development by viewing early childhood as a sequence of stages of development during which there took place important "crisis encounters"—inner conflicts such as trust versus mistrust, autonomy versus shame and doubt, initiative versus guilt. The successful resolution of these crisis encounters depended not only upon the child himself but also upon his milieu. From this can be deduced the importance of careful training for teachers of preschool children.

Theories of Piaget

A second major contribution has been the developmental psychology of Jean Piaget and his followers, who are convinced that children advance through rather regular stages of intellectual development. The first two periods—sensorimotor intelligence (from birth till two) as well as representative intelligence (from two to seven or eight)—relate to the field of early childhood. In the first stage (sensorimotor) the child learns to use his muscles and senses to deal with external objects and events while his language begins to form. He also begins to deal with and know that things exist even if they are beyond his sight and touch. He also starts to "symbolize" (represent things by word or gesture). In the second stage the child experiences the greatest language growth; words and other symbols become a way to represent both the outside world and inner feelings. At this time "magical explanations" make sense to him. He begins to gain a sense of symmetry. His adjustments depend on learning by trial and error, and he manages things by intuition. He begins to integrate symbolization and elementary types of relationships, such as logical and mathematical relationships (grouping, sizes, quantities, and qualities) and spatial and temporal relationships. Piaget's theory laid the groundwork for recognizing the importance of cognitive learning processes and concept formation in the young child. Piaget also stressed the importance of an environment conducive to learning the necessary skills.

New areas of concern. One of the major concerns of nursery schools and kindergartens has been language development. Most investigators agree that true speech starts when the child begins to develop meaningful associations with the words he uses (an infant who imitates the word "mama" without understanding its meaning is not engaging in true speech). For a child between two and six, oral speech is a major task, involving both expression and comprehension. By about the age of four he has mastered the fundamentals of the systematic grammar of his language. By the age of six the average child has increased his vocabulary to about 2,500 words or so—the degree of increase depending on the environment, its variety, and on the willingness of adults to relate to the child. Many studies show that the very young child in an impersonal institution, such as an orphanage, generally lags in language development behind children of the same age in a family setting or in a setting with a close relation between family and nursery school.

Language development of the child

One of the many tasks of early childhood education is to provide training in elementary language skills for all children, but especially for those who need compensatory work. To improve their comprehension and speech there are listening and language games. Educators who find educational games a successful teaching device claim that they stimulate the children's interest in learning.

In recent years a new factor, that of nutrition or malnutrition, has been discovered to have a significant bearing upon early child development. The currently available data from different countries regarding the effects of malnutrition (especially protein and iron deficiency) upon the physical and intellectual growth of children indicate that it can cause mental retardation, especially if it happens during the first year of life. Some studies indicate that permanent changes in the brain may result. Other studies suggest that psychological changes brought about by malnutrition at an early age may persist later in life. In any event, the malnourished child usually appears apathetic, without initiative or energy; previously this "laziness" was stereotyped as a characteristic of certain minority ethnic groups or lower social classes.

Compensatory early childhood education calls for a well-balanced diet for the young child. Furthermore, to assure a good start for the newborn, attention must also be paid to meeting the mother's nutritional needs during pregnancy. In this sense, it may be said that preschool education, as a preparatory measure, has come to be concerned with the child even before birth.

The main goal of early childhood education is usually defined in terms of the prevailing philosophical, political, and religious trends of the given society and from one historical period to another. Thus, the objectives and methods vary from society to society. Nevertheless, there is general agreement that the nursery school is the first very important step in the education of the child because it helps him in achieving later success in school and because it starts to mold him in the image of the society that he is to perpetuate. In societies based on the concept of the collective, the preschool child is already expected to be a participating member; hence great stress is placed upon his socialization. In societies in which individualism predominates, the child is expected to become an individual and to fend for himself at an early age.

BIBLIOGRAPHY. One of the most detailed surveys of the great pioneer figures of early childhood education is ROBERT R. RUSK, *A History of Infant Education*, 2nd ed. (1951; reprinted 1967). Some of the more important classic works are H. SILVER (ed.), *Robert Owen on Education* (1969), containing a collection of Owen's writings; MARIA MONTESSORI, *Il metodo della pedagogia scientifica* (1909; Eng. trans., *The Montessori Method*, 1912, reprinted 1964); MARGARET MCMILLAN, *The Nursery School*, rev. ed. (1930); HARRIET M. JOHNSON, *Children in the Nursery School* (1928, reprinted 1971); and JEAN PIAGET, *Le Langage et la pensée chez l'infant*, 4th ed. (1956; Eng. trans., *The Language and Thought of the Child*, 3rd ed. rev., 1959), and *La Formation du symbole chez l'infant*, 2nd ed. (1959; Eng. trans., *Play, Dreams, and Imitation in Childhood*, (1962). One of the most comprehensive books of readings is JOE L. FROST (ed.), *Early Childhood Education Rediscovered* (1968). The two classic works dealing with communal education in Israel are BRUNO BETTELHEIM, *The Children of the Dream* (1969); and MELFORD E. SPIRO, *Children of the Kibbutz*, new ed. (1965). The most important work on the education of the young in the U.S.S.R. is Н.К. КРУПСКАЯ, *Дошкольное воспитание* (1959), which is vol. 6 of the author's collected writings. An overview of Soviet pedagogy in English is contained in HELEN B. REDL (ed.), *Soviet Educators on Soviet Education* (1964). See also various publications of the UNITED NATIONS EDUCATIONAL, SCIENTIFIC, AND CULTURAL ORGANIZATION (such as the annual *International Yearbook of Education*) and the WORLD ORGANIZATION FOR EARLY CHILDHOOD EDUCATION (such as reports of conferences).

(H.B.R.)

Prescott, William Hickling

A United States historian best known for his dramatic narratives of Spanish conquests of Mexico and Peru, Prescott has been called America's first scientific historian.

William Prescott was born in Salem, Massachusetts, on May 4, 1796, of a prosperous, old New England family. He received three years of rigorous instruction in a preparatory school headed by the Jesuit John Gardiner, who infused in him a love of classical learning. In 1811 he entered Harvard, where his academic record was good but undistinguished; he had serious difficulties with mathematics, and in later life the prospect of appraising the mathematical achievements of the aboriginal Mexicans almost prevented him from completing his work. Near the end of his junior year, a crust of bread thrown during a melee in the student commons caused virtual blindness in his left eye; the weakness of his other eye, caused by

His impaired vision

infection, sometimes prevented him from carrying on any kind of literary work. Throughout his life, Prescott's vision seems to have fluctuated from good to total blindness; and he often resorted to the use of a noctograph, a writing grid with parallel wires that guided a stylus over a chemically treated surface. Substantial portions of all his books and correspondence were composed on this device.



Prescott, daguerreotype portrait, *c.* 1845.
By courtesy of the Metropolitan Museum of Art, New York, gift of I.N. Phelps Stokes, Edward S. Hawes, Alice Mary Hawes, and Marion Augusta Hawes, 1937

Following his graduation from Harvard in 1814, Prescott's health deteriorated after attacks from what appears to have been an acute type of rheumatism, with swellings in his larger joints and lower legs. He convalesced at his grandfather's home in the Azores and then, encouraged by an apparent recovery, toured Europe. After his return to Boston, he embarked upon serious historical studies, shunning a career in business or law because both occupations demanded more stamina than his delicate health and eyesight could allow. At least the youthful Prescott used this argument as late as 1820, when he married Susan Amory. With no obvious occupation, he was known as "the gentleman" by his Boston friends. His wife, as well as other readers, provided the eyes that helped Prescott at this time to embark on a literary career.

His first publication was a number of reviews and essays in the *North American Review* in 1821. Some of these were reprinted in *Biographical and Critical Miscellanies* (1845). His "Life of Charles Brockden Brown" (1834) in Jared Sparks's *Library of American Biography* served notice of Prescott's high abilities as a writer. Largely on the advice of his friend the teacher and writer George Ticknor and the later encouragement from the miscellaneous writer Washington Irving, Prescott turned to Spanish themes for his life-work. The appearance in 1838 of a three-volume *History of the Reign of Ferdinand and Isabella the Catholic,* the product of some ten years of work, was an agreeable surprise to Boston's literary world. This work launched Prescott's career as a historian of 16th-century Spain and its colonies. In his first and only such work, *A History of the Reign of Philip the Second, King of Spain* (3 vol., 1855–58), Prescott produced graceful, authoritative narratives of Spanish military, diplomatic, and political history that had no equal in their time. His modern popularity rests with his epic *History of the Conquest of Mexico* (3 vol., 1843) and his *History of the Conquest of Peru* (2 vol., 1847).

Working with a superb personal library of perhaps 5,000 volumes and with the help of such overseas associates as Pascual de Gayangos, the Spanish aide who discovered manuscripts and rare books for him, Prescott made rigorous use of original sources. His critical use of historical evidence was such that he might well be called the first U.S. scientific historian. The fact that Prescott's histories continue to be popular with scholars as well as lay readers after more than a century of criticism attests to their vitality and readability. Though further research has revised his view of 16th-century Spanish monarchy, Prescott's basic work is still judged to be generally fair and accurate. It was in narrating the Spanish conquests at that time that Prescott's republicanism penetrates his histories so as to colour his picture of the Spanish state and the aboriginal governments of the Aztecs and the Incas. Moreover, his New England Unitarianism made it difficult for him to appreciate the acceptance of the miraculous or supernatural among peoples of another age or to understand the peculiarities of the conquistadors.

Criticisms of his approach

Perhaps the most severe unfavourable criticisms of the *Conquest of Mexico* and the *Conquest of Peru* are based upon his romantic version of native civilizations, which later findings in archaeology and anthropology have found to be distorted. Prescott's failure to visit the historical settings of his narratives and to examine actual remains of the native cultures he described was partly responsible for this defect in his books. Yet the contributions to understanding of pre-Columbian civilizations were made by scientists after Prescott's time, and modern scholars have concluded that Prescott's historical narrative, based upon Spanish chronicles, is essentially sound. What Prescott hoped to do with his histories was to instruct and to entertain. His history was narrative and descriptive rather than philosophical or analytical. His colourful prose dealt with conquests, war, diplomacy, and politics—not with cultural, social, or economic themes. In his Spanish histories his concern was almost exclusively with the Spanish courtiers and other aristocrats.

Despite such criticisms, Prescott's achievements as a historian and as a literary artist were remarkable. For example, the persistent demand for the *Conquest of Mexico* has resulted in its publication in ten languages at least 200 times and that of the *Conquest of Peru* in 11 languages at least 160 times. He was the first English-speaking historian to reach a wide audience outside the Hispanic world with a history expressing the Spanish point of view. Spaniards, in Prescott's histories, were often forerunners of progress. Thus the Moors in Spain and the aboriginal peoples of Mexico and Peru make way for the achievements of Spanish characters. Throughout the conquistador histories, Prescott exposes the reader to vivid landscapes, battles, and processionals as the march of Spanish civilization overwhelms the savage world. Prescott's literary artistry convincingly shows the conquistador Cortés caught up in a series of crises that, on the eve of final victory, tend to become more and more complex. In the end, however, "pusillanimity" of the Aztec emperor Montezuma is the advantage that the forthright Cortés has in determining the outcome of events.

Prescott weaves a dramatic fabric that completely envelopes his narrative. Indeed, much of the same story is repeated in both the *Conquest of Mexico* and the *Conquest of Peru* in respect to descriptions of battles, characterizations, use of metaphor, dramatic encounters, and crises, suggesting that Prescott perhaps manipulated his narratives for literary effect. Yet critics generally agree that he accurately follows his sources. His empathy with the Spanish point of view still makes him the greatest Anglo-U.S. historian of the Hispanic world.

Prescott won for himself and for the United States the distinction of making a significant contribution to historical literature that was recognized throughout the world. Even in his lifetime his achievement was recognized, for, as his biographer George Ticknor maintained, Prescott's visit to England in 1850 was the most distinctive visit ever made to England by an American who was not a government official. By the time of his death on January 28, 1859, Prescott had proved that epic history could be written as literature. In this sense he prepared generations of readers for the polished writings of Boston's other literary historians, such as John Lathrop Motley, George Bancroft, and Francis Parkman. Tall, handsome in appearance, Prescott's personal charm was such that he was popular with the society of his day despite his delicate health and difficulties with sight. His interest in Boston's Perkins School for the Blind, as well as his histories, have served to remind later generations that his triumph over physical obstacles was in itself a heroic achievement.

BIBLIOGRAPHY. C. HARVEY GARDINER, *William Hickling Prescott: A Biography* (1969), is a critical biographical study of Prescott as a writer and as a historian based upon exhaustive research. Gardiner and all other writers on Prescott are indebted to GEORGE M. TICKNOR, Prescott's lifelong friend and associate, for *Life of William Hickling Prescott* (1864). Prescott's talents and faults as a writer are examined in DAVID LEVIN, *History As Romantic Art: Bancroft, Prescott, Motley, and Parkman* (1959); an appraisal of Prescott's scholarship and his significance is made in WILLIAM CHARVAT and MICHAEL KRAUS, *William Hickling Prescott* (1943). Additional sources on Prescott's life and work are C. HARVEY GARDINER (ed.), *The Papers of William Hickling Prescott* (1964) and *The Literary Memoranda of William Hickling Prescott*, 2 vol. (1961); and W.H. MONROE (ed.), *The Works of W.H. Prescott*, 22 vol. (1904).

(W.R.J.)

Pressure and Temperature, Biological Effects of

All biological activities depend upon the chemical reactions that constitute the metabolism of living things. These reactions are influenced by both temperature and pressure, and living things vary greatly in their ability to withstand exposure to pressures and temperatures beyond normal ranges of activity. Most animals are active between about 10°–40° C, although two species of blue-green algae are known to live in thermal springs at temperatures exceeding 85° C and the marine clam, *Yoldia arctica*, is found in waters colder than 0° C.

Physiological ranges of pressure and temperature

Terrestrial organisms at sea level are subjected to one atmosphere of pressure (14.7 pounds per square inch). As altitude increases, atmospheric pressure decreases, until, at the top of Mt. Everest, the highest point on earth at 8,824 metres (29,028 feet) above sea level, atmospheric pressure is less than five pounds per square inch. With the exception of some micro-organisms and insects, there are no living things at this altitude. Since water is 773 times more dense than air, aquatic organisms experience higher pressures than do terrestrial ones. Even though water pressure increases one atmosphere for each 10 metres (33 feet) of depth, a remarkable variety of life exists in the deepest marine trenches (11,000 metres, or 36,000 feet), where pressures exceed 16,000 pounds per square inch.

With regard to temperature, the earth provides a relatively moderate environment compared with outer space, where the temperature approaches absolute zero (0° K, or −273° C), and the surface of the sun, which is 6,000° K. The lowest temperature on earth (−68° C) occurs in Alaska and Siberia; the highest, in hot springs of New Zealand, may exceed the boiling point of water (100° C). Although most of the life on earth is found in regions of less extreme temperatures, some organisms are able to survive the natural temperature extremes cited above.

BIOLOGICAL EFFECTS OF PRESSURE

Effects of pressure on cells. In the 19th century, dredging expeditions disclosed that animals live at depths of 11,000 metres in the sea. Samples of the ocean floor have been brought to the surface under pressure, so that the single-celled organisms that normally live there under great pressures may be studied.

Many species of bacteria are able to grow at pressures of 200–300 atmospheres (atm); higher pressures, however, usually inhibit growth, which is measured as rate of reproduction. For reasons not yet known with certainty, bacteria isolated from ocean depths grow better at low pressures than at high ones, and attempts to grow them at the pressures from which they are obtained (more than 1,000 atm) have not yet been successful. The failure of most bacteria to grow at high pressures stimulated studies of their structure and biochemistry under conditions of high pressure. One of the most frequently observed effects of high pressures (exceeding 300–400 atm) on cell structure is the formation of long filaments, often accompanied by a thickening of cell walls. Several bacterial species develop swollen or bizarre shapes near the threshold of pressure tolerance. Biochemical studies have revealed that the amount of hereditary material, deoxyribonucleic acid (DNA), in the bacterium *Escherichia coli* is considerably less at high pressures than at normal ones; the decreased amount of DNA may be related to inhibition of cell division at high pressures.

The ability of micro-organisms to survive high pressures is influenced not only by the amount of pressure but also by its duration and the temperature at which it occurs; generally, the longer that cells are exposed to high pressure, the more it affects their structure and survival. The harmful effects on cells of high pressure over a long period of exposure, however, often are offset either by lowering or raising the temperature, relative to the temperature for optimum growth at one atmosphere; in a similar manner, high pressure appears to retard cell death at high temperatures. Certain bacteria that reduce sulfate to hydrogen sulfide during metabolism have been found to be able to survive the combination of highest temperature and pressure; sulfate-reducing bacteria have been maintained for several days at a pressure of 1,000 atm and a temperature of 104° C.

High pressures may produce a number of effects on both external and internal chemical environments of cells, which, in turn, may influence the rates of important chemical reactions. High pressure also may influence the structure of cellular proteins, including enzymes (organic catalysts) essential to the chemical reactions that occur in cells. The most complete analysis of the effects of pressure on chemical reactions has been derived from the Arrhenius theory, discussed below.

Arrhenius theory

About 1889, the Swedish chemist S.A. Arrhenius offered a theory to explain the increase in rate of chemical reactions with increases in temperature. The theory proposes that molecules require a certain amount of energy, called the activation energy, before they are able to enter into a reaction; the higher the activation energy of a molecule, the higher must be the temperature before a measurable reaction rate occurs. The theory, when applied to molecular volume changes that accompany chemical reactions, results in a determination of the volume change of activation—*i.e.*, the volume change that chemical reactants undergo in reaching a state of activation sufficient to produce a chemical reaction. The volume changes that occur during chemical reactions at ordinary pressures are small and of little importance; at high pressures, however, volume changes are important in interpreting the effects of pressure on biochemical reactions.

The rates of biochemical reactions may be either retarded or accelerated by increases in pressure, depending on volume changes that occur; a reaction involving increases in the volumes of the reactants, for example, is inhibited by increases in pressure, whereas a reaction involving decreases in the volumes of the reactants is accelerated by increases in pressure. Since most chemical reactions involve volume increases, they are inhibited by high pressure, so that many of the harmful effects of high pressure on cells may be the result of inhibition of essential biochemical reactions.

Despite the explanation provided by the Arrhenius theory, the problem of the ways in which micro-organisms are adapted to live on the ocean floor remains unsolved. The existence of life at such high pressures would appear to involve some mechanism for counteracting the inhibitory effects of pressure on cellular processes.

Effects of pressure on plankton. A number of factors influence the distribution of small organisms in the sea. In addition to high pressure, life in ocean depths is limited by low temperatures, lack of light, and shortages of food. As a result, much oceanic life is found within 50–100 metres (160–330 feet) of the surface; even near the surface, however, pressure, a major indicator of depth, is important. Most plankton remain in a relatively shallow zone. Some species, whose weight equals that of the water they displace, float within this narrow zone; active movements, however, are necessary for most species to stay within the limits at which they are able to live. Although light provides some information about depth near the surface, it varies in intensity, both with distance from the ocean surface and on a daily basis; changes in pressure, therefore, provide organisms with more accurate information on depth. Many species of invertebrate plankton,

which are sensitive to minute changes in water pressure, probably respond to these changes by using information received from specialized receptors that detect small changes in pressure associated with changes in depth.

Effects of pressure on plants. Marine plants must remain near the surface of the ocean since light is necessary for photosynthesis. Most plants, however, are unable to move actively and remain near the surface by increasing their buoyancy (ability to float), usually by secreting low density gases into specialized structures, such as vacuoles or pneumatocysts.

Although pressure influences marine plants, some interesting examples also are encountered in pressure effects in terrestrial plants—*e.g.*, the lifting of sap from lower than ground level to the tops of trees that may attain heights of 100 metres (about 330 feet). A suction pressure of at least −20 atm at the top of a tree is required to raise a column of sap 100 metres, to maintain this pressure, and to overcome the friction in the plant tissue (xylem) in which the sap is carried. Pressures are even more pronounced in other plants—*e.g.*, saltwater mangroves use suction pressures up to −60 atm to pull freshwater into the sap against the pressure of seawater, and, in parasitic plants such as mistletoe, suction pressures 10 to 20 atm greater than those of the host are maintained. Columns of water are maintained under large negative pressures in the small tubelike structures found in the xylem of trees because cohesion, or holding together, is a property of water confined to small spaces.

Effects of pressure on animals. Some animals living in the sea are able to float by depositing low-density substances (for example, fat) in their tissues. Most fishes and some invertebrate animals, however, float by using gases. Gases, because they contract as pressure increases and expand as pressure decreases, require special adaptations. In order to control buoyancy, these animals must maintain a constant volume of gas as the water pressure changes, and some aquatic animals have evolved unique structures for maintaining buoyancy under pressure.

The invertebrate cuttlefish, an agile and rapid swimmer, has a unique structure, called a cuttlebone, along its upper surface. The cuttlebone, composed of layers of many small chambers encased in a rigid network of calcium carbonate, has a bonelike appearance; part of the cuttlebone contains gas (primarily nitrogen), part liquid. A cuttlefish increases its floating ability, or buoyancy, by moving water out of the cuttlebone, thus increasing its proportion of gas-filled space; expansion of gas as pressure decreases is prevented by the rigid matrix surrounding the chambers. Buoyancy decreases when water enters the cuttlebone, reducing the proportion of gas-filled space. Further decreases in buoyancy occur as the cuttlefish sinks because the higher water pressure forces water back into the cuttlebone.

Swim bladder in fishes

In most fishes, buoyancy is maintained by the secretion of gases into a structure called a swim bladder. The swim bladder, normally about 5 percent of the weight of a fish, keeps the fish from sinking and is flexible. If a fish rises toward the water surface (with a decrease in pressure), the swim bladder expands, and gas must be removed if the fish is to regain its buoyancy. In some fish (the Physostomi) the swim bladder is connected to the esophagus, so that gas can pass out through the mouth. In many deep-sea fish (the Physoclisti), however, the swim bladder has no external opening, and gas is removed by the oval chamber (see Figure 1), which has a rich blood supply and is normally separated from the rest of the swim bladder by a ring of muscle. When the muscle relaxes, gases in the swim bladder pass into the blood, thus reducing the volume of gas in the swim bladder.

Fishes that swim into deep waters regain their buoyancy only if the volume of the swim bladder is increased by adding gas, a function that is performed by a special structure known as the gas gland, which transfers dissolved gases from the blood to the swim bladder (see Figure 1). Most fish transfer oxygen into the swim bladder against enormous pressures—*i.e.*, the pressure of oxygen dissolved in seawater and blood is about one-fifth of an atmosphere, but at 4,500 metres (about 15,000 feet),

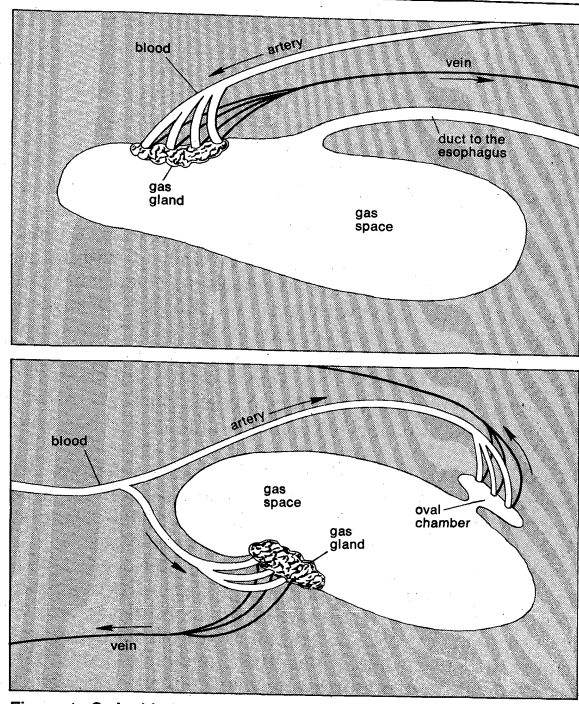


Figure 1: Swim bladders of a physostomid fish (top) and a physoclistid fish (bottom).
From E.J. Denton, *Progress in Biophysical and Biophysiological Chemistry* (copyright 1961); reprinted with permission of Pergamon Press Ltd.

the depth at which fishes with swim bladders live, the gas in the swim bladder is under a pressure of 450 atmospheres. In addition to the oxygen in the swim bladder, there are other gases, such as argon, that are present in measurable quantities. Since no known biochemical reactions involve this element, its transfer by the gland is probably a physical process.

Fishes obtain oxygen from air dissolved in water, and increases in pressure have no effect on this process. Air-breathing animals that enter or live in the ocean, however, face special problems. Air is taken into the lungs at the water surface, and the pressure increases that occur when the animals dive decrease the volume of the lungs, forcing gases into the blood; two things may happen as a result of the high concentration (80 percent) of nitrogen in air. First, large quantities of nitrogen under pressure readily dissolve in tissues, a condition that is toxic to a man diving in deep water, producing an anesthetic effect called raptures of the deep, or nitrogen narcosis. Second, if a diver rises rapidly to the surface, pressure decreases rapidly, the nitrogen in tissues comes out of solution, forms bubbles, and the result is the painful, and often fatal condition, known as bends, or caisson disease.

Although the above-mentioned problems affect man when he descends to considerable depths, many air-breathing animals live in the sea and dive to great depths; the sperm whale, for example, attains depths of 1,000–2,000 metres (3,300–6,600 feet) and returns quickly to the surface unharmed after being submerged for 1–1½ hours. Diving animals have a unique reflex for diving to great depths and staying submerged for extended periods of time without ill effects—before diving, they exhale, thereby reducing the volume of air in the lungs and helping to prevent nitrogen from entering the body; as this occurs, the rate of the heartbeat decreases, and blood flow to most parts of the body, except to the brain and heart, is reduced, allowing the diving animal to conserve the oxygen in the blood for vital tissues.

BIOLOGICAL EFFECTS OF TEMPERATURE

Effects of temperature on cells. As stated previously, Arrhenius proposed a theory to explain the effects of temperature on chemical reaction rates; his concept that molecules require a certain amount of energy (activation energy) before they are able to react was based on the observation that the rate of chemical reactions increases

with increases in temperature. Since higher temperatures provide more energy for reaction, the rate of cell functions increases at high temperatures and decreases at lower ones, within the physiological range. A descriptive expression for activation energy is temperature coefficient, or Q_{10}, which is defined as the change in the rate of a reaction when the temperature changes by 10° C; thus the rate of a reaction with a Q_{10} of two is doubled with an increase in temperature of 10° C and is reduced by one-half with a 10° C decrease in temperature. The normal temperature range of most cells is 10°–40° C; within this temperature range, the Q_{10} (temperature coefficient) for many processes is between one and three. For diverse phenomena, such as certain enzymatic reactions, rate of movement by insects, chirping of crickets, and flashing of fireflies, the Q_{10} is about two. Temperature coefficients for reactions involving physical processes, such as photochemical and diffusion phenomena, are usually less than 1.5.

Temperature coefficient

Determination of temperature coefficients allows descriptions of the effects of temperature on cellular processes within normal physiological limits. If the temperature increases above 45° C or decreases below 5° C, many types of cells die. Heat death, or cell death above 45° C, may be related to the effects of heat on proteins or, more specifically, on the enzymes that control most cellular functions; exposure of proteins to high temperatures causes structural alterations that inactivate many enzymes. High temperatures also may influence the structure of cell membranes. Since high pressure reverses many of the effects of high temperature (see above, *Effects of pressure on cells*), it may retard heat death. Cold death, or cell death below 5° C, may result from a general reduction of metabolic rate as the temperature decreases; if a reaction has a temperature coefficient (Q_{10}) of two, for example, a temperature decrease from 40° C to 0° C results in a 16-fold reduction in metabolic rate. In addition, either the internal structure of cells may be injured by ice crystal formation at the freezing point or the cells may become excessively dehydrated as water in them freezes.

Although most cells live within a temperature range of 10°–40° C, some live at higher or lower temperatures; bacteria and blue-green algae thrive in hot springs at temperatures above 60° C, and some molds and bacteria survive in ice cream at −10° C. Cells that live at high temperatures contain enzymes that are able to resist breakdown by temperature better than enzymes of cells that live within the normal range. The membranes of heat-resistant cells also are more stable at high temperatures than are those of cells within the normal range. The ability of cells to survive low temperatures, however, may be influenced by the rate at which they are cooled—*i.e.*, rapid freezing may result in ice formation, very slow freezing may result in supercooling, which is a decrease in temperature below the freezing point without ice formation. In nature, temperatures usually decrease slowly, often with the result that cells are supercooled; the temperatures of several species of fish that live in deep Arctic fjords, for example, are below the freezing point, but their cells are not damaged by ice formation. Some cells reduce the danger of cold death with an antifreezelike compound, which lowers the freezing point of the internal contents.

Although some cells are able to survive extreme temperatures of heat and cold, many temporarily suspend life activities (*i.e.*, become dormant) if conditions become unfavourable; bacteria form spores, protozoans and some invertebrate animals form cysts. These dormant forms are much more resistant to temperature extremes than are active forms; thus, temperatures up to 170° C for one-half hour may be required to kill spores and cysts, and some spores are capable of germinating even after exposure to temperatures as low as −272° C (see also DORMANCY).

Effects of temperature on plants. Unlike many animals, plants are not capable of generating heat in order to maintain temperatures independent of their surroundings; plant cells, therefore, are subjected to injury from either low or high temperatures, and their ability to function depends on the temperature coefficients (Q_{10}) of their cellular processes. Structural and functional adaptations also enable some plant species to survive harsh environments, although temperature changes may have pronounced effects on plant development.

Daily fluctuations in temperature, called thermoperiods, have dramatic effects on the growth of most plants. Temperature changes at night are particularly important, since most plants grow best when night temperatures are lower than day temperatures. The thermoperiod that produces optimum growth varies among different plants. For a given species, however, slight changes in night temperature can produce extreme changes in growth; if night temperature is not within a few degrees of 17° C, for example, fruit production of tomato plants is severely inhibited. The significance of the decrease in night temperature is not yet completely understood, although lower temperatures may act as triggers, promoting plant growth at the most favourable time of year.

Thermoperiods

Annual temperature changes are also important for plant development. In many species of trees, seasonal exposure to low temperatures is necessary for leaf formation the following spring. Artificial exposure of plants to low temperatures before transplanting, called vernalization, is practiced commercially to enhance fruit and seed production in many important crops.

Although the importance of exposure of plants to cold is known, the mechanisms by which some plants survive freezing temperatures in winter have not yet been established. Cell damage from ice formation and freezing of columns of sap in woody plants may occur. When water freezes to form ice, gases come out of solution, and small bubbles should form in plant tissue called xylem. When ice melts, the movement of sap under negative pressure should be hampered by pockets of gas in the xylem; seasonal changes apparently occur, however, in some plants, enabling them to withstand such freezing. Experimentally induced freezing of trees during summer months may cause damage that does not occur in winter. During the fall months, therefore, trees must become accustomed to, or acclimate to, cold; acclimatization in trees is not yet completely understood, although for some plant species it may involve water loss from cells (see also ACCLIMATIZATION).

Plants are exposed to high temperatures, sometimes up to 50° C, in deserts, which also lack water—a component essential for life and one that evaporates readily at high temperatures. A shortage of water may be a more important factor than high temperatures in limiting the number of successful desert-plant species. Nondesert plants grow at high temperatures as long as sufficient water is provided to compensate for high losses from evaporation; without adequate water, however, plants wilt as evaporation causes excessive water loss. Most desert plants adapt to their environments by storing available water, and they have thick, waxy surfaces that help to prevent excessive water loss at high temperatures. In addition, some plants lose their leaves during periods of drought when the temperature is high, thereby reducing water loss, since water evaporates during respiration through leaf openings (stomata); when water again becomes available, new leaves grow, and photosynthesis resumes. Drought in the hot desert has an effect on plants similar to that of cold in that both result in loss of leaves and temporary suspension of photosynthesis.

High-temperature adaptations of plants

Seeds provide a means by which many plant species are able to survive extreme cold, heat, or drought. Seeds resemble spores and cysts since all are able to remain dormant. The seed coat is a primary defense against the environment; it protects the plant embryo inside from excessive water loss, mechanical injury, and temperature extremes.

Seeds as protective devices

Effects of temperature on animals. Based on their ability to maintain an internal temperature, animals may be divided into two types—cold-blooded, or poikilothermic, animals, including all invertebrate animals, fishes, amphibians, and reptiles, that usually have about the same temperature as their environment; and warm-blooded, or homoiothermic, animals—mammals and birds—that maintain a relatively constant body temperature (about 37° C

for mammals, 40° C for birds) regardless of the environmental temperature.

Poikilothermic animals, greatly influenced by environmental temperatures, are limited, with only a few exceptions, to areas with a temperature range of 10°–40° C; some poikilotherms, however, compensate for extremes in temperature by taking advantage of the fact that most complex environments contain microclimates in which temperatures suitable for life may be found. Many desert reptiles, for example, seek the sun during cool mornings, retreating to cooler, shady burrows when the temperature increases. Some terrestrial poikilotherms maintain remarkably constant body temperatures by seeking out microclimates. Interestingly, the microclimates are usually close to 37° C, the approximate body temperature of homoiothermic animals.

Since homoiotherms maintain constant body temperatures, they are able to remain active in situations in which poikilotherms cannot. Body temperatures of homoiotherms are kept at a constant value by regulatory mechanisms that counteract the effects of the external environment. In cold environments, therefore, regulatory mechanisms maintain body temperature by increasing heat production and decreasing heat loss; in hot environments, regulatory mechanisms maintain body temperatures by increasing heat loss. Within a neutral range (27°–31° C for man) neither heat gain nor heat loss is necessary to maintain body temperature (see also HOMEOSTASIS).

The metabolic rate of homoiotherms exposed to temperatures below the neutral range increases (in contrast to the decrease that takes place in poikilotherms), and the heat thus produced counteracts the increased heat loss due to the low temperatures, so that body temperature remains constant. In order to control the rate of metabolism, homoiotherms have receptors that detect changes in body temperature; some are located in a specific area of the brain called the hypothalamus. When this area of the brain detects a temperature decrease it activates nerves that cause reactions in muscles (*i.e.*, shivering), which increase metabolism and heat production.

Although the increase in metabolism resulting from responses to decreases in temperature enables homoiotherms temporarily to increase heat production, animals in permanently cold areas would have to expend excessively large quantities of energy to maintain a high metabolic rate. Arctic homoiotherms, instead, have evolved two mechanisms to prevent the loss of heat from the body. One is increased insulation; Arctic mammals have thicker fur, Arctic birds thicker feather coats, and whales in Arctic seas thicker layers of blubber than do similar animals from temperate regions. The entire body is not covered with insulation, however; *e.g.*, the feet of most birds have no feathers, the nose of a wolf has no fur. The second mechanism for decreasing heat loss occurs in the extremities and is based on opposite direction of blood flow in arteries and veins. When it is warm, most of the blood returning to the body from extremities is carried in veins close to the skin (Figure 2). When it is cold, however, the blood returning to the body flows in veins that are close to the arteries, and veins close to the skin constrict (Figure 2). Some of the heat from the arteries, which carry warm blood to the body extremities, passes to (and therefore warms) the cooler venous blood coming from the extremities. This mechanism, which is called countercurrent heat exchange, is one of the important ways for Arctic animals to prevent body heat loss.

From K. Schmidt-Nielsen, *Animal Physiology* (1964)

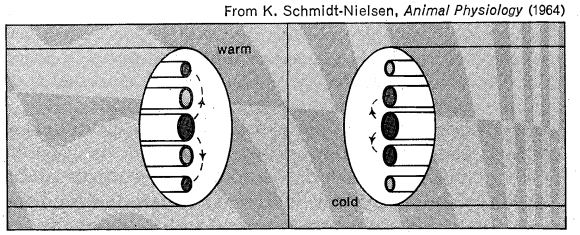


Figure 2: Change in venous blood flow in extremities from near the skin in warm environments to near the arteries in cold environments.

Even though homoiotherms have mechanisms for generating and conserving heat, the energy provided by food, which may be scarce during the winter months, is necessary to maintain a constant body temperature. Some birds migrate to avoid exposure to cold. Some mammals hibernate. A characteristic of hibernation is a decrease in body temperature, which, in turn, means that less energy is required to maintain normal body temperature. By saving energy in this way, however, hibernating animals sacrifice their ability to remain active, since their lower body temperature forces them into a state of lethargy or torpor.

Homoiothermic regulation in a hot environment

When environmental temperatures are above the neutral range (27°–31° C for man), homoiotherms increase heat loss from the body to maintain a constant temperature by one or more of the following methods: conduction, radiation, or convection of heat or evaporation of water. The amount of heat lost by conduction, radiation, or convection is related to the difference between the temperature of the animal and the temperature of the environment. As the environmental temperature increases, this temperature difference decreases, and the amount of heat lost by these methods also decreases; most homoiotherms, therefore, evaporate water to cool themselves when environmental temperatures are high.

Heat is lost when water evaporates because energy, called the heat of vaporization, is needed to change water from a liquid to a gas. About 580 calories of heat are lost whenever one gram of water evaporates from a body. Because terrestrial animals normally contain more water than the environment, water normally evaporates from them by diffusion. Loss of water by diffusion alone, however, is not rapid enough in a hot environment to maintain a constant body temperature. Three mechanisms evolved by homoiotherms to increase the rate of evaporation of water include sweating, panting, and grooming with saliva; all involve an increase in the secretion of watery fluid by specialized glands.

Sweating is used for temperature regulation in some primates (including man), in camels, horses, antelope, sheep, and cattle. Small sweat glands, distributed in large numbers over the body surface, are capable of secreting considerable quantities of fluid. Man sweats profusely in the heat; a 200-pound man in the desert, for example, may lose more than two pounds of water per hour, mostly as a result of sweating.

Panting, which involves an abrupt increase in respiratory rate, is used for temperature regulation by all birds, many carnivores (*e.g.*, dogs, cats), and some antelope and cattle, which also sweat. Inspired air (air breathed into the lungs) is saturated with water before it reaches the lungs, but this water is lost during expiration (removal of air from the lungs). An increase in the rate of respiration, therefore, increases evaporation of water, most of which is from secretory glands and cells in the lining of the respiratory tract. Many animals pant through their noses, thus aiding evaporation of water, since nasal passages, smaller than the mouth, cause more air to come in contact with the moist nasal tissues. In the dog, which pants with its mouth open, air is inspired primarily through the nose and expired through the mouth.

Mammals that cannot pant or sweat groom themselves with saliva to regulate their body temperatures in a hot environment; among this group are most small mammals, such as rodents and bats, and many marsupials. An elephant, which neither sweats nor pants, may spray its body with saliva from its mouth. The use of saliva for cooling has been studied in the laboratory rat, which has three pairs of salivary glands called submaxillary, sublingual, and parotid glands. When the body temperature of the rat increases from 37° to 38.5° C, saliva from the submaxillary glands, secreted rapidly into the mouth, is removed by the rat and placed on the body surface where it evaporates and cools the animal. Saliva is secreted from the parotid glands only if the body temperature of the rat increases to 40° C; secretions from the sublingual glands apparently do not increase with increasing temperatures.

Cooling by saliva

Temperature receptors in the hypothalamus of the rat brain may control the secretion of saliva from different glands as the temperature increases.

The use of water for regulating body temperature in a hot environment is not an efficient mechanism for desert animals because of the scarcity of water. Man, for example, can live in the desert only if he provides himself with large quantities of water to compensate for dehydration, which results from sweating. Some small mammals (for example, the kangaroo rat), however, live in the desert and never drink water. They obtain water from food and have efficient kidneys and other mechanisms adapted for reducing water loss. In addition, they save water by avoiding heat—*i.e.*, their burrows beneath the ground remain cool, and many stay in them during the hot day, coming out only in the cool of the evening. In habit, these small mammals are similar to some poikilotherms, which select favourable microclimates to regulate their body temperatures.

Large desert animals, such as the camel, cannot use burrows to escape from the scorching heat and, contrary to popular myth, do not store water in their bodies. Camels are able to withstand severe dehydration, but their greatest asset in the desert probably is a body of large size, which requires a large amount of heat to cause an increase in temperature. During the hot day, camels store heat by allowing their body temperatures to rise a few degrees; at night, when the environment is cooler, the stored heat is lost, and body temperature falls. Both heat storage and loss occur without the evaporation of large quantities of water.

BIBLIOGRAPHY

General: KNUT SCHMIDT-NIELSEN, *Animal Physiology*, 2nd ed. (1964), an excellent introduction to physiological principles governing adaptations to pressure and temperature; W.S. HOAR, *General and Comparative Physiology* (1966), an advanced text including biological effects of pressure and temperature; ARTHUR GIESE, *Cell Physiology*, 3rd ed. (1970), a discussion of chemical and physical bases for cellular responses to temperature and pressure; D.B. DILL, E.F. ADOLPH, and C.G. WILBER (sect. eds.), *Handbook of Physiology*, sect. 4, *Adaptation to the Environment* (1964), a compendium of review articles including adaptation to aquatic and terrestrial environments.

Pressure: F.H. JOHNSON, HENRY EYRING, and M.J. POLISSOR, *Kinetic Basis of Molecular Biology* (1954), a classic treatment of effects of pressure on chemical reactions and cells; A.M. ZIMMERMAN (ed.), *High Pressure Effects on Cellular Processes* (1970), articles reviewing current knowledge; E.J. DENTON, "The Buoyancy of Fish and Cephalopods," *Prog. Biophys. Biophys. Chem.*, 11:177–234 (1961), a review article summarizing knowledge on the function of cuttlebone and swim bladders; P.F. SCHOLANDER, "Experimental Investigations on the Respiratory Function in Diving Mammals and Birds," *Hvalråd. Skr.*, 22:1–131 (1940), a classic summary of research on physiological adjustments to diving.

Temperature: PETER FREUCHEN and FINN SALOMONSEN, *The Arctic Year* (1958), a delightful account of life in the cold of the Arctic; A.V. WOLF, *Thirst: Physiology of the Urge to Drink and Problems of Water Lack* (1958), classic discussion of temperature regulation, dehydration, and thirst; KNUT SCHMIDT-NIELSEN, *Desert Animals: Physiological Problems of Heat and Water* (1964), an excellent discussion of desert adaptations in vertebrates; G.W. BROWN (ed.), *Desert Biology*, 2 vol. (1968); and J.P. KIRMIZ, *Adaptation to Desert Environment: A Study on the Jerboa, Rat and Man* (1962), two discussions of desert climatology and adaptations of plants and animals to desert environment; T.D. BROCK, "Life at High Temperatures," *Science*, 158:1012–1019 (1967).

(F.R.H.)

Prices, Statistics of

Information about prices is required for a wide variety of purposes in a modern market economy. Perhaps the most important of these is as a tool in the formulation of economic policy at the governmental level. Price statistics are required to measure economic growth; to measure changes in the relative position of particular groups (such as wage earners or recipients of old-age pensions) in the economy; and to determine the degree to which the government will pursue policies that stimulate or dampen economic activity.

Price statistics are important in the business world, too. Sometimes contracts are concluded at a stated price (or wage rate) that is subject to upward revision ("escalation") if prices change. A more widespread and important use of price statistics in business is in connection with planning. In making investment plans for the future, a firm must compare the estimated costs of each investment project with the estimated receipts; this requires information on the prices of materials and wages. Some large firms maintain systematic price records so that cost estimates can be quickly prepared when desired.

HOW PRICES ARE MEASURED

Prices for specific commodities, services, or securities are not hard to gather or to understand. Wholesale price quotations for primary products (coffee, wheat, etc.) and for basic manufactures (iron, steel, aluminum, etc.) are published in trade journals or by government agencies in many countries. Well over 100 such quotations for different products and different countries can be found each month in the United Nations' *Monthly Bulletin of Statistics*. Retail prices for commodities are most usually available on a regular (monthly) basis in government publications. In the United States, for example, the Bureau of Labor Statistics publishes the average urban prices of food and fuels each month. In some countries, including West Germany and Japan, prices are available for a much larger number of nonfood items than in the United States. The Japanese data are particularly detailed and include prices for many commodities in many different parts of the country.

Summary measures of price movements

Indexes. For many purposes the prices of separate commodities are not enough. The reason is that prices for individual commodities (or types of services, or securities) tend to change in different degrees and sometimes even in different directions. What is required is a summary measure of the average amount of the price change for all commodities. This is provided by a price index—that is, a series of numbers arranged so that a comparison between the values for any pair of dates will show the average change in prices from one date to another. Usually a particular date is selected as the base; the index is set equal to 100 at that date, and the values for other dates then indicate at a glance the difference in price level from the base date. The series can easily be shifted to a new base by dividing all numbers in the series by the value for the desired base date and multiplying by 100.

There are a number of practical and theoretical difficulties involved in the making of price indexes. Price indexes are usually based only in part upon quotations that are actually published. The consumers' price index and the wholesale price index typically published in each country by the official price statistics agency usually require the collection of additional price quotations that are used for index number computation but are not made public. The U.S. wholesale price index, for example, is based on unpublished prices for many types of equipment and machinery. One reason, in the case of some machinery items, is that the price information is obtained on a confidential basis. More often, the statistical organization preparing the indexes considers that its sample of price quotations is adequate only for the measurement of changes in the general level of prices.

Data. The central problem of price-data collection is to gather a sample of prices representative of the various price quotations for each of the commodities under study. Sampling is almost always necessary. The larger and the more complex the universe of prices to be covered by the index, the more complex the sampling pattern will have to be. An index of prices paid by consumers in a large and geographically varied country, for example, ideally should be based on a sample representative of price changes in different cities and localities, in different types of outlets (supermarkets, department stores, neighbourhood shops, etc.), and for different commodities. The number of prices chosen to represent each type of city (or metropolitan area), type of outlet, and category of

commodity would ideally be proportionate to its relative importance in the expenditures of the nation. Most price indexes are based on some approximation to such a sampling design.

Once the commodity sample has been chosen, the collection of prices must be planned so that differences between the prices of any two dates will reflect changes in price and price alone. Ideally one would collect the prices of exactly the same items at each date. To this end, commodity prices are sometimes collected in accordance with detailed specifications such as "wheat, no. 2 red winter, bulk, carlots, f.o.b. Chicago, spot market price, average of high and low, per bushel." If all commodities were as standardized as wheat, the making of price indexes would be much simpler than it is. In fact, except for a limited range of goods consisting mainly of primary products, it is very difficult to describe a product completely enough so that different pricing agents can go into stores and price an identical item on the basis of description alone. In view of this difficulty, price collection agencies sometimes rely upon each respondent, usually a business firm, to report prices in successive periods for the same variant of a product (say, men's shoes); the variant chosen by each respondent may be different, but valid data will be obtained as long as each provides prices for the same variant he originally chose. Because a product may vary in quality from one observation to another, even though it retains the same general specification, the usual procedure is to avoid the computation of average observed prices for each commodity for each date. Instead, each price received from each source is converted to a percentage of the corresponding price reported for the previous period from the same source. These percentages are called "price relatives."

Weighting. The next step is to combine the price relatives in such a way that the movement of the whole group of prices from one period to another is accurately described. Usually, one begins by averaging the price relatives for the same specification (*e.g.*, men's high work shoes, elk upper, Goodyear welt, size range 6 to 11) from different reporters. Sometimes separate averages for each commodity are calculated for each city, and the city averages are combined.

A more difficult problem arises in combining the price relatives for different commodities. They must be given different weights, of course, because not all the commodities for which the prices or price relatives have been obtained are of equal importance. The price of wheat, for example, should be given more weight in an index of wholesale prices than the price of pepper. The difficulty is that the relative importance of commodities changes over time. Some commodities even drop out of use, while new ones appear, and often an item changes so much in composition and design that it is doubtful whether it can properly be considered the same commodity. Under these conditions, the pattern of weights selected can be accurate in only one of the periods for which the index numbers have been calculated. The greater the lapse of time between that period and other periods in the index, the less meaningful the price comparisons become. Price indexes thus can give relatively accurate measures of price change only for periods close together in time.

Sources of bias

Another problem of price index number construction that cannot be completely resolved is the problem of quality change. In a dynamic world, the qualities of goods are continually changing (permanent-press clothing, colour televisions, banded tires, self-cleaning ovens, and so forth) to such a degree that it is doubtful whether anyone living in an industrialized economy buys many products that are identical in physical and technical characteristics to those purchased by his grandfather. There is no fully satisfactory way to handle quality changes. One way would be to make price comparisons between two periods solely in terms of goods that are identical in both periods. If one systematically deletes goods that change in quality, the price index will tend to be biased upward if quality is improving on the average and downward if it is deteriorating on the average. A better approach is to attempt to measure the extent to which an observed change in the quoted price represents a change in quality. It is possible, for example, to obtain from manufacturers estimates of the increase or decrease in cost of production entailed in the main changes in automobiles from one model year to the next. The amount added or subtracted from the cost by the changes can then be regarded as a measure of the quality change; any change in the quoted price not accounted for in this way is taken as solely a change in price. The disadvantage of this method is that it cannot take account of improvements that are not associated with an increase in costs.

Whether or not a failure to make sufficient allowance for improvements in the quality of goods causes most price indexes to be biased upward is a matter of dispute. An expert committee appointed to review the price statistics of the U.S. government (the Stigler Committee) declared in 1961 that most economists felt that there were systematic upward biases in the U.S. price indexes on this account. Since the U.S. indexes are usually thought to be relatively good, this view would seem to apply by extension to those of most other countries. The official position of the U.S. Bureau of Labor Statistics has been that errors owing to quality changes have probably tended to offset each other, at least in its index of consumer prices.

Another possible source of error in price indexes is that they may be based on list prices rather than actual transactions prices. List prices probably are changed less frequently than the actual prices at which goods are sold; they may represent only an initial base of negotiation, a sellers' asking prices rather than an actual price. When prices are collected largely by mail, as is generally the case with indexes of wholesale prices, the respondents may give only their list prices, with the result that the indexes will show greater short-run stability than is warranted. One study has shown that actual prices paid by the purchasing departments of government agencies were lower and were characterized by more frequent and wider fluctuations than were the prices for the same products reported for the price index.

CONSUMER PRICE INDEXES

Although consumer price indexes are familiar to everyone, they date only from the final decades of the 19th century. Efforts have been made to reconstruct the movement of consumer prices in earlier periods, but such work has often had to be based on the assumption that changes in the retail price of a given commodity could be inferred from changes in its wholesale price. (Mention may be made, however, of a quite exceptional series on London bread prices dating back to 1545, by B.R. Mitchell and P. Deane.)

What the indexes measure

The indexes that began to be produced in the last decades of the 19th and the first decades of the 20th century were called "cost of living" indexes, reflecting their origin in a concern about the welfare of workers and a consequent interest in the purchasing power of wages. Early indexes tended to concentrate on foods and other necessities. World War I brought a further stimulus to the establishment of official cost of living indexes because of the need for measures of consumer price changes in connection with wartime wage controls. Later, the term consumer price index was generally adopted in order to make it clearer that the indexes were attempting to measure the cost of living at a given standard. The term consumer prices is usually preferred to retail prices since the former more clearly covers whatever it is that the consumer spends his money on, including rents, medical services, taxes, and so forth, whereas the latter has the more limited connotation of measuring only prices paid at retail shops. It may also happen, as in Germany, that consumers purchase significant amounts of some types of goods at wholesale or factory outlets rather than at retail shops. The United Kingdom and a number of other countries nevertheless describe their indexes as "retail price" indexes, and even the term cost of living is still used in a few countries.

Current consumer or retail price indexes are, then, usually defined as being designed to show the changes over time in the prices of the commodities and services pur-

chased by a specified group of consumers. Such indexes are available for more than 100 countries (see the United Nations' *Monthly Bulletin of Statistics*) and are usually prepared by the country's ministry of labour or central statistical office. The indexes of the various countries differ widely in coverage and methods, but there are some general characteristics that can be described.

Limitations in coverage

The population group covered almost always includes families of wage and salary earners, but the indexes vary in the extent to which they include single individuals and other occupational groups such as the self-employed or professionals. The United States index, for example, includes only wage and salary earners, whether living alone or with families, whereas the United Kingdom index covers all households except those headed by a person with income in excess of a stated amount (£40 in 1965) and those with at least three-quarters of their total income derived from pensions. Sometimes special indexes are prepared for particular groups in the population, such as the indexes for pensioners in the U.K.

It is quite common for the indexes to refer only to urban areas, and some are restricted to a limited number of cities (*e.g.*, six state capitals in Australia) or even to the capital city alone (*e.g.*, Mexico City). Some indexes are, of course, broader in coverage; the Japanese index, for example, includes all households, urban or rural, except one-person households and those headed by farmers or fishermen. These limitations in the coverage of the indexes are often overlooked. For broad analytical uses such as measuring changes in the national welfare, a comprehensive index is needed that covers the whole population, including persons living in rural areas, single persons, and high-income families in urban areas.

In most countries, the consumer price indexes, usually published monthly, include indexes for major categories of expenditures such as food, clothing, rent, and so forth. Sometimes they include more detailed categories (such as dairy products) and even individual items (such as milk). This depends partly on the number of products for which prices are gathered. In countries that can afford more extensive price collection and that need it because of the diversity of products purchased by their consumers, the number of items covered is usually between 250 and 450; but in smaller countries or poorer ones, the number is often between 100 and 150, and in a few countries even less than 50.

The formula. The most common formula used in calculating consumer price indexes is a weighted arithmetic mean of price relatives. The price relatives described earlier are weighted according to the amounts consumers spend on each product; the resulting figures are summed for all commodities and divided by the sum of the base year expenditures for the same collection of commodities.

The expenditure weights reflect the relative importance of each item in the expenditures of the population group covered by the index. This information is usually obtained from a survey of a sample of the covered families, in which data are gathered about the expenditures upon each commodity and service. The weights need to be revised from time to time as new products appear and as spending habits change. In the U.S., the Stigler Committee recommended a comprehensive revision of the weights at least once a decade and revisions for the more volatile categories at least every five years. The main obstacle to frequent weight revision is the cost of the family expenditure surveys, although modern sampling methods have made it possible to obtain reliable results with smaller numbers of observations than were used in the early days of consumer price indexes. In some countries, however, family expenditure surveys are conducted continuously, as they are in the annual surveys in the U.K. or the monthly surveys based on overlapping panels of households in Japan.

Conceptual difficulties. A fundamental difficulty in existing consumer or retail price indexes is that they are defined in terms that are not really appropriate for the uses to which they are put. The official compiling agencies describe them as measures of the change in price of a fixed basket of goods and services. This seems clear enough, but difficulties arise as soon as the index is put to use. One may wish, for example, to ascertain whether an urban family with an income of $10,000 in 1969 was better off than a similar family with an income of $8,000 in 1959. One may find that the consumer price index rose by 27 percent in the ten-year period. This more than offset the 25 percent increase in money income, and the family is therefore slightly worse off. Although this is a fair illustration of the way the indexes are generally employed, the underlying reasoning is defective. In the first place, the implicit assumption is that the family bought exactly the same goods in 1969 as in 1959. Even if all goods in the market were identical at the two dates (no new products and no quality change), the family would not necessarily purchase the same items; it could logically be expected to substitute goods that had smaller-than-average price increases for goods that had larger-than-average price increases. As a result of these shifts, the family might actually prefer the new bundle of goods at the new prices to the old bundle at the old prices. Thus, the cost of acquiring a bundle of goods that yielded the same satisfactions would have risen not by 27 percent but by some smaller percentage. An index that takes into account the subjective preferences of consumers is called a constant-utility index, since it measures not the change in price of a constant bundle of goods but the change over time in the costs of purchasing bundles of goods that yield a constant level of utility or satisfaction. Even this formulation is not completely rigorous because the utilities or satisfaction obtained by a given individual from a given bundle of goods will vary over time as his tastes change. All that an index can do in strict theory is to measure differences in the money incomes required at different price structures to make a given individual with given preferences indifferent as to the price structures.

Conceptually, there is little doubt that the constant-utility index is superior for virtually all of the analytical purposes for which indexes are used, but there are great difficulties in constructing one that would command wide support and confidence. Some authorities believe that the present indexes could be improved if the physical units were selected so as to represent the characteristics or utilities really sought by buyers. They hold, for example, that it would be more appropriate to compare the costs of an appendectomy or of an obstetrical delivery than the costs of an office visit to a doctor or eight hours of nursing care. The Stigler Committee felt that some improvement could be made if new commodities were introduced into the indexes when they first begin to be purchased by the population; the indexes would then reflect the gains to consumers that come with reductions in prices after a product achieves wide acceptance.

The Stigler Committee was concerned lest the important economic interests that make use of the consumer price index institutionalize it and hamper improvement of it. The Committee feared, for example, that neither labour nor management would welcome changes in the methods of preparing an index, the movements of which automatically result in wage adjustments for millions of employees under labour–management agreements.

WHOLESALE PRICE INDEXES

Indexes for commodities

Wholesale price indexes measure the changes in commodity prices at a selected stage or stages before goods reach the retail level; the prices may be those charged by manufacturers to wholesalers or by wholesalers to retailers or by some combination of these and other distributors. In the U.S., the index measures the price movements of all commodities flowing into primary markets of the U.S.—whether domestically produced or imported. Primary markets are those in which a good in a given stage of fabrication is first sold in substantial quantities. Since primary markets include goods of all degrees of fabrication, the same commodity is often priced at several stages of processing. Cotton, for example, is priced in the form of raw cotton, cotton yarn, cotton gray goods, cotton piece goods, and cotton clothing.

One of the earliest wholesale price indexes was produced for Great Britain in 1886, covering the period after 1846.

The official wholesale price index in the United Kingdom, produced by the Board of Trade, goes back to 1871. In the United States, the first major effort to summarize wholesale price changes through index numbers was published in a report of the U.S. Senate in 1893. The present wholesale price index of the United States, maintained by the Bureau of Labor Statistics, has been computed for the period since 1890. In both the U.K. and the U.S., economic historians have attempted to reconstruct wholesale price indexes for the 19th century that are superior to the early efforts.

The number and character of the commodities included in wholesale price indexes vary widely from country to country. In large industrial countries like the U.K., the U.S., and Germany, the number of commodities included numbers in the thousands; but for most countries it is much smaller, often only 100 or 200. The smaller numbers of products will serve well enough if only a general all-commodities index (or a few subindexes at most) is wanted. Greater numbers are required when many subindexes are desired. The U.S., for example, publishes indexes for commodities classified according to stage of processing (crude materials, intermediate materials, and finished goods), according to the durability or nondurability of the products, and according to the economic sector for which goods are intended (consumers, producers, etc.). The commodities are also grouped into 15 categories and nearly 100 subgroups (fresh fruits, grains, etc.) and a large number of product classes (apples, bananas, barley, corn, etc.), for each of which monthly price indexes are published. In addition, there are a number of indexes for special commodity groups such as various categories of pharmaceutical preparations. The number of commodities included in the U.S. index has expanded from 250 when the index was started in 1902 to about 2,400 in 1970. The new commodities have tended to be more highly fabricated and to have more stable prices, and they have therefore dampened the fluctuations in the index. One reason for the inclusion of more commodities was a gradual shift in the conception of the function of the index. Originally it was regarded as a measure of movements in the general price level, but as other indexes became available, such as the consumer price index and the implicit national income deflators (see below), less reliance was placed on the wholesale price index for this purpose. At the same time, there was a growing demand for subindexes pertaining to particular classes of products for various business and analytical purposes.

Countries in which industrial production is not highly variegated usually have smaller numbers of product classifications; these serve to distinguish between price movements of domestic goods and price movements of imports and between those of food or agricultural products and those of industrial products. Raw materials and standardized products in early processing stages that are easy to price tend to be well represented in the wholesale price indexes of all countries; whereas more complex types of producers' goods, such as heavy electrical equipment, tend to be underrepresented or omitted even in the indexes of the advanced industrial countries. This is a source of upward bias in the general wholesale indexes since there is reason to believe that technological change has been particularly important in bringing about improvements in complex goods.

Many of the technical problems of wholesale price indexes are similar to those discussed in connection with consumer price indexes, although it is probably fair to say that the wholesale indexes have been subject to much less criticism and have been produced with fewer resources than consumer price indexes. Prices are usually gathered from business firms by mail, less frequently from trade journals and trade associations, and also from government purchasing agents. Weights are generally based on relative sales volume. Data from censuses of production (manufacturing, mining, agriculture, etc.) are used for weights when they are available. The formula used for computing the indexes is most frequently the base-weighted arithmetic average of price relatives already explained in connection with consumer price indexes.

IMPLICIT PRICE DEFLATORS

Comprehensive price indexes

Measures of price change that are more comprehensive than either the consumer or the wholesale price index have evolved as a result of the recent development of national income statistics. After statisticians had compiled estimates of gross national product in terms of current prices (*i.e.*, the prices prevailing in each year), they turned their attention to estimating them in terms of constant prices (*i.e.*, estimating the gross national product, for example, for a whole succession of years in terms of the prices of a single year). This was necessary if the figures were to distinguish between real changes in physical output and changes in prices. This is done by expressing each component of gross national product in terms of the prices of the selected base year, using whatever available price series seems most suitable for that component. Most of the price series are apt to be elements of the official wholesale (*e.g.*, agricultural machinery) and retail price indexes (*e.g.*, food), but a wide range of other price information is also employed. When these "deflated" components of gross national product are added up, the result is an estimate of gross national product in terms of base period prices. In most countries, national income and product statistics are published both in current prices and in constant (*i.e.*, base-period) prices. There is, of course, a comprehensive price index implicit in the comparison of the current and constant-value series. For example, from the fact that the 1966 Japanese gross national product was 35,092,000,000,000 yen in current (1966) prices and 26,822,000,000,000 yen in 1960 prices, an implicit price index of 130.8 can be computed (35,092 divided by 26,822 and multiplied by 100) for 1966 on a 1960 base. Implicit deflators are usually available for various components of the gross national product both in terms of the output of particular industries and in terms of the final products purchased by consumers, investors, and government. Implicit price indexes for many countries are published in the United Nations' *Yearbook of National Account Statistics.*

Implicit price deflators are generally superior as a measure of the overall behaviour of prices because, in addition to being more comprehensive than the major indexes previously discussed, they refer to a universe that can be more readily defined than that of the wholesale price index. They are usually available, however, only on a quarterly or annual basis, and therefore the other indexes are still used as month-to-month indicators of price behaviour.

The implicit deflators obviously contain within themselves whatever defects characterize the underlying consumer, wholesale, and other price series upon which they are based. If, as suggested above, there has been an improvement in the quality of many goods that escapes measurement so that the price indexes have tended to overestimate the extent of price increases, then the use of implicit deflators leads to a corresponding underestimation of the increase in real national product and real national income.

INDEXES FOR SPECIAL PURPOSES

Export and import price indexes

There are various ways of measuring price changes in international trade. A few countries of the world, including Germany and Japan, have export and import price indexes. The U.S. and most other countries have unit-value indexes for their exports and imports. A unit-value index differs from a price index in that it measures the changes in the average value per physical unit regardless of whether the change is due to a change in price or to a change in size, quality, or other circumstance.

Unit-value indexes are computed from foreign trade data on values and quantities. The unit value for each component series is derived by dividing quantities into values; the changes in the individual unit-value series are then collected into an overall index through the use of export or import values as weights. Unit-value indexes are usually constructed according to Fisher's ideal index number formula: that is, each index represents the square root of the product of base-year weighted and given-year weighted indexes. This formula is used rather than the

one discussed above because the composition of exports and imports, and therefore the appropriate weights, changes much more rapidly than does the market basket of goods in the consumer or wholesale price indexes. The Fisher formula averages the results of using current-year weights and base-year weights. Usually the previous year is taken as the base year, and the percentage changes calculated in this way are linked to form a continuous series.

Unit-value series are not reliable approximations of price changes because even for relatively simple types of products, like steel pipe, changes in the mix of export or import products (as from narrow to wide pipe, thick to thin, etc.) can change the unit value even though no single export or import price has changed. Another handicap is that quantity data are simply not collected and published for many types of manufactures because the number of units would be meaningless when the size, quality, design, and other characteristics of the product vary widely from one unit to another.

Despite these objections, unit-value indexes continue to be produced. They are relatively easy to compute; true price indexes require much more money and personnel. Unit-value indexes are often used for analytical purposes as if they were price indexes, although wholesale price data, reweighted to take account of the relative importance of different products, may be better indicators of price changes in international trade. Wholesale prices are not fully satisfactory either because domestic prices (which are what the wholesale price indexes mainly measure) do not always move parallel to export prices.

The indexes discussed in the preceding sections are the most important and most widely used. There are, however, a wide variety of others. Most industrial countries have indexes of hourly earnings and wage rates; the most common types are indexes of average hourly, weekly, or annual earnings in manufacturing.

Construction-cost indexes are found in many countries. Virtually all construction-cost indexes are put together by pricing the inputs of labour and materials or a combination of labour and materials and units of construction work (excavation, brickwork, concrete pouring, etc.). The disadvantage of this approach is that it does not take account of changes in the efficiency with which the inputs are combined to accomplish the construction task.

Indexes of security prices are available wherever there are active securities markets such as in New York, London, Paris, Tokyo, Brussels, Frankfurt, and Zürich.

The general user of price indexes can usually find what he needs in most countries by looking in the monthly bulletins of the central statistical office; the ministry of labour, commerce, or economics; or the central bank. Daily newspapers and trade journals in each country also contain price quotations for many commodities, services, and securities. The United Nations and its specialized agencies, particularly the Food and Agricultural Organization and the International Labour Organization, include compilations of prices from many countries in monthly bulletins and in yearbooks of statistics.

PLACE-TO-PLACE PRICE INDEXES

City-to-city and country-to-country comparisons

Comparisons of the levels of prices in different places may sometimes be required. One of the complaints of the New York City postal workers in their 1970 strike, for example, was that the practice of paying uniform nationwide salaries was unfair to them because a given number of dollars purchased substantially fewer goods in New York City than in certain other parts of the United States. The problems and techniques of preparing place-to-place price indexes are similar to those of time-to-time indexes. The price universe has to be defined, a sample chosen, careful pricing made of equivalent qualities of goods in different places, and weights selected to collect the (place-to-place) price relatives for different goods into an overall index. In the actual operation, however, place-to-place indexes are more difficult to prepare because it is harder to assure that identical goods are being priced in every locality.

The U.S. Bureau of Labor Statistics compared the 1966 costs of a city worker's family budget at a moderate standard of living in 39 metropolitan areas. The standard budget was prepared for a family consisting of an employed husband, age 38, a wife not employed outside the home, a girl, age 8, and a boy, 13. (A standard budget was also prepared for a retired couple.) The budget quantities were derived from scientific or technical judgments about the requirements for physical health and well-being (*e.g.*, for food and housing) and from analytical studies of choices actually made by consumers (clothing and recreation)—all adjusted to climatic and regional differences. With the U.S. urban average marked as 100, the indexes ranged from 94 in Dayton to 110 in Honolulu, San Francisco, and Los Angeles.

International comparisons. Price comparisons between countries are needed because the exchange rates between currencies do not necessarily reflect their relative purchasing powers. Efforts to calculate "purchasing-power parities" between currencies have a long history, but because of the extensive work required to compare prices item by item (and to be certain that prices for equivalent items are being compared), this area remains a major gap in the world statistical system.

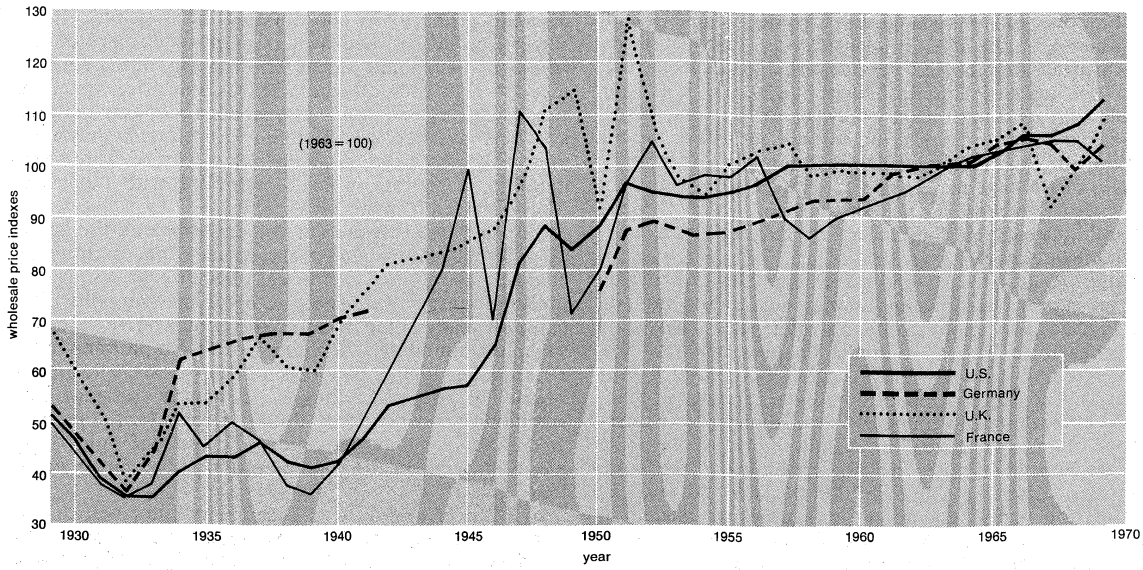
Several of the purchasing-power comparisons were made in order to compare the standard of living of workers in different countries. The early cost of living study carried out by the U.K. Board of Trade in 1908 compared wages, food prices, and rents in the U.K. with those in Belgium, France, Germany, and the U.S. In 1954 the European Coal and Steel Community compared wages received and prices paid by steelworkers and coal miners in its six member countries.

Other comparisons have arisen out of the need of employers such as the United Nations, foreign ministries, and international firms to pay the same real salaries to their employees of comparable grades regardless of country of assignment. This requires price comparisons so that money salaries can be adjusted to compensate for differences in the purchasing power of different currencies. A readily available comparison designed to serve this purpose is published each June and December in the United Nations' *Monthly Bulletin of Statistics*. The indexes compare the cost of purchasing a basket of goods and services bought by a typical international civil servant in about 100 foreign cities with the cost in New York City, where the UN headquarters is. The indexes are based on prices for 150–200 items; housing is excluded from some of the published figures because the rents paid by international civil servants are sometimes subsidized. Weights for the indexes are derived from the expenditure patterns of the covered personnel.

Comparisons of GNP. Price indexes may also be used to compare the gross national products (GNP) of different countries. The UN and other sources currently publish such comparisons, derived mainly or entirely from the use of exchange rates. These are known to be inaccurate because of the differences between exchange rates and purchasing-power parities already alluded to. The Statistical Office of the UN has set out to develop actual purchasing-power comparisons of GNP and its major components (consumption, capital formation, and government expenditure) among different countries.

Difficulties in making international comparisons of GNP

The kind of work that is involved in such comparisons may be seen in a study carried out in the early 1950s at the Organization for European Economic Cooperation. Purchasing power equivalents and comparisons of real product per capita were prepared for four European countries (France, Germany, Italy, and the United Kingdom) in relation to the U.S. It was found that, for the bundle of goods included in the U.S. GNP, the purchasing powers of the European currencies in 1950 were from 8 to 24 percent higher than the exchange rates would indicate; for the bundle of goods in the European GNP, they were from 57 to 91 percent higher. The purchasing power of the British pound, for example, was not 2.8 times that of a dollar, as the 1950 exchange rate of $2.80 = £1 would have suggested, but 3.47 times for the U.S. bundle of goods (24 percent higher than 2.8) and 4.59 times for the European bundle of goods (64 percent higher than 2.8).



Fluctuation of wholesale price indexes in France, Germany, the United Kingdom, and the United States during a 40-year time period.

These differences between the purchasing-power parities and the exchange rates meant that to translate the European GNP's into dollar terms by using exchange rates underestimated them relative to the U.S. GNP. The per capita GNP of the U.K., for example, when converted to dollars at the official exchange rate, was 37 percent of the U.S. per capita GNP, but the percentage rose to 53 when both U.K. and U.S. products were valued at European prices and to 63 when valued at U.S. prices. Differences of this magnitude can be important when member countries in international organizations have to agree upon sharing financial burdens, as they do in aid programs for poor countries or in military alliances such as NATO.

Without purchasing-power comparisons, no one can be sure what the real differences are in the levels of income (or GNP) attained by various countries. The same difficulty applies to important subaggregates of GNP. One of the major variables affecting economic growth, for example, is the amount of fixed capital formation (*i.e.*, investment in new construction and equipment). In order to know how much real investment there has been in each country, purchasing-power comparisons are necessary. Not only will the conversion of capital expenditures into dollars or some other common currency by means of exchange rates be misleading but even the ratio of capital expenditures to gross national product, in terms of each country's own currency, may not be comparable from one country to another; the comparability of the ratios requires that the relative purchasing powers of each pair of currencies be the same for capital goods as for all other goods. In fact, capital goods are relatively expensive in some countries and relatively cheap in others. In 1960, for example, it took about 165 Argentine pesos to buy one U.S. dollar's worth of capital goods (the currency rate was 82.7 pesos per dollar). In terms of pesos, Argentina devoted about 22 percent of its gross product to investment; but when all outputs were revalued to reflect the average of U.S. and Latin American prices, Argentine investment was equivalent to about 11 percent of its gross output. Thus, by international standards Argentina's real new capital formation relative to other components of real product was much more modest than the Argentine figures taken alone would suggest.

THE MOVEMENT OF PRICES OVER TIME

Price movements in different countries are linked through international trade, and, if due allowance is made for movements in exchange rates, the broad changes in prices will be found to be very similar for countries maintaining relatively open economic relations with the rest of the world. This can be seen in the Figure, which shows the behaviour of prices in some leading countries over a period of nearly half a century.

Long-term changes. The course of prices in the 19th century was characterized by relatively moderate changes except for the rather violent fluctuations associated with war and postwar periods. Prices rose during wartime, when governments made large purchases of war materials, and fell afterward. The end of World War II, however, was not followed by a period of declining prices. This may be due in part to the fact that it was as much a period of continued war as one of peace; the Korean War in the early 1950s and the Vietnam War in the late 1960s were clearly associated with rising prices, and the Cold War served to buoy up prices in between these periods of actual fighting.

The historical record of prices has been cited as evidence for the existence of long swings in economic activity. The first of the upswings, 1789–1814, has been associated with the effects of the Napoleonic wars and the Industrial Revolution; the second, 1849–65, with the discovery of gold, the building of railroads, and the American Civil War; the third, 1896–1920, with "electrification and motorization" and World War I; and the fourth, from 1932 on, with World War II, its unpeaceful aftermath, and the greatly expanded sphere of governmental activity in the economy. Downswings have generally been explained in terms of the eventual loss of impetus in the technological innovations that had played the major role in the preceding upswings.

Cyclical changes. Another pattern of price change is that seen in the familiar business cycle. The fluctuations observed in the business cycle are, of course, changes in aggregate values produced, traded, or consumed; these changes in values are the result of changes in both prices and quantities. Thus, even as a symptom of what is happening in the business cycle, price changes must be considered in connection with changes in quantity. Studies have indicated that prices and quantities fluctuated about equally over the course of a cycle. There has, however, been a significant difference in timing: on the upswing, prices rose only 80 percent as much as quantities did; but on the downswing, they have declined more than quantities. Also, quantities have tended to rise in the final stage of contraction, while prices have still been falling. Not all prices have followed this pattern, nor has every cycle corresponded to the generalized picture. For farm products, foods, consumer goods, and nondurable goods, prices varied more than quantities during the course of the cycle, but the combined variation of both price and quantity (*i.e.*, value) was characteristically low. On the other hand, for durable goods, capital equipment, and nonfarm products, prices were relatively stable, and cyclical influences were reflected in relatively large changes in quantities.

Some notion of the nature of the changes in price rela-

Changes in relative prices

tionships over time may be gleaned from a comparison of components of the U.S. wholesale price index at three cyclical peaks roughly 20 years apart. Farm prices rose relative to industrial prices between 1929 and 1948 and declined in absolute terms between 1948 and 1969, whereas the rate of price increase in industrial prices accelerated. Among industrial prices, lumber rose sharply in price in the first period and continued with a greater than average increase in the second period. Rubber prices lagged in the early period, and rose relatively to the average in the later period.

Changes such as these—and the even greater diversity of changes in price relationships among individual commodities—reflect two major groups of influences. One set of influences is to be found in the changes affecting the supplies of and demands for the various goods and groups of goods. The other is represented by differences in market structure that cause some goods to respond more readily than others to a general movement of prices. The great increase in lumber prices may be attributed to the large quantities of lumber demanded for the war and postwar construction booms and to the difficulties in expanding the quantity supplied. The prices of tires and tubes and of other rubber products did not rise very much in the two decades following 1929 because these years saw important cost-reducing technological changes in rubber manufacturing, particularly the development of synthetic rubber.

The growth of productivity with the consequent rise in consumer incomes is also likely to have an impact on relative prices. Higher incomes, including higher wages, mean relatively greater price increases for things that utilize labour intensively relative to things that require larger amounts of materials and equipment for their production. Prices of services, for example, have increased considerably more than commodity prices in recent decades.

Changes in price relationships are attributable also to the varying responsiveness of different prices to a general shift in the price level. In the case of farm products, government policy has a great deal to do with price movements. Changes in industrial prices, on the other hand, may reflect the structure of particular industries. Certain economists have held that, in industries with a small number of firms, businessmen have the power to set their own prices. These "administered" prices are not changed very often; responses to changes in demand take the form of output adjustments. The existence of administered prices has been widely accepted by economists, although there are some who argue that their rigidity has been exaggerated.

BIBLIOGRAPHY. A classic study in index number construction is W.C. MITCHELL, "The Making and Using of Index Numbers," *Bull. U.S. Bur. Labor Statist. 284*, pt. 1 (1921; reprinted as *Bull. 656*, 1938). A more recent review of the problems of price statistics, particularly with reference to the U.S., may be found in PRICE STATISTICS REVIEW COMMITTEE, NATIONAL BUREAU OF ECONOMIC RESEARCH, *The Price Statistics of the Federal Government* (1961). Analytical studies of price behaviour include: F.C. MILLS, *The Behaviour of Prices* (1927), and *Price-Quality Interactions in Business Cycles* (1946); A.L. BOWLEY, *Wages and Income in the United Kingdom since 1860* (1937); A.C. NEAL, *Industrial Concentration and Price Inflexibility* (1942); and G.J. STIGLER and J.K. KINDAHL, *The Behaviour of Industrial Prices* (1970). Historical price series may be found in U.S. DEPARTMENT OF COMMERCE, CENSUS BUREAU, *Historical Statistics of the United States, Colonial Times to 1957* (1960); and B.R. MITCHELL and P. DEANE, *Abstract of British Historical Statistics* (1962). Current sources of price statistics include the U.S. DEPARTMENT OF LABOR, *Monthly Labor Review*; the U.S. DEPARTMENT OF COMMERCE, *Survey of Current Business* (monthly); and the UNITED NATIONS, STATISTICAL OFFICE, *Monthly Bulletin*. For international price comparisons, see M. GILBERT and I.B. KRAVIS, *An International Comparison of National Products and the Purchasing Power of Currencies* (1954); and I.B. KRAVIS and R.E. LIPSEY, *Price Competitiveness in World Trade* (1971).

(I.B.K.)

Price System

The price system, as its exists in western Europe and the Americas, is a means of organizing economic activity. It does this primarily by coordinating the decisions of consumers, producers, and owners of productive resources. Millions of economic agents who have no direct communication with each other are led by the price system to supply each other's wants. In a modern economy the price system enables a consumer to buy a product he has never previously purchased, produced by a firm of whose existence he is unaware, which is operating with funds partially obtained from his own savings.

Prices are an expression of the consensus on the values of different things, and every society that permits exchanges among men has prices. Because prices are expressed in terms of a widely acceptable commodity, they permit a ready comparison of the comparative values of various commodities—if shoes are $15 per pair and bread 30 cents per loaf, a pair of shoes is worth 50 loaves of bread. The price of anything is its value in exchange for a commodity of wide acceptability (a money): the price of an automobile may be some 50 ounces of gold or 25 pieces of paper bearing the picture of an eminent statesman.

A system of prices exists because individual prices are related to each other. If, for example, copper rods cost 40 cents a pound and the process of drawing a rod into wire costs 25 cents a pound, then, if the price of wire exceeds 65 cents, it will be profitable to produce wire; and if the price of wire falls below 65 cents, it will be ruinous to produce wire. Competition, therefore, will hold the price of wire about 25 cents per pound above that of rods. A variety of such economic forces ties the entire structure of prices together.

The system of prices can be arranged to reward or penalize any kind of activity. Society discourages the production of electric shoe-string-tying machines by the simple expedient of making such a machine's attainable selling price less than the prices of the resources necessary to produce it. Society stimulates men of large athletic promise to learn golf (rather than polo or cricket) by the immense prizes (= prices) that are given to tournament winners. The air in many cities is dirty because no one is charged a price for dirtying it and no one can pay a price for having it cleaned.

THE BASIC ECONOMIC TASKS

Three functions of a price system

Every economic system has three functions. In a decentralized (usually private enterprise) economic system, the price mechanism is the instrument by which these functions are performed.

One function is to determine what is to be produced and in what quantity. Even a primitive economy must choose between food and shelter, weapons and tools, priests and hunters. In a modern economy the potential variety of goods and services that may be produced is immense. Consider simply the 10,000 new book titles published each year or the hundreds of colours of paint or the thousands of styles of clothing—each of these actual collections being much smaller than modern technology permits.

A price system weighs the desires of consumers in terms of the prices they are willing to pay for various quantities of each commodity or service. The payment of $1,000 for the services of a skilled surgeon (a price much influenced by the number of surgeons) reflects the great importance of his services to the buyer-patient, whereas the offer of 75 cents for a month's use of an additional telephone outlet reflects the minor convenience it provides. Of course the offers of consumers are influenced by their wealth as well as their desires, but for any one consumer relative desires are proportional to price offers.

Universal laws are most uncommon in social life. Economists nonetheless place immense confidence in the proposition that the consumer will buy less of any commodity when its price rises. This *law of demand* is by no means a necessary fact of life; rather it is an empirical rule to which there are no known, reliable exceptions. Bread, caviar, education, narcotics—a man will buy more of each when its price falls. These demand prices are the guides to producers that in effect tell them which commodities to produce and in what quantities.

The second function an economy must perform is to decide how the desired goods are to be produced. There is more than one way not only to skin a cat but also to grow wheat, train lawyers, refine petroleum, and transport baggage. The efficient production of goods requires that certain obvious rules be followed: no resource should be used in producing one thing when it could be producing something more valuable elsewhere; and each product should be made with the smallest possible amount of resources.

A functioning price system steers resources into their most important use by appealing to the desires of their owners for large incomes. The man capable of being a surgeon is drawn to this occupation from, say, that of a high school teacher by the promise of annual earnings (= price of labour) much more than those of the high school teacher. Capital is drawn from a faltering trade to a booming new industry in which it receives a higher return.

This same price system seeks to achieve production efficiency through the sanction of competition. If one firm can produce shoes with fewer resources than its rivals, it will make larger profits; so it is stirred to discover more efficient combinations of inputs and location of plant, to devise wage systems to stimulate workers, to use computers to reduce inventories, and so forth without end.

The third function of an economy is to determine who gets the product. Family A gets $5,000 worth of goods this year, family B five times as much—how is the division to be decided? The incomes of individuals are determined by the quantities of resources (labour skills, capital in all its forms) they own and the prices they receive for the use of these resources. Workers are incited by the price system to acquire new skills and to exercise them diligently, and families are encouraged to savings (capital accumulation) by the payment of interest or dividends. The inheritance of both personal ability and wealth also enter into the distribution of income.

If the price system is working reasonably well (some of the common failures will be noted later), it performs all of these economic functions with remarkable subtlety and precision. Society desires not only the correct amount of wheat but also that it be consumed more or less evenly over the crop year, with a surplus to carry over in case of a partial failure of the next year's crop. The price system provides a seasonal price pattern that encourages the holding of inventories rather than early splurging and richly rewards speculators who correctly anticipate a crop failure and hold grain that will alleviate it. In the same way, the desires of every sizable group of consumers (or resource owners) are registered through the price system; entrepreneurs are incited by price offers to provide opera and musical comedy, kosher food, and Persian delicacies. One might almost say that the price system is devoted to minority rule, since the only pressure toward uniformity is in the possibility of lowering costs of production by standardizing goods.

High prices in a properly functioning price system thus serve as incentives to produce more and consume less, and lower prices serve as corresponding deterrents. In addition the price system is a method of communicating information. Herbert Spencer once stated, rather ponderously, that only by constant iteration can alien truths be impressed upon reluctant minds: the price system, with its capacity for infinite repetition, is well suited to this sometimes unpleasant task. A higher price of steel scrap, for example, tells thousands of owners and collectors of scrap that more scrap is wanted and that more exhaustive search for abandoned rails, boilers, radiators, and machines is worth undertaking. A higher price of gasoline tells thousands of automobile drivers that gasoline should be used more sparingly, and the message is repeated each time each driver purchases more gasoline.

HOW THE PRICE SYSTEM WORKS

The complexity and variety of tasks performed by the price system will be illuminated by an examination of three specific economic problems.

The choice of occupation. Individuals must be distributed among occupations in such a way as to serve two basic purposes. First, the labourer must be placed where he is most productive—making certain that Enrico Fermi becomes a physicist rather than a chef and that there are not too many plumbers and too few electricians. Second, the individual worker should be given an occupation that is congenial to him: since he will spend a large part of his life at work, it will be a better life if he can choose the type he prefers.

The price of labour

The price of labour is the instrument by which workers are distributed among occupations: wages in rapidly growing occupations and rapidly growing parts of the nation are higher than in corresponding employment in declining occupations and areas. The choice of occupation involves, however, much more than simply a comparison of wage rates. The following are a few of the complications: (1) The wages of an occupation must as a rule be sufficient to compensate the costs of training. (2) The wages of an occupation must be sufficient to compensate special disadvantages (such as a large chance of unemployment). (3) Wages must be higher in large cities than in small because living costs are higher in large cities. (4) Wages must compensate men for their additional skill as they acquire experience (they usually reach a peak of earnings between ages 40 and 55) and thereafter decline as the worker's efficiency declines. (5) Wages will reflect differences in taxation, fringe benefits (pensions, vacations), etc. Accordingly the wage structure even for a single occupation in a single city is elaborate. When a single wage (price) is imposed upon an occupation, labourers are no longer properly distributed by wages; for example, a city school system that pays all teachers of given experience the same wage finds it difficult to staff its less attractive schools.

The preferences of the individual worker cannot be given full play, or each man would become president of the corporation at a sumptuous salary. Yet the labourer may choose to live in California rather than Maine; then the price system will incite employers to move their operations to California, where they can hire this labourer more cheaply. The labourer may prefer to work long hours or short hours, and employers are induced by wage offers to cater to these diverse preferences. In fact it is equally appropriate to speak of the worker buying conditions of work and of the employer buying the services of the worker.

The conservation of resources. A society has some resources that can be replaced by investment: timber is now largely grown as a commercial crop. Farmland is a more ancient example: the fertility of soil can be increased by prudent cultivation. Other resources are not replaceable, such as coal and petroleum. How does the price system conserve these exhaustible resources?

Pricing and conservation

The method of using a resource is independent of the pattern over time of income and expenditures that the owner of the resource desires. Suppose that a farm will have a value of $100,000 if it is maintained at a constant level of fertility and yields a yearly income of $10,000 forever but that it can be cultivated ("mined") intensively to yield $12,000 a year for five years at the cost of a much reduced yield thereafter, with a value of $90,000. Even if the farmer is in urgent need of immediate funds and does not expect to live more than five years, he will still cultivate the farm at the uniform rate. Only then is it worth its maximum value to him, and only then (by sale or mortgage) can he obtain the largest possible funds even in the near future. In short, one need not adapt his expenditure pattern to his income pattern so long as he can borrow or lend.

If the growth of consumption or the decline of reserves threaten the exhaustion of supplies of a resource, then the price of that resource will rise and promise to rise more in the future, and this rise will serve to reduce current consumption and to reward the owner of the resource for holding back much of the supply for the future. This rise in price will therefore also stimulate buyers to find more economical ways of using the commodity (for example, burning the fuel more efficiently) and stimulate producers to find new supplies or substitute products. The price system will, therefore, ensure that the supply of the re-

source will be stretched out so that the resource will be available in both the present and the future.

The special case of the automobile. The number and severity of automobile accidents would seem to lie outside the system of prices, but it can be shown that they are much influenced by prices. The driver may buy protection by owning a new and heavy automobile rather than an old and small one; for the heavier the car and the better its tires and brakes, the lower the rate of serious accidents. Every driver, in effect, accepts some rate of accident: his driving habits and the highway conditions will on the average lead to some rate of accident per thousand miles of driving. The frequency of these accidents could, of course, be reduced but only at the cost of slower travel, fewer trips, and other costs that are higher than the cost of accidents. The fines imposed upon speeders are, in effect, a price charged for fast driving.

If the price system for drivers were not supplemented by extensive regulation and licensing, each driver would adjust his manner of driving so as to take those risks of accident that on balance "paid": the expected benefit from the driving would equal the expected cost of accidents. He might wish to insure himself against these risks, just as he insures his house against fire; but then his insurance premiums would reflect the level of accidents to which his driving led.

Pricing and accidents

Here it is necessary to take account of a previously neglected factor: the use of a price system involves substantial costs. When an accident occurs, under the price system the fault of each party must be determined, and this can give rise to lengthy and expensive litigation. Some recent proposals in the U.S. to settle small claims by a state fund, without determination of fault, are designed precisely to eliminate these costs of a price system. Of course the alternative system of compulsory insurance will have its own costs, one of which is the reduction in the incentive to drivers to be careful.

Minor accidents may be viewed as a form of waste, and it is never economically feasible to eliminate all waste. To pursue efficiency to its very limit is of course inefficient; the economist J.M. Clark commented on the "irrational passion for dispassionate rationality." A penny saved is indeed a penny earned, but this may be a low wage.

LIMITATIONS AND FAILURES OF THE PRICE SYSTEM

The price system is an extraordinarily powerful instrument in organizing an economic system, but it is subject to three broad classes of limitations.

Private and public price control. Sometimes prices are not permitted to do their work. Monopolies are able to exert control over prices; and they use it, sensibly enough, to raise their profits above the level allowed by competition. The monopolist (or group of colluding enterprises) sets prices at a level such that prices are above costs or, to use words of identical significance, such that resources earn more in the monopolized industry than they can earn elsewhere. The basis of the monopoly is its ability to prevent outsiders from entering the industry to share in the unusual profits and, by the act of producing, actually serve to eliminate them.

The cost of monopoly

The fixing of prices by monopolists reduces the income of society. This is, in fact, the only well-established criticism (on grounds of efficiency) to be levied against monopolies: there is no reason to assume that they will make products less suited to consumer tastes or innovate more slowly or pay lower wages or otherwise misallocate resources. But the basic inefficiency led, first in the United States in 1890 and then increasingly in European nations, to governmental policies to maintain or restore competition.

Public price control has two aspects. A large part of public regulation is intended to correct monopolistic pricing (or other failures of the price system); this includes most public-utility regulation in the United States (transportation, electricity, and gas, etc.). Whatever the success of these endeavours—and on the whole there has been a substantial decline in confidence in the regulatory bodies—they are usually instructed to achieve the goals of an efficient price system.

Other public price controls are designed to serve ends outside the reach of the price system. Prices of farm products are regulated (raised) in most nations with the intention of improving farmers' incomes, and the fixing of interest rates paid by banks is undertaken to improve bank earnings. Such policies are invariably defended on various economic and ethical grounds but reflect primarily the political strength of large and well organized producer groups.

Externalities and the price system. A second class of limitations consists of things that should be done but are not performable by a price system. Even when prices are freely established by competition, there is a class of economic relationships called "externalities" not efficiently controlled by prices. These may be illustrated by the air pollution caused by automobiles. Since no single automobile makes a significant contribution to air pollution, the owner has no incentive to bear the cost of installing antipollution devices even though all drivers would be better off if each did so. Yet if there are many automobiles in a region, it would be prohibitively expensive for drivers to contract with one another to have each install devices in his automobile to reduce pollution. The external effects of any one automobile's exhaust fumes are so diffuse and affect any one person so triflingly that they cannot be regulated by the price system.

The class of "externalities" is as broad as the class of actions that have effects upon people who are not parties to the contracts governing the actions. An attractive garden pleases passers-by, but they cannot be charged a portion of its cost. A new piece of scientific knowledge will prove useful to unknown persons. These two examples indicate that some externalities are economically trivial and some are highly important.

When the price system cannot deal with diffused effects, other social controls often take its place. The state invokes a whole arsenal of policies to deal with externalities, of which the following are only examples: (1) The state may subsidize activities that do not end in a product that can be sold. Thus basic scientific research that does not lead to patentable processes is subsidized. (2) Individuals may be compelled to act uniformly in areas where contracts would be too expensive; traffic laws, zoning laws, and compulsory vaccination are examples. (3) The state may itself undertake an activity that cannot be financed by sale of services, the most obvious example being national defense.

Traffic congestion and prices

An interesting type of externality is the problem of highway congestion. Any one person's presence on a highway at a time and place of peak density has only a negligible effect upon others, so that, except on toll roads, private contracts have not been feasible. The state itself has not been able to deal effectively with highway congestion. More highways can be built until no highway is ever crowded, but this would be intolerably expensive. The state has lacked a method of inducing drivers to shift to less crowded hours and routes by charging fees to those drivers who impose high congestion costs by driving at peak times. Recent developments in technology may make it feasible to use the price system to reduce congestion. For example, cameras at appropriate points could photograph automobile licenses and a computer could accumulate the charges based on route and time for each automobile. Then only a person for whom travel at peak times was worth, for example, 25 cents per mile would impose (and pay for) the congestion he created.

Imperfect knowledge and tastes. A third class of limitations to which the price system is subject has to do with the control of knowledge and tastes. To the extent that an economic actor, whether a consumer, a labourer, or an investor, is poorly informed, he is likely to make decisions whose consequences are much different from those he desired and expected. What follows relates only to consumer decisions, but parallel issues arise in labour markets, securities markets, etc.

A consumer can satisfy his desires only if he makes intelligent purchases—that is, only if the goods he buys are what he believes them to be. How can the consumer know whether the meat is free of disease or whether the

washing machine will function well and long or whether the fabric of the garment is one synthetic fibre or another? To ascertain these facts personally, the consumer would have to be a versatile scientist equipped with a superb laboratory—and then he would need to spend so much time testing goods that he would have little time to enjoy them.

In some measure the consumer does experiment in his buying: whenever he buys a thing repeatedly, experience tells him much concerning its properties. Direct experience is a sufficient guide in buying celery or hiring domestic servants, but usually the purchase of information takes a less direct form. The city's premier department store can sell at prices somewhat higher than less well-known retailers; and the difference represents the payment of a price for reliability, responsibility, and the guarantee of quality. In parallel fashion the consumer buys the washing machine of a company that made his excellent refrigerator. Occasionally, information is bought directly: the advice of a lawyer, the knowledge of an appraiser, the taste of an interior decorator.

The need for information

The most important and controversial method of informing consumers is by advertising. Many critics are outraged by the self-serving statements of sellers, some of whom indubitably provide irrelevance and deception rather than information. Yet the informational content of advertising may not be as deficient as its critics believe: advertising itself meets two market tests. In the first place, the direct sale of information by consumer advisory services has never become important, although there are no obstacles to entering this business. In the second place, there has been a general, sustained improvement in the quality of consumer goods over time: the automobile tire goes many more miles than formerly; the airplane flies more safely.

Nevertheless, recent public policy has paid great attention to increasing the safety of products and to raising the accuracy of advertising claims.

Knowledge is sometimes difficult to distinguish from taste: does the consumer who persists in smoking cigarettes have inadequate knowledge or simply different comparative values for the pleasures and risks of smoking? Censorship, in any event, is fairly common in every economic system: no society allows young children or incompetents full freedom of action or allows the unlimited sale of narcotics. Since the price system never forbids an effective demand (a demand backed by a willingness to pay the supply price), some form of restriction of prices is, therefore, necessary if certain tastes are to be forbidden or restricted. Compulsory school attendance can be viewed as, in effect, a form of censorship; and so are the controls on sale of firearms and the taxes on tobacco and liquor.

OTHER PRICE SYSTEMS

The foregoing discussion has been confined to the price system as it exists in capitalist economies. The Communist countries have prices, but not autonomous price systems; in those countries the direction of economic activity is largely in the hands of the central authorities, and prices are used mainly as a marketing device. None of the three allocative functions of an economy—determination of what will be produced, of how it will be produced, and of who will get the product—is performed by the price mechanism in the Soviet-type economies.

The relative scarcities that money prices measure exist, of course, in all countries and would exist in a world where no money or exchanges were allowed. Robinson Crusoe had a problem of allocating his time between sleep, garnering food, building shelter, etc.; and he confronted implicit costs of extending any one activity, for more food meant less of other things. The economist calls these implicit exchange ratios "shadow prices," and they appear in one form or another in all areas of life in which men make deliberate choices.

Price systems are therefore the result of scarcity. The basic proposition of economics, that scarcities are essentially ubiquitous, is often phrased as "there is no such thing as a free lunch"; and it reminds one that the price of the lunch may be future patronage, a reciprocal lunch, or a boring monologue. The task of economic organization is the task of devising price systems that allow a society to achieve its basic goals.

BIBLIOGRAPHY

Major historical treatises: ADAM SMITH, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vol. (1776); JOHN STUART MILL, *Principles of Political Economy*, 2 vol. (1848), in his *Collected Works*, vol. 2 and 3 (1965); ALFRED MARSHALL, *Principles of Economics*, 9th ed., 2 vol. (1961; pub. orig. in 1890).

Contemporary textbooks: A.A. ALCHIAN and W.R. ALLEN, *University Economics*, 2nd ed. (1967), for a wealth of analyses of prices; P.A. SAMUELSON, *Economics*, 8th ed. (1970), long the most popular elementary textbook; F.H. KNIGHT, *The Economic Organization* (1965), a classic statement of the functions of an economic system; G.J. STIGLER, *The Theory of Price*, 3rd ed. (1966).

Selected analyses: W. VICKERY, "Pricing as a Tool in Coordination of Local Transportation," in *Transportation Economics*, pp. 275–296 (1965); M. FRIEDMAN, *Capitalism and Freedom* (1962), esp. ch. 6–12; R.A. RADFORD, "The Economic Organization of a P.O.W. Camp," *Economica*, 12:189–201 (1945), an account of the evolution of a cigarette-based price system.

(G.J.S.)

Priesthood

Throughout the history of religion, the official institution that mediates and maintains a state of equilibrium between the sacred and the profane aspects of human society, and that exercises a stabilizing influence on the social structures and cultic organizations, is the priesthood. The term priesthood is derived etymologically from the Greek word *presbyteros* ("elder"), of which it is a contraction, and is equated with the Latin word *sacerdos* (the Roman officiant at the sacrifices and sacred rites).

NATURE AND SIGNIFICANCE

The role of the priest

The primary role of the priest is that of the ritual expert, the one who has a special and sometimes secret knowledge of the techniques of worship, including incantations, prayers, sacrificial acts, songs, and other acts that are believed to bridge the separation between the divine or sacred and the profane realms. The priest gains such knowledge through the institution known as the priesthood, which may be composed of various groups or guilds devoted to all or only a few aspects of the priestly craft. Because the priest gains his special knowledge from a school for priests, he is differentiated from other religious and cultic leaders, such as the magician, shaman (healer and visionary), diviner, or prophet, who obtain their positions by means of individual efforts (*e.g.*, learning from a master magician or diviner; individual ecstatic experiences that are publicly recognized). As a member of the institution that regulates the relationship between the divine or sacred and the profane realms through ritual, the priest is the accepted religious and spiritual leader in his society.

At various times in the history of a culture or society the priestly institution may be attacked by other institutions or groups that vie for the religious leadership (and thus sometimes the social, political, and economic leadership) of a people. Such anticlericalism is a phenomenon not only of modern society (as noted in the Russian Revolution of 1917, the Mexican Civil War that began in 1857, and other less dramatic movements) but also in the ancient world, such as in Egypt in the 14th century BC, when the priesthood of the god Amon and the priesthood of the god Aton changed positions. Anticlericalism may be fostered by battles for religious leadership between two or more opposing priestly groups, or by prophets and others who are concerned with religious experiences in their personal rather than in their institutional forms. Among Protestants, the doctrine of the priesthood of all believers (*i.e.* all believers have direct access to God) militates generally against strong anticlerical tendencies within their own ranks. In Islām, there is, technically, no priesthood, though there are local spiritual and community leaders, such as the *imām*, the *mullah*, the *muftī*, the *qāḍī*, and others.

THE PRIEST AND HIS OFFICE

The function of the priest as the mediator and maintainer of the equilibrium between the sacred and the profane in human society, and as the stabilizer of the social structures and the cultic organizations, determines the various criteria for holding the priestly office. In preliterate society the functions are accomplished by ritual experts who are trained in the special knowledge and techniques of magico-religious disciplines in which sacred power is believed to be inherent. They also are trained in disciplines that enable them to gain a supernormal psychic knowledge and in the techniques of mystical experience. As agents of the sacred power, they are believed to have the ability to control and manipulate through ritual natural processes and events, and to engage in a sacramental relationship between man and the sacred order on which the community and its members depend for their sustenance, survival, and well-being.

Functions and criteria

Because religious institutions and beliefs are intimately connected with other social institutions, those who are set apart to establish efficacious relations with the transcendental powers that are believed to control the universe, as well as human affairs and destinies, occupy a key position at the dynamic centre in the social structure. Thus, in certain types of societies the office and functions of the priest may be limited to those having a particular ancestry, those belonging to certain tribes (such as the Levites in Judaism), families (such as the Eumolpids of the Eleusinian mystery religion of Greece), or castes (such as the Brahmins in Hinduism), and those initiated into certain professional orders (such as the cure doctors among the Maya).

Besides such sociocultural criteria, there are also certain personal requirements in various cultures for those who would become members of a priesthood. Celibacy (as in Roman Catholicism and the Arcakas of the Digambara sect in Jainism), asceticism (as in various Buddhist groups), and religious experiences (as among some holiness Protestant sects) are among the personal requirements for those who aspire to or are chosen to assume the priestly office.

The religious functions of priests are quite varied. In his specific role as the officiant of the rites that unite the sacred and the profane realms, the priest as a *pontifex* (from the Latin word that means maker of a bridge) celebrates or administers at the rituals of initiation into the cult or church, presides over ritual re-enactments of creative, redemptive, or salvatory (salvation-working) events, and offers sacrifices to the gods or to one God. He also functions as a perpetuator of the sacred traditions, practices, and beliefs and as a teacher, healer, counsellor, and diviner.

In all their respective offices, functions, and capacities, those who have exercised and manipulated sacred power have attained a uniquely prestigious position as the spiritual and social leaders *par excellence*. When perplexing and emotional situations that appear to be beyond human control, knowledge, skill, or techniques have arisen in a community, the people have a recourse in the priest, who has the special knowledge of the relationship between the divine or sacred and the profane realms. The priest is often called upon at critical junctures in the lives of individuals (such as birth, puberty, marriage, and death) and in the life of a community (such as seasonal changes or at times of flood, drought, and famine).

PRIESTHOOD IN THE RELIGIONS OF THE WORLD

Primitive cultures. The office of priesthood in primitive society ranges from the medicine man, who magically manipulates the sacred power as a quasi-impersonal force, to the shaman, who, as an agent of divine or spiritual beings, may be at once a medicine man, a visionary, occult diviner, and a genuine sacerdotal cult leader. Not infrequently, chiefs or headmen of a clan or village may become sacred men or ritual experts displaying supernatural insight and knowledge as the agents of spiritual beings. As such, they may eventually become prophets or sacral kings occupying a unique status in the community. This status may be acquired by accepted claims of descent from a mythical divine ancestor or god on whom the fertility of the crops and the welfare of the community in general are thought to depend. Occult power and insight may be derived from spirits with whom the prophet or diviner is related in a trance or ecstasy, or else in dreams, visions, or in auguries and oracles (divination techniques) that make known the divine will and predictions. To occupy such a position, a strenuous course of training is frequently required involving some knowledge of therapeutics (healing), leechcraft (use of leeches in healing), trephining (operating on the skull), herbs, poisons, and perhaps sleight of hand and similar techniques, together with the development of psychical and occult qualities. While charlatanry (pretentions to medical knowledge) is often practiced, the office also demands an understanding of the technical equipment calculated to bring about the results that are sought, and the proper type of temperament, conditions of mind, and state of emotion.

Medicine men and cult leaders

Unlike the shaman and the medicine man, both of whom exercise their respective functions while relying largely on their own initiative and psychic and occult powers of divination, healing, and direct access to the spirit world at their command, the priest supplicates and conciliates supernatural forces superior to himself, guards the sacred tradition in his care, and acts as the master of sacrifice. The shaman or magician officiates in his own name and by his own methods and techniques; the priest serves the altar, in the temple or shrine, as the representative of the community in his relations with the gods and the sacred order by virtue of the status and its functions that have been conferred upon him at his ordination, bestowing its sacredness and attendant taboos.

The ancient Near East. In the Nile Valley the occupant of the throne (the pharaoh) was regarded as a god incarnate, the sacred king. Because he was believed to be the epitome of all that was divine, he alone was the intermediary between mankind and the gods whom he summed up in his complex personality. Before the rule of Upper and Lower Egypt was centralized by the founder of the 1st dynasty (*c.* 3100–*c.* 2890 BC), Menes, the high priest and king "Scorpion" was traditionally considered to be the incarnation of Horus (the sky god). Under the powerful influence of the priesthood of Heliopolis, the sacred kingship was given a solar significance; the pharaohs were represented as the sons of the sun god Re, who was identified with the god Atum. When Memphis became the imperial city, Re-Atum was brought into relation with its god Ptah (the creator) and subsequently with Osiris, a fertility god who also was regarded as a god of the dead. Osiris was equated with the life-giving waters of the Nile and believed to be the first civilizing king, reigning in the person of the pharaoh as his posthumous son Horus. Therefore in Egypt the divine origin and status of the monarchy was so firmly established that it became the stabilizing force of the civilization in the Nile Valley.

Sacred kingship and the priesthood in Egypt

In theory, the king of Egypt was the high priest of every god, and in all important ceremonies he alone was depicted in the temple scenes as the officiant. For practical purposes, however, he delegated his functions to the particular professional local priesthoods. In due course the priests, courtiers, and officials shared in his divine powers and privileges. Even though they only acted on behalf of the divine ruler, they shared in some measure his personality and potency. As the priest *par excellence*, the pharaoh remained unique, and it was from him that the priesthood and nobility exercised their sacerdotal functions when the solar theology of Heliopolis was established in the 5th dynasty (*c.* 2494–*c.* 2345 BC). After the 5th dynasty he was accredited exclusively with having built all the temples, and on their walls he was portrayed officiating as the mediator between mankind and the gods in all important rites. In fact, nevertheless, his place was taken by one of the retinue of priests attached to the royal household or to the local temples. Only very occasionally did he perform his sacerdotal duties in person. Every temple had its high priest who, in addition to his sacred offices, might also be a high judicial official, thereby combining sacred and profane roles, varying with the actions performed.

From the number of titles assigned to priestly officials in the Old Kingdom (*c.* 2686–*c.* 2160 BC), it appears that one priest must have served simultaneously a multiplicity of cults in the shrine to which he was attached and received a share of the offerings and emoluments. In wealthy great temples in the New Kingdom (*c.* 1567–*c.* 1085 BC) there was an elaborate organization, and a large staff engaged in administrative, civil, and educational duties, as well as in their sacerdotal and mortuary functions. There were also a number of priestesses associated with the mother goddess Hathor (wife of the sun god Re), who were mainly concerned with playing the sacred sistra and other musical instruments. From the 18th to the 21st dynasties (1567–1085 BC), however, under the Theban priesthood of Amon-Re, the priestly hierarchy was able to create an absolute "state within a state."

Priestly classes and functions in Mesopotamia

In Mesopotamia, where kingship occupied a less prominent position than it did in the Nile Valley, powerful priesthoods and highly organized temples were firmly established in and after the 4th millennium BC. The temples were centres of sacred learning of the content and methods of incantation, prognostication, exorcism, and of political and economic administration, and their attendent priests were divided into classes with special sacerdotal and secular functions. The valleys of the Tigris and Euphrates rivers contained a series of city-states loosely bound together under the rule of a governor (*patesi* or *ensi*) and of a high priest (*sangu mah*). Their main functions were to integrate the temple communities of the city, their revenues and boundaries, and the *unigallu*, or head of the priesthood, was responsible for the performance of the seasonal rites, with the assistance of those who sang the liturgies and directed the temple music and those who uttered the lamentations, invoked the oracles, and interpreted the astrological signs and omens. The duties of other members of the priesthood included making incantations and pronouncing spells and exorcisms. The seers, or *baru*, foretold the future by means of soothsaying, interpreting dreams and visions, and revealing revelations.

An important function of the priesthood in the ancient Near East was that of exorcism. All down the ages good and evil have been viewed as two contending forces in perpetual conflict, and in the dual task of the riddance of evil and the induction of good the exorcist, the shaman, and the seer have exercised complementary functions hardly distinguishable in the higher cultural levels in Babylonia and Assyria. When the function of the driving out of demons from human beings and building by incantations and ritual expulsions was separated from the interpretation of omens and astrological portents, the office of the exorcists was in ever increasing demand. According to cuneiform texts from around the 3rd millennium BC, exorcists held a position of supreme importance, assisting at the consecration of temples, at funerals, and at seasonal ceremonies. In the temple schools, exorcists, diviners, and astrologers, together with physicians, scribes, and judges, fostered the study of astronomy, medicine, and jurisprudence. In their sacerdotal functions and capacities they were regarded as the agents and exponents of the gods, having a monopoly on sacred knowledge, controlling for good or evil the destinies of the community and its members. Therefore, their power was enormous, and they maintained the relation between religion and culture by their all-embracing hierarchy that determined the present welfare and future destinies of both human beings and the social order.

Canaanite priesthood

Similar beliefs and practices occurred in northern Syria in the middle of the 2nd millennium BC, before the Israelite settlement in Canaan. Here, again, the priesthood was responsible for the dramatic rituals on which the social structure and the well-being of mankind were believed to depend, especially in the climax of the autumnal festival that culminated in the enthronement of the year god. The priestly movement was centred in the temple of Ugarit, and the religious texts discovered there (Ras Shamra) were essentially liturgical documents devised to make the sacred drama enacted efficacious. Baal, the weather god, was regarded as "the lord of the furrows of the field," responsible for the rain and the production of the fruits of the earth. A series of temples was erected in his honour in Syria and Palestine. The temple at Ras Shamra was a considerable structure with a lattice, the opening of which was supposed to produce the autumnal rains, and its liturgical ritual was celebrated by the priests and priestesses in their official capacities as those who personified the gods and goddesses responsible for providential bounty and beneficence.

With the advent of Persian sovereignty in Mesopotamia, a new approach to cosmology and mythology was introduced. In the religion founded by Zoroaster (*c.* 7th century BC), the ultimate ground of the universe was reduced to a single supreme deity, Ahura Mazdā, the All-wise Lord. Astrology became widely practiced by the Iranian Magi (priests of the Zoroastrian religion) who were highly skilled in the knowledge of astral lore and the signs of the zodiac. In the three centuries after the Persian occupation of Mesopotamia (6th century BC), the Babylonian priests under Iranian influence made remarkable strides in astronomical learning that ultimately dethroned astrology from its exalted position in sacred tradition. Beyond the Euphrates Valley, however, it survived until after the Muslim rule, when the astrologer and the astronomer became indistinguishable.

Greece and Rome. The ancient Greeks were devoid of hierarchic institutions composed of men and women through whom the gods were approached, though priests and priestesses could be found in many places engaging in specific sacerdotal functions and ritual acts. Some of them attained considerable social and civic prestige and importance and were attached to particular temples or shrines such as the oracle at Delphi, which was consulted on private and state matters. Their duties, however, were generally those of members of a household engaged in everyday affairs, rather than of a caste or sacerdotal order set apart and consecrated for the performance of sacrificial and other rites, functions, and practices. Though not regarded as mediators between the gods and men, they did act in such ritualistic capacities in certain civic and administratve areas. There was no specific distinction, however, between them and lay members of society. On the contrary, such officials as magistrates might be priests and vice versa. Some exercised considerable influence if they were regarded as outstandingly efficient, wise, or distinguished in their respective civic or religious capacities.

Civic importance of the priesthood in ancient Greece and Rome

Similarly, in ancient Rome when the agricultural religion of Numa (the legendary second king) was transformed into an institutional state cult in the republic, it was organized as a hierarchy with the *rex sacrorum* ("king of the sacred things") inheriting the office and attributes of the former priest-king. He had to be a patrician and was chosen for life, subordinate only to the *pontifex maximus*, who was the head of the college of *pontifices* ("advisors on the sacred law") and *flamines* ("priests devoted to a particular god"), three of whom were assigned to the gods Jupiter, Mars, and Quirinus, the remaining 12 to other deities. The flamen Dialis, dedicated to the supreme sky god, Jupiter, occupied a unique position socially, politically, and sacerdotally and was subject to strict taboos and regulations because of his sacred office. The *flaminica*, the wife of the flamen Dialis, participated in his sacredness and official status, and so vital was her association with him and his office that if she died he ceased to perform his functions.

Attached to the temple of the goddess Vesta on the Forum in Rome were the six Vestal Virgins dedicated from childhood to the service of the sacred fire in the *atrium Vestae* (hearth temple), and to the care of the store house (*penus*). Originally they were selected from patrician families by the *pontifex maximus*, but later plebians were eligible for election. During their 30 years of service to the goddess, beginning in childhood, chastity had to be strictly observed on pain of death by starvation, but after the completion of the period of service the virgins were free to marry. The duties assigned to them at very ancient festivals, such as the Lupercalia (a fertility festival) on February 15, and at other occasions, indicate their unique position and significance in the state cult going back, in all

probability, to their origins in the family tradition, associated with the *pontifices* as the officiating priests. Augurs (divinatory personages) had a powerful influence on state religious beliefs and practices, especially in divination to ascertain the will of the gods and the blessing of the crops. They also interpreted signs in the sky as good or bad for the guidance of the magistrates. At the end of the republic (in the 1st century BC) this practice led to abuses that were ridiculed by the politician and orator Cicero. Among other groups of Roman priests were the Salii on the Palatine Hill (the 12 Leapers of the god Mars), and the Luperci, whose sacerdotal functions were confined to the Lupercalia. The *fetiales* were Roman officials employed in making treatises or declarations of war, whose work gradually fell into disuse at the beginning of the empire (late 1st century BC) when the state cult was in decline and losing its vitality.

Ancient Judaism. When Christianity became the legal religion of the Roman Empire after AD 313, it had already inherited from its Jewish background a concept of an organized priesthood. The Jewish priesthood had been centralized in the Temple at Jerusalem (destroyed by the Romans in AD 70) since the 10th century BC. The Hebrew designations for those who exercised oracular, divinatory, and ecstatic functions in ancient sanctuaries that were prominent cultic centres prior to the building of the temple, such as Mamre, Hebron, Bethel, Shechem, and Gilgal, were *cohen, levi, navi'*, and *ro'e*, corresponding to priest, Levite, prophet, and seer, respectively. *Cohen* is the equivalent of the Arabic word *kāhin*, diviner, and in Hebrew it has the meaning of "priest," denoting the occupant of the office concerned with obtaining oracles by the aid of the ephod ("pouch") containing the Urim and Thummim (sacred lots) and by inspiration, as well as with officiating at a sanctuary. After the 7th century BC, when worship was concentrated in Jerusalem, the capital, the priesthood was restricted to the Levitical house of Aaron (brother of Moses, the 13th-century-BC lawgiver) after having been previously drawn from other lines of descent, such as those of David, Nathan, Micah, and Abinadab (royal, prophetic, and priestly families).

Priests, Levites, prophets, and seers

Whether in fact the Levites ever were members of a sacerdotal tribe is open to debate, but in any case they represented a special fraternity set apart to be guardians of the sanctuary and to engage in oracular and prophetic function, over against the rival priestly *cohenim* in their respective independent confraternities. It was not until after the Exile of the Jews to Babylon in 586 BC, when the Priestly Code was drawn up, that the distinction between priests and Levites became absolute. The priesthood was confined exclusively to those claiming succession from Aaron, in spite of the Zadokites claiming priestly descent from Eleazar as an "everlasting covenant" (Num. 18:2–7, 25:13; I Chron. 24:37). The Zadokites may have constituted the survival of an ancient Jebusite (Canaanite) royal priesthood, giving them special duties and privileges in the Temple worship above those of the Levites. Later, when the priesthood became reserved for the descendants of the family of Aaron alone, the title was restricted to members of the non-Aaronic families of the tribe acting as the servants of the Temple.

The oracle given by the priests as the inspired word of the Law, called the Torah, which was referred back to Moses in postexilic Judaism, acquired a new significance, involving a rigid observance of its ritual and legal commands that permeated every aspect of life, worship, and conduct. The cessation of the daily sacrifice and other Levitical priestly ministrations in the Temple after the fall of Jerusalem (AD 70) gave a new emphasis to and interpretation of the Torah in the synagogue and in domestic rituals. The prerogatives of the high priest, and those of the priesthood in general, with its exclusive lineage, were maintained after the revolt of the Jews under the leadership of the Maccabees against the Hellenistic Syrians in the 2nd century BC, and the priestly blessing (*dukhan*) in the synagogue remained the exclusive right of the *cohenim* claiming descent from Aaron. They also have had the right to be the first called upon to read the Torah in the synagogue, followed by a Levite. Their privileges, however, have been questioned by some rabbinical (nonpriestly Torah scholars and religious leaders) authorities. The Sadducees (deriving their name from the Zadokites) were the high priests in Jerusalem during and after the time of the Hasmoneans, the descendants of the Maccabees (135–104 BC). They exercised considerable influence in the Jewish Sanhedrin (supreme rabbinic court) as the conservative class of the religious aristocracy, favoured accommodations to Greek culture, and maintained the importance of the letter of the Written Law over against the oral tradition of the rival Pharisees. The high priesthood, however, was declining in status under the increasing control of the Roman authorities.

Emphasis on interpretation of the Torah

Christianity. At this critical juncture in Judaism, Christianity, with its own particular conception of priesthood and sacrificial redemption, began in Palestine and rapidly spread throughout the surrounding regions in the Greco-Roman world. In the New Testament, the imminent destruction of the Temple in Jerusalem and its worship is predicted, and the culmination of its high priesthood in the person of Christ, after the order of Melchizedek (the priest-king of Jerusalem revered by Abraham in the Old Testament), is proclaimed. The Jewish Aaronic priesthood and its ritual are represented in the Letter to the Hebrews as imperfect shadows, in a Platonic sense, of the archetypal order of the eternal sacrifice of Christ. Only Christ, who was described as "beyond the veil" (referring to the veil that separated the "Holy of Holies" section from the other areas of the Temple), was believed to be able to save those who came to God through him, since he had removed the barrier of sin that separated man and God for those in a state of grace. In Christ's reconciling offering as both priest and victim on the cross, the writer of the Letter to the Hebrews stated that he accomplished the removal of the barrier in the heavenly tabernacle. This was interpreted in terms of the Hebrew Day of Atonement, a ceremony in which the Jewish high priest made expiation annually for himself, the priesthood, and the whole congregation of Israel. The view of Christ as king, high priest, mediator, and victim influenced the establishment and gradual development of the Christian priesthood in the church, which shares in and makes accessible to its baptized members the all-sufficient priesthood of Christ.

The relationship between presbyter and bishop

Originally the term *presbyter* ("elder") and *episkopos* ("overseer"), current in the New Testament and the early church, were probably identical. From the 2nd century on, however, the sacerdotal hierarchy developed along the lines of the Hebrew priesthood, the title *episcopus*, or bishop, becoming reserved for those who presided over the presbyterate, then called *sacerdotes* because they shared in the episcopal *sacerdotium* ("priesthood"), which included the offering of the eucharistic sacrifice of bread and wine. But the conferring of holy orders (ordination of presbyters) and administering the sacrament of confirmation were confined to the episcopate, together with administration of the diocese (jurisdictional area). In due course the threefold ministry of bishops, priests, and deacons (administrative and liturgical assistants in a parish) became organized on a diocesan basis. This remained the norm in the Western Church until the Reformation in the 16th century, when it was repudiated by the continental Reformers (*e.g.*, Luther, Calvin, and Zwingli). In Roman Catholicism, Anglicanism, Swedish Lutheranism, and Eastern Orthodoxy, apostolic succession and jurisdiction has been maintained, especially in the Roman Catholic papacy and in Eastern Orthodox patriarchates.

In European Christianity in the Middle Ages (*c.* 6th to 14th centuries), the deeply laid tradition inherited from the theocracies and priesthoods in the ancient Near East of a common social, judicial, political, and religious structure was sustained and gave expression to medieval civilization, especially in the Latin West. Church and state became so closely associated that they were virtually identical, as in the cases of the sacred and secular in preliterate societies. As Dante (1265–1321) contended in *De Monarchia* ("Concerning the Monarchy"), the pope, as the head of the spiritual aspects of society, and the emperor, as the ruler of the temporal areas of concern, were equal-

ly ordained by God to exercise their functions in their respective spheres of power and influence for the welfare of mankind. If this duality of control did not endure, because the church gradually usurped more and more of the civil jurisdiction and dominated emperors, kings, and other ecclesiastical rulers, the unification of the body politic was rendered more complete as an integrated whole, and the life and character of medieval civilization was determined through the papacy and its priesthood.

Hinduism. In Vedic India, the early period of Hinduism, when the priestly caste (Brahmin, or Brāhmaṇa) was vested in a particular tribe or special class, it occupied the primary place of importance in the segmentation of Hindu society. The king was subordinate in some respects to the Brahmins, though at one time both sometimes were chosen from the Kṣatriya, or warrior caste. Nevertheless, because the existence of the universe and all cosmic processes were made to depend upon sacrificial offerings, the king delegated such functions to the priests before the end of the 7th century BC, the priests having usurped that position previously held by the kings. The priesthood then exercised supreme control over the fortunes of the gods and men, of heaven and earth, and of the state, though not infrequently priests worked in the service of princes. The Brahmins, however, were so firmly established in the caste system as the twice-born masters of sacrifice and of the sacred knowledge that they alone possessed that they were viewed as holding the universe in their grasp. As "lords of creation" by divine right they were divided into ten tribes, above all other castes. They were required to pass through four *āśramas* (the celibate religious student, married householder, forest hermit, and wandering ascetic), or conditions of life, prior to and after marriage, to become anchorites (or hermits), and to attain the plenitude of their status, vocation, and authority, thus renewing the creative process by the due performance of the sacrificial offering.

The role of the Brahmin caste

Against this Brahmanic sacerdotalism and its caste organization, the reaction noted in the *Upaniṣads* (writings representing the end of the Vedic period) about 600 BC introduced a mystical conception of the priesthood in Hinduism, and subsequently in Buddhism. Living as hermits in the forest, groups of mystics, reacting to Brahmanic ritualism, gathered around them disciples to learn and propagate a philosophical doctrine based upon the quest to discover a new function for the priesthood, in which the identification of the inner eternal self of man (*Ātman*) with the divine ground of the universe (Brahman) was achieved by asceticism, renunciation of the world, and mystical experience and realization, rather than by the sacrificial offering. This quest was eventually associated with the meditative techniques of *yoga* (mental and physical exercises) and opened the way for the rejection of the exclusive claims of the Brahmins in favour of the mystical insights and esoteric knowledge accessible to all who adopted the Upaniṣadic teaching and way of life.

With the establishment of the four *āśramas*, sacrifice and Brahmanic study of the Vedas were rehabilitated and brought into relationship with the Upaniṣadic tradition. The Brahmin was thus regarded as occupying the highest state of life; he was

Revival of sacrifice and Vedic study

> one who has sensed the deepest self and acts out of that consciousness, [communicating it to others], . . . gives moral guidance, . . . lays down the science of values, draws out the blueprints for social reconstruction, and persuades the world to accept the high ends of life. (From S. Rādhākrishnan, *Eastern Religions and Western Thought.*)

This view has lacked, however, a conception of the godhead with whom personal relations are possible, thereby making the priesthood an aspect of an impersonal divine power (Brahman), of a pantheistic principle executing its functions with automatic precision by virtue of its sacerdotal equipment, of an alternating mystical and ascetic priestly sublimation.

Jainism and Buddhism. In Jainism and Buddhism, arising within Brahmanism as nontheistic sectarian movements, the Brahmin priesthood was sublimated and the Vedic caste and sacrifice eliminated. In their place a monastic system was evolved, with monks and nuns devoted primarily to rigorous asceticism in the quest of perfection, including the prohibition of taking life in any state or form, and in the pursuit of chastity and truthfulness. Complete detachment from all phenomenal possessions and connections in Jainism (founded by Mahāvīra in the 6th century BC) made paramount the mendicant life of meditation and spiritual exercises dependent upon the fulfillment of vows of poverty. The functions of the priesthood were sublimated in a process of self-salvation, centred around the purpose of the deliverance of a suffering humanity from the cycles of rebirth. Since in Buddhism *taṇhā* ("desire") was regarded as the fundamental cause of *dukkha* ("the burden of existence"), priestly intervention and the sacrificial offerings were considered to be of no avail in the pursuit of the Eightfold Path leading to the passionless peace of Nirvāṇa (the state of bliss).

In the absence of any conception of a deity in Buddhism, the question of sacerdotal mediation could be ignored, though, in the Mahāyāna ("Greater Vehicle") school, and in the Tantric (esoteric, magical) school, some elements of the priestly tradition survived. The earliest converts to Buddhism were Brahmins for the most part, and a religious organization in monasteries developed, with various prescribed roles for their inhabitants, with daily routines, certain periods devoted to alms and quests, and periods for sacred learning and the translation of literary and theological works. To these activities were added other functions, such as recitations of the sacred texts at births, marriages, and in sicknesses to keep evil influences at bay. In the temples, shrines were erected to the honour of the eternal Buddha, and the image of the Blessed One on a lotus bedecked with flowers has become the central object of worship in certain Buddhist groups. The recitation of the ancient Pāli *sūtras* (divine revelations) is believed to transmit the merit inherent in the texts, as is the endless repetition of the sacred formula *Namu O-mi-to* (*Amida-butsu*), "Homage be to the Buddha of infinite light." This, however, is not a sacerdotal devotion performed by, or requiring the presence of, a priest. Buddhism, in fact, has never been able to produce a strong and lasting ecclesiastical organization or a hierarchical segmentation (as in Hinduism), because it has interpreted unity in terms of Becoming, instead of in terms of Being.

Priestly traditions in Tantric and Mahāyāna Buddhism

Buddhism, Taoism, and Shintō in China and Japan. In the Mahāyāna Buddhist sects, the monks, and those who are popularly known as bonzes, can hardly be said to exercise definitely sacerdotal functions in the temples, monasteries, and shrines. For the most part, these functions have been confined to recitations and invocations, which all of the believers share. In China, the Taoist "priesthood" emerged as an organized institution at the beginning of the Christian Era. Some were celibates and others were married, living ordinary domestic lives. A number were mendicants and some engaged in alchemy and astrology; others were illiterate. There were also those who assisted in ceremonies and collecting the revenues. In the 6th century AD, in imitation of Buddhism, the Taoist celibates lived in monasteries with a patriarch as the head and interchanged facilities with their Buddhist counterparts. In the Zen contemplative sect in Japan, an attempt was made to attain a state of enlightenment (*satori*) by a strict discipline and training in quasi-*yoga* intuitive methods, without priestly intervention or divine grace. The Zen temples and halls of meditation have become centres of learning, art, and education as well as of vigorous austerity and contemplative mysticism, which has not been without attractions to various persons of the West in recent times.

When Buddhism reached China, Japan, and Tibet in the opening centuries of the Christian Era, it came under the influence of the indigenous faiths, cults, and social structures, and, reciprocally, it became a most important influence, adapting its beliefs and customs to those already established in these regions. In the second half of the 6th century AD, after Buddhism had acquired official recognition, pagodas, temples, and monasteries were erected with ornamentations of Buddhist origin. Buddhism adapted itself to Shintō, the native religion of Japan, and to its

shrines, festivals, and rites. The functions of the four priestly classes (*e.g.*, as ritual experts, diviners, musicians, female dancers, and "abstainers" to ward off pollution) that emerged from the family or tribal cults of Shintō were absorbed by Buddhism.

When Shintō was restored as the national religion of Japan in the 19th century, after a period of decline, the Shintō and Buddhist priests were assigned their respective duties and offices by the State Department of Religion without discrimination, for the maintenance of reverence for the gods and love of country (the Truth of Heaven and the Way of Humanity) and proper respect for the sacral emperor (the mikado). This dual sacerdotal combination lasted only until 1875, because Buddhism and Shintō were basically incompatible. This resulted in Shrine Shintō becoming the national faith under the Imperial family, maintaining its divine status, cultic practices, and priesthood, but leaving Buddhism free to propagate its *dharma* (or law) in its own way. New rituals and ceremonies were composed by the government for use at the Shintō shrines, and the duties and grades of the priests were fixed.

THE PRESENT SITUATION

In recent years, in Christianity especially, notwithstanding the doctrinal divergencies and modes of expression in the nature and function of the priesthood, new approaches have been made by both Catholic and Protestant theologians, liturgical scholars, and laymen. This is particularly apparent in the ecumenical and liturgical movements in Western Christendom, in the administration of the sacraments, and in the participation of the laity in the liturgy and in the other offices. In Roman Catholicism, especially under the influence of the second Vatican Council (1962–65), and in the contemporary Anglican Convocations, the "priesthood of the laity" has been more widely recognized and practiced, though this has been a cardinal doctrine of Protestantism since the Reformation. Lay persons trained in liturgical functions have assumed functions, such as reading of the Scriptures and administering the eucharistic elements of bread and wine, in various Protestant churches and, in some instances, in Roman Catholic celebrations that in the past have been the prerogatives of priests. In the Eastern Orthodox churches, such movements and influences have been less effective and operative, largely because of the heavy pressure they have endured from governments that have been hostile in the Communist countries where Orthodox membership is concentrated. They have always, however, preserved a living unity of faith, worship, and organization. Priesthood is inherent in these institutions; it has proved to be the unifying, stabilizing force in Eastern Orthodoxy, the second largest Christian body (after Roman Catholicism), beset as it has been by so many hazards and hostilities in its long and checkered history.

Participation of the laity in liturgical and other functions

BIBLIOGRAPHY. The nature, characteristics, and significance of priesthood in primitive cultures ancient and modern are discussed by GRAHAME CLARK in *Archaeology and Society*, 3rd rev. ed. (1957); EMILE DURKHEIM, *Les Formes élémentaires de la vie religieuse* (1912; Eng. trans., *Elementary Forms of the Religious Life*, 1915, reprinted 1964); and the cultural and scientific development by JACQUETTA HAWKES and LEONARD WOOLLEY, *History of Mankind*, vol. 1, *Prehistory and the Beginnings of Civilization* (1963). The prominent features of priests and kings in the rise of civilization are considered by H.J.E. PEAKE and H.J. FLEURE in the 4th volume of their "Corridors of Time" series (1927). G. LANDTMAN, *The Origin of Priesthood* (1905); and E.O. JAMES, *The Nature and Function of Priesthood* (1956), are comparative and anthropological studies of the subject in its wider aspects with full bibliographies. In ancient Egypt and the Near East the organized priesthoods are treated by J.H. BREASTED (ed.), *Ancient Records of Egypt*, 5 vol. (1906–07), and in *Development of Religion and Thought in Ancient Egypt* (1912, paperback edition 1959); by HENRI FRANKFORT in *Ancient Egyptian Religion* (1948), and in his important exposition of the *Kingship and the Gods* in Egypt and Mesopotamia (1948). The available Sumerian evidence is produced by S.N. KRAMER in his *Sumerian Mythology* (1944). The priestly aspects of the Annual Festival are surveyed by S.H. HOOKE in the *Origins of Early Semitic Ritual* (1938), and in *Myth, Ritual and Kingship* (1958). The Canaanite counterparts are recorded in G.H. GORDON, *Ugaritic Literature* (1949); and in G.R. DRIVER, *Canaanite Myths and Legends* (1956). For the Hellenic conceptions of priesthood, a full bibliography is appended to the article on "The Religion and Mythology of the Greeks" by W.K.C. GUTHRIE in *The Cambridge Ancient History*, rev. ed., vol. 2, ch. 40 (1961). The origin and status of the Levites and the Levitical code in the Hebrew priesthood are investigated by T.J. MEEK in *Hebrew Origins*, rev. ed. (1950). For the origin of the Christian priesthood, see K.E. KIRK (ed.), *The Apostolic Ministry* (1946). In India the Vedic, Brahmanic and Upaniṣadic conceptions of priesthood, and the predominance of the Brahmin caste in Hinduism are discussed in A.B. KEITH, *Religion and Philosophy of the Veda and Upanishads*, 2 vol. (1926); J.H. HUTTON, *Caste in India*, 2nd ed. (1951); S. RADHAKRISHNAN, *Eastern Religions and Western Thought*, 2nd ed. (1940), and *The Hindu Way of Life* (1963); and R.C. ZAEHNER, *Hinduism* (1962), with a full bibliography. For the sublimation of priesthood in Buddhism in India, China, and Japan, see EDWARD CONZE, *Buddhism: Its Essence and Development* (1951); and P. DAHLKE, *Buddhism* (1927). D.H. SMITH, *Chinese Religions* (1968), is an introduction to religious thought and sacerdotal practice in China. *Religious Studies in Japan* (1959), is a very informative composite volume by a group of Japanese scholars in English. Concerning the Zen sect, see D.T. SUZUKI, *The Manual of Zen Buddhism* (1950); A.W. WATTS, *The Way of Zen* (1957); and reference to it in R.C. ZAEHNER, *Mysticism, Sacred and Profane* (1957). The priesthood in Shintō is discussed by D.C. HOLTOM in *The National Faith of Japan* (1938). For studies on the present situation of the priesthood in Christianity, the following works should be consulted: J.D. BENOIT, *Liturgical Renewal* (1958); J.H. SRAWLEY, *The Liturgical Movement* (1954); E.B. KOENKER, *The Liturgical Renaissance in the Roman Catholic Church* (1954); B. LEEMING, *The Vatican Council and Christian Unity* (1966); B.C. BUTLER, *The Theology of Vatican II* (1967); and B.C. PAWLEY (ed.), *The Second Vatican Council* (1967).

(E.O.J.)

Priestley, Joseph

A clergyman, political theorist, educator, and scientist, Joseph Priestley represented the spirit of English dissent that, in the 18th century, produced major advances in liberal thought and experimental science. A prolific writer on many subjects, he exercised a wide influence on politics, religion, and science in Europe and North America. His textbooks, teaching, and classroom demonstrations in science were major innovations in education. Modern Unitarianism owes much to the theological ideas he developed, which stressed man's spiritual autonomy. Known also as one of the discoverers of oxygen, he performed many original experiments with gases, which provided much of the empirical data that contributed to the theoretical foundations of modern chemistry. A man of buoyant optimism, he believed that human progress would be the natural result of political and religious freedom and of the applications of science.

Priestley was born at Birstall Fieldhead, near Leeds, Yorkshire, on March 13, 1733, the oldest of six children. His father, a cloth dresser by trade, employed a few workmen in his home in the finishing treatment of a cloth known as homespun. For several years during his childhood Joseph was often sickly; but, perhaps as compensation, he became an avid learner on his own. Although his parents were Calvinists, they were open to other religious views and wanted him to enter the ministry of the Dissenting church, the diverse congregations of which, such as Presbyterian or Independent, did not conform to the Church of England. To prepare himself, he studied "Hebrew, Chaldee, Syriac, and a little Arabic" with a local Dissenting minister.

With improving health, in 1752 he entered a new Dissenting Academy at Daventry, Northamptonshire, one of the educational institutions that, in the 18th century, was established independently by Nonconformists. These academies offered high-quality education and attracted the best teachers and students. As a theological student, his unorthodox opinions developed even more as he added to the required curriculum his own intensive study of history, philosophy, and science, and a daily translation of 10 folio pages of Greek. In 1755 he left Daventry Academy to become assistant minister to the independent

Early liberal views

Presbyterian congregation in Needham Market, Suffolk; but having become a "furious freethinker," his unorthodox and even heretical opinions gradually lost him the confidence of his orthodox congregation. "After much pain and thought," he renounced the doctrine of the Atonement; moreover, his examination of St. Paul's epistles in the New Testament satisfied him that the Apostle's reasoning "was in many places far from being conclusive." Priestley's intellectual development passed from the Calvinism of his family, through Arianism, with its denial of Christ's divinity, to a rational Unitarianism, with its complete denial of the Trinity. At no stage of this development, however, did he ever waver in the sincerity of his religious convictions or in the genuine piety of his life.

In 1758 Priestley transferred to a more sympathetic congregation in Nantwich, Cheshire, where he opened a day school with 36 students. Becoming interested in science at this time, he provided them with "philosophical instruments," such as an air pump and a static generator for electrical demonstrations. His teaching success led to his appointment in 1761 as a tutor in language and literature at Warrington Academy in Lancashire. Because the universities and learned professions were barred to Dissenters, Priestley developed new courses that were suitable for students preparing for careers in industry and commerce. Textbooks were not available, and these he set himself to write. In 1761 he published the *Rudiments of English Grammar*, a work that was radical in its description of English usage as opposed to prescription on the basis of dead classical languages; it remained in use for 50 years. Ordained a Dissenting minister in 1762 at Warrington, in the same year he married Mary Wilkinson, aged 18, the only daughter of an ironmaster at Bersham in Wales. They had a daughter and three sons.

Contributions to education

Priestley enriched the curriculum with his *Theory of Language and Universal Grammar* (1762), *Chart of Biography* (1765), *Essay on a Course of Liberal Education for Civil and Active Life* (1765), and *Lectures on History and General Policy* (1765, 1788). By emphasizing history, science, and the arts rather than the classics of university curricula, he sought to prepare students for a practical life. His educational activities made Warrington Academy the most distinguished school of its kind in England. The degree of LL.D. was conferred on him by Edinburgh University in 1765.

Priestley became even more interested in science while attending lectures and demonstrations on practical chemistry given in 1763–65 by the surgeon Matthew Turner. Beginning in 1765, Priestley spent a month of every year in London, where he met the leading men of science, including the American statesman and inventor Benjamin Franklin. On the basis of his electrical experiments, Priestley in 1766 was elected to membership in the Royal Society of London. The next year, with Franklin's encouragement and generous loan of the requisite books, he published *The History and Present State of Electricity*, an original work in which he summarized the knowledge of electricity to his time and described his own experiments. He anticipated the inverse square law in electrical attraction with his observation that when he electrified a hollow sphere there was no charge inside. He also discovered that charcoal conducts electricity and noted the relationship between electricity and chemical change, whereby one could cause the other. This work "drew him into a large field of original experiments" that were not in electricity. As he became more interested in chemistry, he reflected on the similarity between the processes of burning and respiration.

Contributions to electricity

In 1767 Priestley was appointed minister of Mill Hill Chapel, in Leeds, Yorkshire, where he had more leisure for writing and experimenting. Although not primarily theoretical in his approach to science, he was able to perceive what research would be most fruitful. He studied gases, or "airs," as they were then called, a subject that led him to perform many significant experiments. The facts he discovered became central in the theoretical development of chemistry. His work on gases began while he was living next to a brewery in Leeds, where he observed "fixed-air" (carbon dioxide) as it effervesced from vats of fermenting liquor. Before his work, only three gases were known: air, carbon dioxide, and hydrogen. With zeal and enthusiasm he discovered 10 new "airs," four in 1767–73 while at Leeds: nitric oxide (nitrous air), nitrogen dioxide (red nitrous vapour), nitrous oxide (diminished nitrous air, later called "laughing gas"), and hydrogen chloride (marine acid air). His success resulted in large part from his ability to design ingenious laboratory apparatus, particularly an improved pneumatic trough, and his skill in its manipulation. Moreover, by collecting gases above mercury in the trough, instead of in water, he was able to isolate, by trial and error, those that were water-soluble.

First gas experiments

Concurrently with these experiments, Priestley developed rapidly as a political theorist. Representative of 18th-century liberal thought was his *Essay on the First Principles of Government, and on the Nature of Political, Civil, and Religious Liberty* (1769), in which he emphasized individualism; he believed that people should have a voice in their government and power over their own actions. The English economist Jeremy Bentham acknowledged that this influential book inspired his phrase "the greatest happiness of the greatest number." Throughout his life Priestley was an ardent believer in human progress and perfectibility.

Political theory

In his gas experiments, Priestley showed that the volume of air decreases by one-fifth during respiration, putrefaction, and combustion. In 1771 he observed that plants restored air used by a live mouse or a burning candle. The next year he discovered that mixing colourless nitric oxide and ordinary air resulted in a diminution in total volume, which he measured in a eudiometer, and a reddish colour (nitrogen dioxide). This was an important new technique; it substituted nitrous oxide for mice as a measure of the "goodness of air" and made possible quantitative tests of chemical changes that involve gases. His summary of this work in the *Philosophical Transactions*, "On Different Kinds of Air" (1772), promptly drew the attention of the French chemist Antoine-Laurent Lavoisier, who proceeded to give a theoretical interpretation of Priestley's observations.

In 1772 Priestley's *History and Present State of Discoveries Relating to Vision, Light, and Colours* was an original contribution to optics. It brought an invitation to join Capt. James Cook's second voyage of exploration (1773–75) as an astronomer, but he was obliged to decline when opposition mounted to his Unitarian views. His technique for "impregnating water with fixed-air" (using sulfuric acid and chalk), however, was suggested for use on the voyage to make drinking water palatable —and also later made possible the soda-water industry.

In December 1772, William Fitzmaurice-Petty, 2nd earl of Shelburne, later 1st marquis of Lansdowne, appointed Priestley as librarian, literary companion, and tutor to his two young sons, on generous terms, with the freedom to preach and write as he wished. He settled at the Shelburne estate at Calne, Wiltshire, in July 1773. The following November he was awarded the Copley Medal by the Royal Society for his experiments on gases, on which he continued to report in a series of volumes entitled *Experiments and Observations on Different Kinds of Air* (6 vol., 1774–86).

Priestley's most famous discovery occurred on August 1, 1774, when he obtained a colourless gas by heating red mercuric oxide (he called it *mercurius calcinatus per se*). Finding that a candle would burn in it "with a remarkably vigorous flame," he called it "dephlogisticated air" because he believed, accepting prevailing theory, that ordinary air became saturated with phlogiston when it could no longer support combustion or life. (Phlogiston was thought to be a material that was transferred during burning and respiration; a unifying idea in 18th-century chemistry, it avoided quantitative considerations but was the reverse of the oxidative interpretation of combustion and respiration established by Lavoisier in 1789.) Priestley was not yet sure, however, that he had discovered a "new species of air." (Oxygen was also discovered by the Swedish chemist and apothecary Carl Wilhelm Scheele,

The discovery of oxygen

probably by 1773.) The following October Priestley accompanied Shelburne on a journey through Belgium, Holland, Germany, and France, where in Paris he informed Lavoisier how he obtained his new "air." The meeting of the two scientists was highly significant for the future of chemistry: Lavoisier required no more than the barest intimation of the success of Priestley's experiments to appreciate their significance. He immediately repeated them and in 1775-80 conducted intensive investigations from which he correctly deduced the elementary nature of oxygen, recognized it as the active "principle" of the atmosphere, interpreted its role in combustion and respiration, and gave it its name (1789). Priestley, however, did not accept all of Lavoisier's views and continued in particular to uphold the phlogiston theory until, in his old age, he became its last champion.

Continuing his studies of the atmosphere, Priestley discovered ammonia (alkaline air), sulfur dioxide (vitriolic acid air), silicon tetrafluoride (fluor acid air), nitrogen (also discovered by Daniel Rutherford in 1772), and a gas later identified as carbon monoxide. Of striking significance was his observation that light was important for plant growth and that green plants gave off "dephlogisticated air" (really oxygen). These observations became fundamental to the systematic work on photosynthesis by the Dutch physician Jan Ingenhousz (begun in 1779) and by the Swiss clergyman-naturalist Jean Senebier during the next two decades. Drawing then on his earlier knowledge of electricity, Priestley decomposed ammonia by sparking with electric sparks and noticed the formation of dew when mixtures of hydrogen and oxygen are exploded.

In 1779, for reasons not entirely clear, Priestley left Shelburne's employ and settled in the industrial town of Birmingham as minister of the New Meeting congregation. He received a small annuity from Shelburne, and friends raised a subscription to defray the expenses of his experiments. The years at Birmingham were the happiest and busiest of his life. He wrote books on religion and theology and published sermons, tracts, and catechisms. In his *History of the Corruptions of Christianity* (1782), he rejected most of the fundamental doctrines of Christianity, including the Trinity, predestination, and the divine verbal inspiration of the Bible and traced them to their historical sources of error. This work aroused another storm of protest. By the time of the French Revolution he had acquired a reputation as the antagonist of all establishments, both political and religious.

Priestley regularly took part in meetings of Birmingham's Lunar Society, which met monthly at the full moon. Flourishing independently of the Royal Society of London, this organization in 1766–91 actively promoted science and its applications to industry and crafts. There, Priestley was the honoured colleague of the naturalist Erasmus Darwin, the pottery manufacturer Josiah Wedgwood, and the inventor of the steam engine, James Watt. Concluding his scientific work, he observed that "calces" (oxides) are changed to the metallic state (reduced) when heated in hydrogen; but he did not notice that water is also produced, and he gave a phlogistic interpretation.

During these years, Priestley was widely known as the defender of the principles of the French Revolution and an ardent advocate of civil and religious liberty. He angered the anti-revolutionary populace by publicly disagreeing with the *Reflections on the Revolution in France* (1790), written by the British statesman Edmund Burke, who opposed the Revolution. On July 14, 1791, on the second anniversary of the fall of the Bastille prison in Paris, an outbreak of mob violence occurred in Birmingham, in the course of which Priestley's house, library, and laboratory were destroyed. He was driven from Birmingham, never to return.

For the next two years Priestley resided at Hackney, near London, where he taught at the New College (Hackney College); but the progress of the French Revolution, the execution of Louis XVI in 1793, and the declaration of war against France the same year excited further rancour against him. His three sons had emigrated to the United States in August 1793; Priestley and his wife followed in April 1794, taking up residence in Northumberland, Pennsylvania. He was welcomed by various learned societies, including the American Philosophical Society, and offered the chair of chemistry at the University of Pennsylvania; but he refused to accept any public office. Move to America

During his years in America, Priestley continued his literary and religious activities, but he was hampered in his scientific work by the limitations of communicating with his friends in England. He published in the United States a four-volume *History of the Christian Church to the Present Times.* He was a friend and correspondent of two leaders of the American Revolution, John Adams and Thomas Jefferson, both of whom held views on religion similar to his own. His *Doctrines of Heathen Philosophy compared with those of Revelation* appeared after his death. It had been written at the suggestion of Jefferson, who felt, however, that "it did not do justice to the undertaking," adding that Priestley "felt himself pressed by the hand of death." He died at Northumberland on February 6, 1804.

BIBLIOGRAPHY. The best book-length biography, which does justice to all aspects of Priestley's work, is F.W. GIBBS, *Joseph Priestley: Adventurer in Science and Champion of Truth* (1965). A complete bibliography of Priestley's published writings is R.E. CROOK, *A Bibliography of Joseph Priestley, 1733–1804* (1966). By far the best account of Priestley's discoveries in chemistry, with an extensive list of references to the Priestley literature, is contained in J.R. PARTINGTON, *A History of Chemistry*, vol. 3, pp. 237–301 (1962). A more popular description is K.S. DAVIS, "Priestley the Experimenter," in *The Cautionary Scientists* (1966). A sympathetic biography, which stresses the significance of the socio-economic background of Priestley's England is found in J.G. CROWTHER, *Scientists of the Industrial Revolution* (1962), The relation of Priestley's religious philosophy to the literary and philosophical background is described in BASIL WILLEY, *The Eighteenth Century Background* (1940). For an admirable brief account of Priestley's theology and writings on religious subjects, see CAROLINE ROBBINS, "Honest Heretic: Joseph Priestley in America," *Proc. Am. Phil. Soc.*, 106:60–76 (1962). Older but still useful essays are PHILIP HARTOG, "The Newer Views of Priestley and Lavoisier," *Ann. Sci.*, 5:1–56 (1941); and R.E. SCHOFIELD, "The Scientific Background of Joseph Priestley," *ibid.*, 13:148–163 (1957).

(S.R.)

Primates

The mammalian order Primates, which includes the prosimians, monkeys, apes, and man, has long excited human interest, because it contains man's closest relatives. The obvious similarity of monkeys and apes to man was noted long before any genetic relationship was postulated. Recognizing that monkeys possess intelligence substantially exceeding that of other familiar mammals, man has valued these animals as companions and pets for centuries. In recent decades, in addition to becoming increasingly popular as pets, nonhuman primates have served as substitutes for man in situations, such as experimental medicine and space science, which require near-human biological reactions.

The two main groups of the order Primates are the prosimians (lemurs, lorises, and tarsiers) and the anthropoids (monkeys, apes, and man). As classified at present, the suborder Prosimii consists of six families; the Tupaiidae (tree shrews, though these are considered by some authorities to belong in the order Insectivora, a view elaborated in INSECTIVORA), Lemuridae (lemurs), Daubentoniidae (aye-aye), Lorisidae (galagos and lorises), Tarsiidae (tarsiers). Collectively, the prosimian suborder is frequently referred to as lower primates.

The suborder Anthropoidea, sometimes called "higher" primates, also comprises six families: Callitrichidae (marmosets and tamarins), Cebidae (South American monkeys other than marmosets), Cercopithecidae (African and Asian monkeys), Hylobatidae (lesser apes: siamangs and gibbons), Pongidae (great apes: orangutan [orangutan], gorilla, and chimpanzee), and Hominidae (men, living and extinct). The two geographically separated stocks, the Old World and the New World monkeys, are often referred to as catarrhines and

platyrrhines, respectively, terms which derive from the shape of the nose (see below, *Form and function*).

I. General considerations

HISTORICAL BACKGROUND OF PRIMATE STUDIES

The inclusion of man in the order provides the rationale for the vigour with which this group has been studied by scientists since the time of Galen of Pergamum. Aristotle and Hippocrates, in the 3rd and 4th centuries BC, recognized the similarity of man and apes, but it was Galen who demonstrated the truth of this kinship by dissection. He wrote, "the ape is likest to man in viscera, muscles, arteries, veins, nerves and in the form of bones." It should be noted that Galen was in fact referring to monkeys, not the true apes, which were unknown to western man until the 15th century. None of these early scientists saw any evolutionary significance in the similarity of man and "apes," a correspondence that they regarded as purely coincidental. An inkling of the truth of man's relationship with primates must have penetrated the mind of Albertus, a leading naturalist of the Middle Ages, who produced a classification of animal life in his book *De Animalibus*. Albertus' classification, which placed man between "apes" (monkeys) on the one hand and "animals" on the other, provides the first whiff of the "missing-link" concept, which later was to befog the issue of man's place in nature.

The Dark Ages were aptly named as far as knowledge of primates is concerned. The first evidence of a renaissance of interest was in the time of Vesalius, the great Belgian anatomist of the 16th century, who published a comparative anatomy of man and "apes" in order to confound the precepts of Galen. He did not succeed in disproving Galen's assertion that "apes were likest to man" but, unwittingly, he succeeded in stirring up an interest in the biology of primates that has never since flagged. The first true ape studied as a scientific specimen was a chimpanzee dissected by Edward Tyson, an English anatomist, in 1699. Tyson's specimen, which he called the "Orang-Outang, sive Homo Sylvestris," is to this day housed in the British Museum (Natural History), mounted in a standing position that reflects Tyson's belief that he had discovered the pygmy, a race of humans known since the time of the ancient Greeks. Tyson wrote of his "pygmie" that it was "no man, nor yet a common ape but a sort of animal between both." It never occurred to Tyson or his contemporaries, who believed that all animals had been created independently in their current image, that man, apes, and monkeys were connected by common evolutionary descent. In 1758, Linnaeus—the father of animal and plant classification—added the lemurs and bats to the monkeys, apes, and man and called the whole assemblage the Primates. It is to be noted, however, that Linnaeus was sufficiently perceptive to see that man was a primate. His conclusion was regarded as a grave blow to human dignity, and it was followed by new classifications such as that of Blumenbach in 1776, placing man in a separate order. Man was not again considered part of the primate order until a century later when the English anatomist St. George Mivart in the climate of post-Darwinian thought published his classification of primates.

Discovery of the first true ape

The first evolutionist was a French scholar of the late 18th century, the Chevalier de Lamarck, who saw animal life as an uninterrupted continuity in which old species were transformed into new species in a sequence of increasing complexity and perfection. In 1821 Baron Georges Cuvier, a rabid anti-evolutionist, had the historic distinction of describing *Adapis*, the first fossil primate genus ever recognized. Fossils such as *Adapis*, Cuvier believed, were the remains of animals destroyed by past catastrophes (floods, earthquakes, etc.) and living animals were new stocks divinely created to fill the vacuum, a view consistent with the widely held notion that species were immutable. During the early 19th century a number of geologists and biologists questioned the doctrine of immutability, but it was not until 1859, with the publication of Charles Darwin's "On the Origin of Species by Means of Natural Selection," that positive evidence was provided, along with a sound alternative theory. The Darwinian contention that not only had man evolved, but he had evolved from a simian ancestor resulted in acrimonious debate among scientists, theologians, philosophers, and laymen. As influential zoologists and anatomists rose to support Darwin, the truth of man's primate consanguinity began to be accepted, if not actually relished.



Figure 1: Representative prosimians.
Drawing by R. Keane

Today, few scientists deny that man and the lower primate belong in the same order; in fact, much current research is directed toward closing the apparent gap between the highest of the nonhuman primates, the chimpanzee and gorilla, and man.

In times past, the public image of primates was largely dictated by prevailing religious beliefs. In Asian countries, where primates abound, monkeys have for a long

Religious views toward primates

time been regarded with various degrees of deference that, among Hindus in India, for instance, amounts to worship. In Europe and North America, where monkeys and apes are totally absent, no religious sect has attached divine significance to them and, in fact, the reverse has been the case, monkeys at various times having been regarded as the personification of evil and depravity, familiars of the devil. This image, however, is fading as a result of enlightened instruction in schools and the advances in naturalistic presentation of primates in zoos. The nonhuman primate is becoming generally accepted by laymen as an animal of peculiar interest to man with many amusing and endearing qualities.

The qualities of primates, however, are less endearing to farmers and agriculturists in certain parts of the world. In South Africa, the chacma baboon (*Papio ursinus*) competes with domestic sheep for grazing lands and is an occasional predator of lambs; in West and Central Africa native crops are subject to daily assaults by forest-living monkeys; and, in India, macaques, which have been accorded a semi-sacred status, live alongside man in towns and villages and are parasitic upon him for their food and shelter.

Scientific interest in nonhuman primates, their structure at all levels and their way of life, is currently in the ascendancy. Their value as research animals increases year by year, so that at present over a quarter of a million wild monkeys are consumed annually by laboratories in the study of human diseases, in the production of vaccines, in the experiments of organ transplantation, in the testing of drugs and, even, for the clinical trials of new cosmetics. Clearly their scientific usefulness raises important problems of conservation of primate stocks in the wild. Other disciplines whose researches depend upon observations and upon experimentation with nonhuman primates include those of endocrinology, neurology, psychology, and sociology. Much is being learned as a result of such studies that is of great significance for man and his future betterment.

SIZE RANGE AND ADAPTIVE DIVERSITY

Size and weight

Members of the order Primates show a remarkable range of size and adaptive diversity. The smallest primate (in linear dimensions) is probably the dwarf or mouse lemur (*Microcebus murinus*) of Madagascar, which measures approximately 13 centimetres (about five inches) in head and body length; the most massive is certainly the gorilla (*Gorilla gorilla*), whose head and torso length may reach 100 centimetres (39 inches; the comparable dimension in man may exceed this figure, but the gorilla is considerably more robust). In terms of body weight, the range is even more striking. The pen-tailed tree shrew (*Ptilocercus lowii*) weighs on average 46 grams (about 1½ ounces); the gorilla's weight may be more than 3,000 times as great, varying from 140 to 180 kilograms (about 300–400 pounds).

In terms of habitat, primates are tropical animals and occupy two major vegetational zones: the tropical forest and woodland–grassland vegetational complexes. Each of these zones has produced in its resident primates the appropriate adaptations, which are not remarkable for their diversity. There is perhaps more diversity of bodily form to be seen amongst forest-living species than among savanna inhabitants. One of the explanations of this difference is that it is the precise pattern of locomotion rather than the simple matter of habitat that governs overt bodily adaptations. Within the forest there are a number of ways of moving about. An animal can live on the forest floor or in the canopy, for instance, and within the canopy it can move in three particular ways: (a) by leaping—a function principally dictated by the hindlimbs; (b) by arm swinging (brachiation)—a function particularly of the forelimbs; (c) by quadrupedalism—a function equally divided between the forelimbs and the hindlimbs. On the savanna, or in the woodland-savanna biome, which substantially demands adaptations for ground-living locomotion rather than those of tree-living, the possibilities are limited. If bipedal man is discounted, there is a single pattern of ground-living locomotion, which is called quadrupedalism. Within this category there are at least two variations on the theme: (a) knuckle-walking quadrupedalism, and (b) digitigrade quadrupedalism. The former gait is characteristic of the African apes (chimpanzee and gorilla), and the latter of baboons and macaques, who walk on the flats of their fingers. After man, the Cercopithecinae are the most successful colonizers of nonarboreal habitats.

The structural adaptations of primates resulting from locomotor differences are considered below in more detail, but they do not prove to be very extensive. Primates are a homogeneous group in terms of morphology, and it is only in the realm of behaviour that differences between primate taxa are clearly discriminant. It can be said that the most successful primates (judged in terms of the usual criteria of population numbers and territorial spread) are those that have departed least from the ancestral pattern of structure, but furthest from the ancestral pattern of behaviour. "Manners makyth man" is true in the widest sense of the word; in the same sense, manners delineate primate species.

DISTRIBUTION AND ABUNDANCE

The nonhuman primates have a wide distribution throughout the tropical latitudes of Africa, India, Southeast Asia, and South America. Within this tropical belt, which lies between latitudes 25° N and 30° S, they have a considerable altitudinal range. In Ethiopia the gelada baboon (*Theropithecus*) is found living at altitudes up to 5,000 metres (16,000 feet). Mountain gorillas of the Virunga Volcanoes are known to travel across mountain passes at altitudes of more than 4,200 metres (13,800 feet) when travelling from one high valley to another. The howler monkeys of Venezuela (*Alouatta seniculus*) live at 2,500 metres (8,000 feet) in the Cordillera da Merida and in northern Colombia the dourocouli (*Aotus*) is found in the tropical montane forests of the Cordillera Central.

Temperate zone species

A few primate species occur naturally in temperate latitudes. The rhesus monkey (*Macaca mulatta*), for example, is found near Peking, and Japanese macaques (*Macaca fuscata*) are to be found on the Shimokita Peninsula, a hatchet-shaped promontory of the Japanese island of Honshu on latitude 41° N, where snow covers the ground for four or five months of the year. Macaques are particularly hardy animals and can withstand cold climates in the wild and captivity extremely well. The limiting factor to the occupation of temperate zones by monkeys is the winter food supply, combined with the short daylight hours available for foraging.

Primates in general are still plentiful in the wild; for the majority of common species there are few serious conservation problems at present. A number of species, particularly the orang-utan, the gorilla, some of the Madagascan lemurs, and some South American species, such as the golden marmoset, however, are in serious danger of extinction unless their habitats can be preserved in perpetuity and predatory habits of man kept under control.

II. Natural history

REPRODUCTION AND LIFE CYCLE

The stages of the life cycle of primates vary considerably in duration. Among the most primitive members of the group, these stages are broadly comparable to those of other mammals of similar size. Higher in the phylogenetic scale they are substantially extended. The life span of a lemur is about 15 years, of a monkey 25–30 years, of a chimpanzee 40 years, and of man 70–75 years. The greatest differential between man and prosimians, indeed between the prosiminans and apes and between apes and man, is in the duration of the infant and juvenile stages combined. The least differences occur in gestation period which, despite the general belief, can not be consistently correlated with adult body size. Gibbons, which weigh considerably less than macaques, have a 20 percent longer gestation period.

The clear trend toward prolongation of the period of juvenile and adolescent life is probably to be associated with the corresponding trend towards a progressive elabo-

ration of the brain. The extended period of adolescence means that the young remain under adult (primarily maternal) surveillance for a long period, during which time the juvenile acquires, by example from its mother and peers, the knowledge that will allow it to become properly integrated as a fully adult member of a complicated social system. One might expect, therefore, a close correlation between the period of adolescence, the brain size, and the complexity of the social system; and, insofar as the latter factor can be assessed, this appears to be the case.

Breeding periods. The reproductive events in the primate calendar are copulation, gestation, birth, and lactation. Owing to the long duration of the gestation period, these phases occupy the female primate (among higher primates anyway) for a full year or more; then the cycle starts again. The female does not usually come into estrus (physiologically receptive period) until the infant of the previous pregnancy is weaned.

The reproductive patterns of prosimians differ from those of anthropoids. Prosimians show one or more discrete breeding seasons during the year, during which time they may undergo more than one reproductive estrous cycle. The breeding seasons are separated by periods of anestrus, which in galagos are accompanied by changes in the skin of the vulva (the external organ of the genitalia), which closes over, completely sealing the vagina. In the Sudan estrus occurs only twice yearly in the African bush baby (*Galago senegalensis*), during December and August; but in captivity, breeding seasons may occur at any period in the year. In the wild, birth seasons are closely correlated with the prevailing climate, but in captivity under uniformly equable laboratory conditions, this consideration does not apply. In its native Madagascar the ring-tailed lemur (*Lemur catta*) has only a single breeding season during the year; conception occurring in autumn (April) and births taking place in late winter (August and September); but in zoos in the Northern Hemisphere a seasonal inversion of the birth period occurs in which it shifts to late spring and early summer. These examples suffice to indicate the influence of environmental factors on the timing of the birth seasons.

Reproductive cycles in the Anthropoidea continue uninterrupted throughout the year, though seasonality in births is characteristic of primate species living either outside the equatorial belt (5° north and south of the Equator) or at high altitudes in equatorial regions, where dry seasons and seasonal food shortages occur. Seasonality of births in macaques (*Macaca* species) has been documented in Japan, on Cayo Santiago in the Caribbean (where an introduced population thrives under semi-natural conditions), and in India. Observations on langurs in India and Ceylon, on geladas in Ethiopia, and on patas monkeys in Uganda have also demonstrated seasonality in areas with well-marked wet and dry seasons. Anthropoids within the equatorial belt display birth peaks rather than birth seasons. A birth peak is a period of the year in which a high proportion of births, but not by any means all, are concentrated. Equatorial primates such as guenons, colobus monkeys, howlers, gibbons, chimpanzees, and gorillas might be expected to show a pattern of births uniformly distributed throughout the year, but population samples are as yet too small to make this assumption. In captivity, under both laboratory and zoo conditions, anthropoids breed throughout the year with little evidence of seasonality. Even in man there is evidence of high birth peaks. In Europe, the highest birth rates are reached in the first half of the year, and in countries in the Southern Hemisphere in the second half. This may, however, be a cultural rather than an ecological phenomenon, for marriages in certain western countries reach a peak in the closing weeks of the fiscal year, a fact which undoubtedly has some repercussions on the birth period.

Distribution of births through the year

Gestation period and parturition. The period during which the growing fetus is protected in the uterus is characterized by a considerable range of variation among primate species, but shows a general trend towards prolongation as one ascends the evolutionary scale. The range of prosimians overlaps that of the anthropoids, which is never shorter than 140 days. Tree shrews, for example, have a gestation period of 41–50 days, lemurs 132–134 days, macaque monkeys 146–186 days, gibbons 210 days, chimpanzees and gorillas 250–290 days, and man (on the average) 267 days. Even small primates like bushbabies have a gestation period considerably longer than those of nonprimate mammals of equivalent size, reflecting the increased complexity and differentiation of primate structure compared with that of nonprimates. Although in primates there is a general trend towards evolutionary increase in body size, there is no absolute correlation between body size and the duration of the gestation period. Marmosets, for example, are considerably smaller than spider monkeys and howler monkeys but have a slightly longer pregnancy (howler monkeys 139 days; marmosets 140–150 days).

An extraordinary and somewhat inexplicable difference exists between the dimensions of the pelvic cavity and the dimensions of the head of the infant at birth in monkeys and man on the one hand, apes on the other. The head of the infant ape is considerably smaller than the pelvic cavity, so that birth occurs easily and without prolonged labor. When the head of the infant monkey engages in the pelvis the fit is exact, and labour may be a prolonged and difficult affair as it is in man, in whom similar conditions prevail. Human parturition, however, is generally a much more extended process than that of monkeys. Like the human infant, the monkey is born head first. Twin births are as rare in Old World monkeys and apes as they are in man, but certain New World monkeys (*e.g.*, marmosets) habitually produce twins.

Condition of the newborn primate

The degrees of maturation and mother dependency at birth are obviously closely related phenomena. Newborn primate infants are neither as helpless as kittens, puppies, or rats, nor as precocial as the newborn gazelles, horses, and other savanna-living animals. With a few exceptions, primate young are born with their eyes open and are fully furred. Exceptions are the lowly tree shrews and mouse lemurs, which carry their young in their mouths in the manner of many mammals which bear altricial infants. Primate life being peripatetic, it is axiomatic that the infants must be able to cling to the mother's fur. The young of most higher primates have fully prehensile hands and feet at birth and are thus able to grasp the maternal fur without assistance; only man, chimpanzee, and gorilla need to support their newborn infants. In man this stage persists throughout the nursing period (six months or more), but in gorillas it lasts only four weeks and in chimpanzees for an even shorter period.

It seems likely that the difference between the African apes and man in respect of postnatal grasping ability is related to the acquisition in man of bipedal walking. One of the anatomical correlates of the human gait is the loss of the grasping function of the big toe, which is aligned in parallel with the remaining digits. Such an arrangement precludes the use of the foot as a grasping extremity. The human infant and, to a lesser degree, the gorilla infant must depend largely on its grasping hands to support itself unaided. The fact that man is habitually bipedal and that, consequently his hands are freed from locomotor chores, may also be a contributory factor; the human mother can move about and at the same time continue to support her infant. Selection for postnatal grasping, therefore, has not had the high survival value that it has in nonhuman primates in which the survival of the infant depends on its ability to hold on tightly. It is a well-known fact, of course, that newborn human infants can support their own weight, for short periods, by means of their grasping hands. Clearly adaptations for survival are not wholly lacking in the human species. Perhaps there are involved cultural factors which have the effect of suppressing natural selection for early infant grasping ability. The first may be that the social evolution of a division of labour between the sexes and a fixed home base has allowed the mother to park her infant with other members of the family as baby-sitters; the second that in more peripatetic communities, the in-

vention of infant-carrying devices, such as the papoose technique of North American Indians, has made it unnecessary for the infant to support itself. Whatever the biological or cultural reasons, the human infant is less precocial, than the young of any other primates.

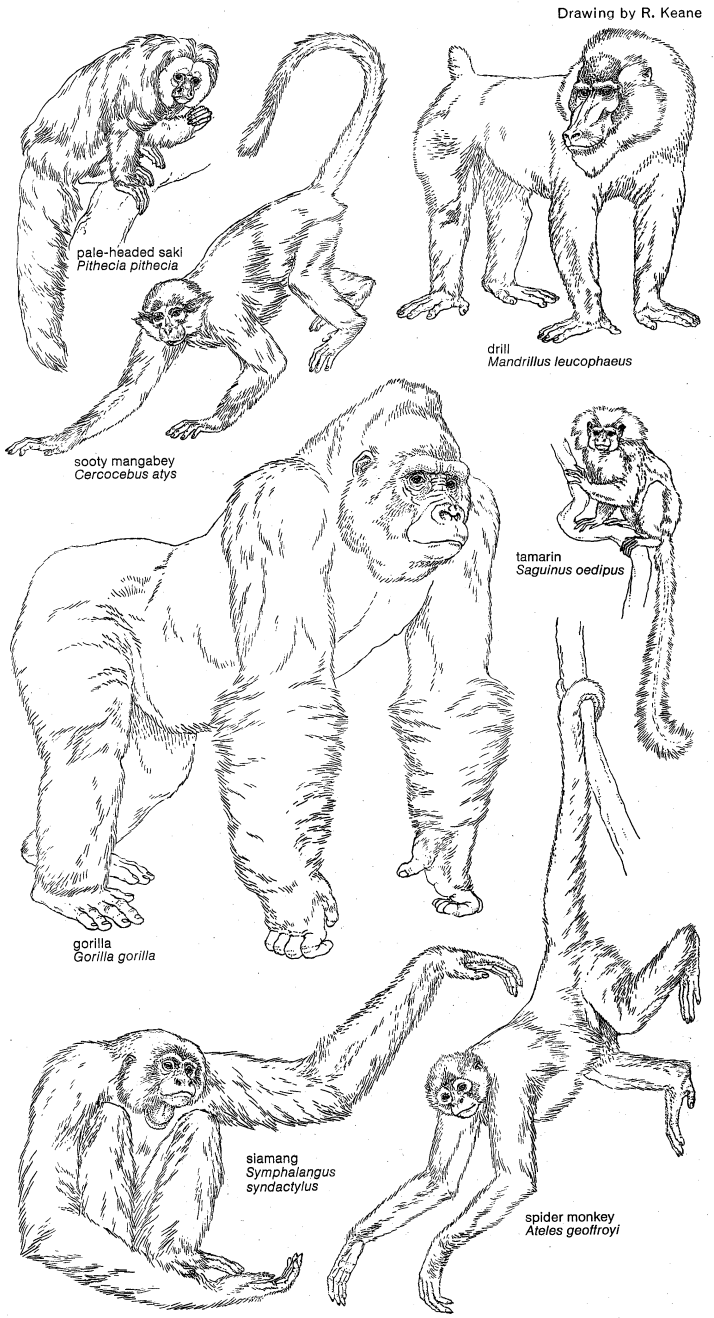


Figure 2: Body plans of representative anthropoids.

Once the primate infant learns to support itself by standing on its own two (or four) feet, the physical phase of dependency is over; the next phase of psychological dependency lasts much longer. Man, metaphorically, is tied to his mother's apron strings for much longer periods than are the nonhuman primates. The reasons for this are discussed below. According to Adolph Schultz, the anthropologist whose comparative anatomical studies have illuminated knowledge of nonhuman primates during the last 40 years, the juvenile period of psychological maternal dependency is 2½ years in lemurs, 6 years in monkeys, 7–8 years in apes, and 14 years in man.

GROWTH AND LONGEVITY

The characteristic pattern of weight-growth curves of man is also shared by higher nonhuman primates such as chimpanzees (*Pan troglodytes*) and rhesus monkeys (*Macaca mulatta*). There seems little reason to doubt that similar curves will be found to exist in other higher primates not yet studied. It has been suggested that the growth curves of primates are closely akin to the basic curves of mammals in general.

The trend toward the prolongation of postnatal life as a characteristic of primates affects all postnatal life periods including infantile, juvenile, adult, and the period of senescence. The period which shows the least change in apes and man is that of gestation. The relative stability of the gestation period is a clear indication that, developmentally speaking, in the degree of structural complexity and cellular differentiation, there is little difference between apes and man. The differences that undoubtedly exist can probably be attributed to cultural factors. There is no reason to suppose that apes are not built to last equally as long as man; it is simply that the natural hazards of gorilla or chimpanzee life are not buffered by the blessings of civilization, except under conditions of captivity, in which case life spans of 40 years and more are known. The potential life span of the chimpanzee has been estimated at 60 years.

Average and potential life span

The characteristic growth spurts of human infants in weight and height also occur in nonhuman primates but start earlier in the postnatal period and are of shorter duration. Primates differ from most nonprimate mammals by virtue of a delayed puberty in both sexes until growth is nearly complete; in male macaque monkeys, as in man, the growth spurt precedes the onset of puberty. To postpone this latter event is of selective advantage because it ensures that the juvenile males in a primate society are not sexually competitive with adult males until they are large enough and strong enough to contribute fully to the responsibilities of the society in which they live by taking part, for instance, in defense against predators.

LOCOMOTION

Primate locomotion, being an aspect of behaviour which arises out of anatomical structure, shows much of the conservativeness and the opportunism that characterizes the primates. Primates with remarkably few changes in their skeletons and musculature have adopted a bewildering variety of locomotor patterns. The "natural" habitat of primates—in the historical sense—is the canopy of the forest. Although many primates have adopted the ground as their principal foraging area during the day, given the opportunity, they return to the trees to sleep at night. Trees provide cover from the climate and protection from predators; they are of course also a source of food. Only the gelada (*Theropithecus gelada*), the hamadryas baboon (*Papio hamadryas*) of the mountainous regions of Ethiopia, and the chacma baboon (*P. ursinus*), which lives on the rocky coast of the Cape of Good Hope, South Africa, are ground sleepers; even these animals seek the protection of the cliffs and rocky precipices of their habitats at night. No primate sleeps totally unprotected; as a consequence of their relative immunity from predation, primates are heavy sleepers.

The essential arboreality of primates has guaranteed the relative uniformity of the locomotor apparatus. Even man, who has long since abandoned the trees as his principal lodging place, has only partially lost the physical adaptations for tree climbing; his hands remain in the arboreal mold. Only his feet have lost their primitive prehensility in adapting to bipedal walking. Primate locomotion can be classified on behavioral grounds into four major types (see Table). Within these major categories there are a number of subtypes, and within these subtypes there are an infinite number of variations between species and, by virtue of individual variability, within species. The differences between the four major categories lie principally in the degree to which the forelimbs and hindlimbs are used to climb, swing, jump, and run.

Locomotor types

Vertical clinging and leaping, for instance, is primarily a function of the hindlimbs, as is bipedalism, whereas brachiation is performed exclusively with the forelimbs. Quadrupedalism involves both forelimbs and hindlimbs,

Locomotor Classification

category	subtype	activity	primate genera
Vertical clinging and leaping		leaping in trees and hopping on the ground	*Avahi, Galago, Hapalemur, Lepilemur, Propithecus, Indri, Tarsius*
Quadrupedalism	slow climbing type	cautious climbing—no leaping or branch running	*Arctocebus, Loris, Nycticebus, Perodicticus*
	branch running and walking type	climbing, springing, branch running and jumping	*Aotus, Cacajao, Callicebus, Callimico, Callithrix, Cebuella, Cebus, Cercopithecus, Cheirogaleus, Chiropotes, Lemur, Leontideus, Phaner, Pithecia, Saguinus, Saimiri, Tupaia*
	ground running and walking type	climbing, ground running	*Macaca, Mandrillus, Papio, Theropithecus, Erythrocebus*
	New World semi-brachiation type	arm swinging with use of prehensile tail; little leaping	*Alouatta, Ateles, Brachyteles, Lagothrix*
	Old World semi-brachiation type	arm swinging and leaping	*Colobus, Nasalis, Presbytis, Pygathrix, Rhinopithecus, Simias*
Brachiation	true brachiation	gibbon type of brachiation	*Hylobates, Symphalangus*
	modified brachiation	chimpanzee and orangutan type of brachiation	*Gorilla, Pan, Pongo*
Bipedalism		striding	*Homo*

Source: From J.R. and P.H. Napier, *A Handbook of Living Primates.*

of course, although not to an equal extent. Some quadrupeds are hindlimb dominated; in others, the forelimb and the hindlimb are equally important. The hindlimb-dominated primates, such as the langurs and colobus monkeys, employ a large element of leaping in their movements, a less notable feature of the more generalized quadrupeds such as the species of the genus *Cercopithecus*. The quadrupedal category is inevitably somewhat of a grab bag, and the gaits included in it have not yet been studied critically. One subtype, here designated as slow climbing, differs profoundly from the other subtypes of the category, being somewhat ponderous and devoid of elements of leaping or jumping. The species in this category are all arboreal nocturnal prosimians.

As many authorities who have studied locomotion in free-ranging primate species have pointed out, the classifications of locomotion into categories is a somewhat artificial procedure. A chimpanzee shows a variety of different gaits according to the circumstances of the environment, and brachiation may form only a minor part of its repertoire; this holds true also for the langurs and colobus monkeys, which are placed in the subtype Old World semibrachiation. Although the categories are phrased in behavioral terms, their implications are strictly anatomical. Brachiation is the mode of locomotion for which the animal is specifically adapted; the anatomical correlates of brachiation are quite unmistakable and can be determined equally as well in fossil bones as in living animals. A purely behavioral classification, based on observations in the living, can have little value in paleontological diagnosis. It is quite arguable that although pure arm swinging is rare in Old World semibrachiators, it is nevertheless one of the principal survival mechanisms of the group, an adaptive advantage that in critical circumstances provides these species with an "edge" over their competitors. In some instances it may well be the case that the particular category is of evolutionary significance only. The gorillas, for instance, whose anatomy is that of a brachiator, in fact, seldom brachiate; adult males probably never do. Gorillas are large animals and bulk-food eaters, forced by their dietary demands to obtain their food on the ground; living gorillas have adapted to a terrestrial life by adopting a compromise gait in which they walk quadrupedally bearing their weight on the backs of their knuckles. This is a form of quadrupedalism (inasmuch as all four limbs are employed) but, once again, the anatomy of their bodies indicates that in past time gorillas (or their evolutionary forebears) were smaller animals which lived and moved in trees. The modern gorilla differs widely from its historical counterpart in behaviour but, structurally, its kinship with its arboreal ancestor is overtly apparent.

Changes in climate and geography during the evolutionary history of primates may also have led to structural atavisms in the anatomy of living primates. Many chimpanzees now living in woodland-savanna conditions in Africa, where the trees are widely spaced and generally unsuitable for the classic arm swinging progress of a brachiator, have adopted a largely ground-living life. Like the gorilla, the chimpanzees are also knuckle walkers, but given an environment like that of a zoo with a cage specially designed with a plentitude of overhead bars and ropes, they brachiate freely and frequently. Habitat changes in the past undoubtedly account for many apparent anomalies in the locomotor classification of present-day primates.

The relation of locomotion to habitat changes

When the subject of primate arboreal locomotion is studied in evolutionary terms, through the medium of fossils, it becomes clear that locomotor categories are not discrete, but that they constitute a continuum of change from a hindlimb-dominated gait to a forelimb-dominated one. The best single indicator of gait, one that has the added advantage of being strictly quantitative, is the intermembral index. Briefly, the index is a ratio expressed as percentage of arm length to leg length; an index over 100 indicates relatively long arms. The intermembral index provides a model by means of which the locomotion of an early primate can be inferred by determination of the intermembral index of the fossil skeleton. Animals do not necessarily fall discretely into categories. Species with indices lying between those of clearly recognizable locomotor types represent transitional types, whose style of locomotion manifests features of both of the bracketing categories. Some lemurs have indices that fall between 65–75, but their gait is a combination of vertical clinging and quadrupedalism. The ring-tailed lemur in fact shows an intermediate form of behaviour; it is generally quadrupedal but, given the right environmental milieu, can be observed in vertical clinging and leaping. The South American spider monkeys (*Ateles*), whose index lies between 100–108 show a type of locomotion that contains the elements of both quadrupedalism and brachiation.

The intermembral index

Applying the intermembral index to fossil primates, it has become apparent that the earliest primates living in the Eocene Epoch, from about 54,000,000 to 38,000,000 years ago, moved about in the manner of modern vertical clingers and leapers. Quadrupedalism was well-established during the Miocene Epoch (about 26,000,000 to 17,000,000 years ago) when the two major environmental types of quadrupedal gait—the terrestrial and the arboreal—were established, with indices in the region of 85–100 and 75–85, respectively. Brachiation, associated with a high intermembral index, was established as a way of arboreal life at the end of the Miocene and the beginning of the Pliocene.

The origin of bipedalism, which implies the appearance of manlike creatures, is indeterminate. There is direct evidence of bipedalism extending back 1,500,000 to 2,000,000 years, and certain indirect evidence (see below *Evolution and paleontology*) suggests that bipedalism might have evolved in a modified form some 10,000,000 to 14,000,000 years ago.

Locomotion of primates is an area of primate biology of great interest to scholars. It is not only that locomotion is a basic factor of primate life (primates must move to eat and must eat to live) but it is also an indicator of evolutionary progress. To study the origins of primate locomotion is to study primate evolution. Because of its immediacy for survival, natural selection has acted strongly on locomotor behaviour and the locomotor apparatus; changes in locomotor pattern may thus be said to provide principal milestones on the road to modern primates.

Bipedalism, far from being a unique possession of mankind, as is generally thought, is a basic possession of the order Primates. With one or two exceptions (one of which is the tree shrew, which may not even be a primate), all primates sit upright. Many stand upright without supporting their body weight by their arms, and some actually walk upright for short periods. The view that the possession of uprightness is a solely human attribute is untenable; man is merely one species among the 189 that constitute the order who has exploited the potential of his ancestry to his ineffable advantage.

Walking

Chimpanzees, gorillas and gibbons, macaques, spider monkeys, capuchins, and others are all frequent bipedal walkers. To define man taxonomically as "bipedal" is not enough; to describe him as habitually bipedal is nearer the truth, but habit as such does not leave its mark on fossil bones. Some more precise definition is needed. Man's walk has been described as striding, a mode of locomotion defining a special pattern of behaviour and a special morphology. Striding, in a sense, is the quintessence of bipedalism; it is a means of travelling during which the energy output of the body is reduced to a physiological minimum by the smooth, undulating flow of the progression. It is a complex activity involving the joints and muscles of the whole body in its performance and it is likely that the evolution in the human gait took place gradually over a period of 10,000,000 years or so.

The pattern of locomotion of man's ancestors immediately preceding the acquisition of bipedalism has long been a matter of controversy, and the question has not yet been resolved. The evidence derived from anatomical, physiological, and biochemical studies for the close affinity of chimpanzees and gorillas with man would imply that he evolved from a brachiating ancestry. But the many differences between the apes and man, particularly in the structure of the limbs and teeth, have persuaded the majority of leading anthropologists that the theory of a brachiating origin of man is untenable. The issue is still hotly debated, the more so since current immunological studies of serum proteins are tending to emphasize the recentness of the separation of pongid and hominid stocks (the lines that led, respectively, to apes and man) in phylogeny. An alternative to the "brachiating theory" must inevitably be some sort of "quadrupedal theory." Various authorities have proposed other solutions: semibrachiation (a form of quadrupedalism with emphasis on arm swing), a formal quadrupedal gait, and even a form of locomotion similar to that of *Tarsius* and other clingers and leapers. At the present time, there is insufficient information to elucidate the phylogeny of man's bipedal gait. It can, however, be assumed that it must have involved a large measure of truncal uprightness, a good deal of leaping, and a sizable element of arm swinging. There is no living primate that demonstrates, precisely, this combination of locomotor characteristics, although the arboreal monkeys of the subfamily Colobinae theoretically fit the bill better than most.

ECOLOGY AND DIET

Few primate species are found today in temperate latitudes. The limiting factors appear to be the general climate and the duration of the daylight hours. The higher the latitude, the greater is the fluctuation between summer and winter temperatures and the shorter are the hours of winter daylight. The combination of quiescent growth periods of fruits, leaves, and grasses, a frozen top soil, and a restricted diurnal period all make conditions extremely hard for monkeys, which depend on foraging in the trees, on the ground, or under it for their food supply. No monkey lives, or has ever lived, in the zones of permafrost, but it is probable that in times past they flourished in temperate regions. Without the harassment of man, monkeys no doubt could have survived in the temperate zones; their foraging, albeit condensed into a temperate-zone day of seven hours or so, would have supplied them with the food they required, providing they were free to forage without the threat of human persecution. Dietary adaptability of primates is greater than realized. On the Shimokita Peninsula of the Japanese Island of Honshu and in the Shiga Mountains of the same Island, Japanese macaques survive from late November until April on incredibly meagre rations. The diet of these monkeys, during the months when snow covers the ground and when the last remaining summer seeds and leaves have vanished, is bark and the fleshy cambium layer that lies beneath it. The Japanese macaques, cold and hungry, ride out the worst that winter, at the latitude of 41° N offers. Near Moscow, several species of African primates have been kept for several years in outdoor facilities in which winter temperatures go well below −18° C (0° F). It is true that they are well provisioned but, nevertheless, it is almost incredible that a tropical species should be sufficiently adaptable and possess adequately effective homeostatic mechanisms to survive in this degree of climatic adversity. Ruggedness of this sort is a primate characteristic, a survival mechanism that operates in the worst conditions, and one which provides a clue to the real reason for the primacy of the primates among mammals.

Resistance to cold

Diet. The diet of primates is a factor of their ecology, that, during their evolution, has clearly played an important role in their dispersion and adaptive radiation as well as in the development of the teeth, jaws, and digestive system. Diet is also closely related to locomotor pattern and to body size.

The principal food substances taken by primates may be divided into vegetable (fruits, flowers, leaves, nuts, barks, pith, seeds, grasses, stems, roots, and tubers) and animal (birds, birds' eggs, lizards, small rodents and bats, insects, frogs, and crustacea). The flesh of larger mammals (including primates) is not listed as an important item of nonhuman primate diet, although it is taken by certain species such as baboons and chimpanzees in special circumstances that are not yet fully understood.

While in many mammalian groups, diet is selective and specific to the order, among primates it is difficult to establish any hard and fast rules. Although there are decided preferences for certain food items, catholicity is more characteristic than specificity. Generally speaking, primates are omnivorous, as the physiology of their digestive system attests. Relatively few examples of dietary specialization are to be found. The so-called leaf-eating monkeys, a sobriquet that embraces the whole of the subfamily Colobinae, are by no means exclusively leaf eaters and according to season include flowers, fruit, and seeds in their diet. The howler monkeys of the New World have a similar dietary preference.

Broadly, however, certain overall dietary preferences are discernible. The leaf-eating langurs have already been mentioned. The apes (other than the gorilla) are substantially fruit eaters; the percentages of fruit in the diet of the gibbon and the three great apes are as follows: gibbon 62 percent, orang 50 percent, chimpanzee 67 percent, and gorilla 2–15 percent.

Many of the smaller nocturnal prosimians such as tarsiers, galagos, dwarf lemurs, sportive lemurs, the aye-aye, slender lorises, and the South American night monkey are substantially insectivorous.

The larger diurnal prosimians (*e.g.*, typical lemurs, the sifaka, and the indri) and the large nocturnal prosimians (the Asian slow loris and the African potto) are more vegetarian, including both fruit and leaves in their menu. It seems apparent that size, rather than activity rhythm, governs the nature of the primate diet. The small marmosets of the South American genus *Callithrix* have exclusively diurnal rhythms and are insectivorous, while the slightly larger, but equally diurnal, tamarins (*Saguinus*) are more omnivorous.

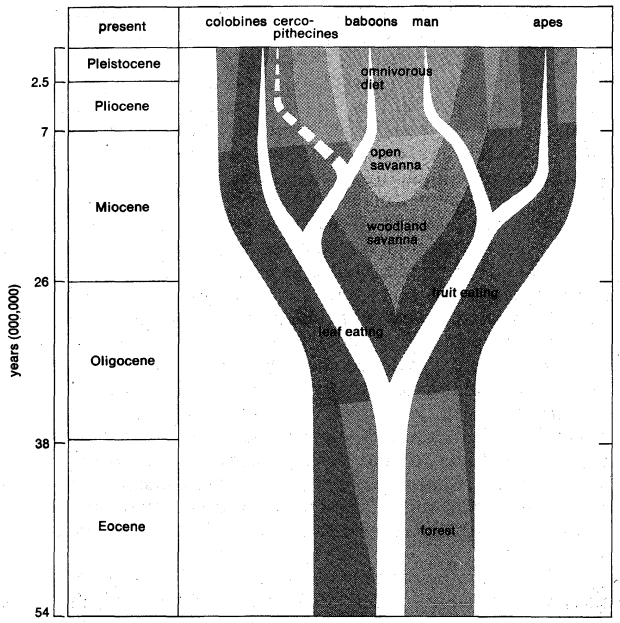


Figure 3: Hypothetical evolution of primate feeding habits in relation to vegetational areas.
From J.R. Napier and P.H. Napier, *A Handbook of Living Primates* (1967); Academic Press

The most widespread monkey populations are those of the subfamily Cercopithecinae. Many ground-living species are included in this group, the members of which are notable for the extreme catholicity of their diet; meat eating as a regular pattern, however, has not been reported among them.

Size in relation to food habits. In evolutionary terms, increase in size has probably played a large part in determining the direction of primate evolution. Fossil evidence of early primates of the Paleocene (about 65,000,000 years ago) suggests that these were small, forest-living creatures, not yet specialized as tree climbers. Their bodily form suggests that they possessed the build and the gait of squirrels. Their molar teeth bore high, pointed cusps but were neither as tall nor as pointed as those of the insectivore-like ancestors of the preceding Cretaceous epoch, whose molars were ideally adapted for cracking the hard, chitinous exoskeletons of insects. This fact suggests that the reduction of the molar cusps was associated with the adoption of a fruit-eating habit. Although this has some validity as a generalization, it should not be taken too literally. All primates include insects in their diet and there are many almost exclusively insectivorous forms, which have reduced the height and acuity of their molar cusps. Increasing body size, a trend which is clearly apparent throughout primate evolution, would have been associated with the adoption of supplementary sources of food. Insect eating is a laborious way for a large animal to obtain nourishment unless it happens to possess the extensible tongue of those mammalian vacuum cleaners, the anteaters.

The shift from insect to fruit eating

An increase in size and the gradual addition of bulk foods to the diet would, in turn, have affected the habitat and the pattern of locomotion. Fruit eating is more easily accomplished in trees than on the ground and this growing preference amongst primates may have been accompanied by the tree-climbing habit. Once such a habit was established, natural selection would have operated to espouse the suitable physical adaptations which would facilitate such a behaviour. Suitable adaptations in this case would have been the facility to climb, leap, and balance in trees.

It is noteworthy that during evolution the development of a prehensile foot preceded that of a prehensile hand. Vertical clinging primates such as the tarsiers or small squirrel-like quadrupeds like the marmosets and tamarins —all of which have prehensile feet, but not prehensile hands—by remaining small, have never been subject to the same evolutionary pressures that have impinged on larger primates. A large arboreal primate without prehensile hands is at a considerable disadvantage in moving about in the canopy of trees, but a small one suffers little disadvantage. Amid the large and firm branches, size is no particular hazard, but at the periphery of the crown, where the fruit is most abundant and the branches are slender and flexible, the risk of falling is increased. It is, therefore, likely that the combination of an increase in body size, associated with the inevitable shift towards a bulk diet, led first to the evolution of a grasping hand, then to the appearance of a prehensile hand, and finally to an opposable thumb. Four prehensile extremities are more effective than two in defying gravity.

Such adaptations of the forelimbs would have had the effect of equalizing the role of the limbs. The limbs of vertical clingers are functionally disparate, the lower pair being dominantly propulsive, the upper, secondary and purely supportive. The limbs of quadrupeds, on the other hand, are more homogeneous, both pairs having a propulsive function during running. Thus, it would seem that the transition in locomotor grade between vertical clinging and leaping and quadrupedalism came about as adaptation to increased body size. Size, diet, ecology, locomotion, and anatomical structure provide a constellation of causes and effects that are critical factors in the evolution of the primates.

Effect of forest stratification. An aspect of primate ecology that has not yet been fully investigated is the effect of the stratification of forests on the adaptations (particularly the locomotor adaptations) and on the speciation of primates. In recent years it has become apparent that forests are not simply a random consociation of trees but are an ordered system of plants under the control of climatic, photobiotic, and edaphic (soil) factors.

The chief physiognomic features of rain forests, the ancestral home of primates and the principal habitat (in numerical terms) of nonhuman primates today, are the evergreen broad-leaved trees that, collectively, form a closed canopy, so opaque to sunlight that the forest floor is in perpetual twilight. Epiphytes and thick-stemmed lianas drape the trees, linking one crown to another and providing aerial pathways for monkeys to pass from tree to tree through a continuum of interlacing branches, a three-dimensional maze that provides home, restaurant, shoping districts, and highways for primates. Three strata of the canopy of rain forests are usually recognized by botanists: an understory, a middle story, and an upper story. The understory is often "closed," the crowns of the constituent trees overlapping one another to form a dense, continuous, horizontal layer. The middle story is characterized by trees that are in lateral contact but do not overlap; and the highest story by tall trees, some 50 metres (160 feet) or more, which form a discontinuous layer of umbrella-shaped crowns. The occasional "emergent" forest giant may tower above the highest layer of the canopy. There is some evidence, much of it conflicting, that zonation of forest primates occurs within the forest canopy. The stratification of forest is extremely variable and the number of layers tends to diminish from three to two in secondary forest, dry deciduous forest, and montane forest; and from two to one as temperate zone, tropical woodland, or montane woodland supervenes.

Tropical grasslands, or savannas, are also the homes of primates in Africa and Asia; no savanna-living primates exist in South America. Tropical grasslands comprise a mixture of trees and grasses. The proportion of trees to grass varies directly with the rainfall; areas of high seasonal rainfall support single story woodlands of tall trees while lush grasses form the ground vegetation; but where rainfall is both seasonal and low, the trees consist of stubby xerophilous (dry-loving) shrubs and short, tussocky grasses. The principal primate fauna of the savanna biome are the ground-living species; in Africa, the grivet (*Cercopithecus aethiops*), baboons (*Papio* species), and the patas monkey (*Erythrocebus patas*); and in Asia the macaques (*Macaca* species) and the Hanumān langur (*Presbytis entellus*).

Tropical montane forests or tropical rain forests at high altitude also abound in primates in Africa, Asia, and South America. In equatorial Africa certain primate species have colonized the montane-savanna regions, or moorlands, where the rugged mountainous terrain and seasonal food scarcity support herds of geladas and hamadryas baboons. These high mountaineers of Africa have no ecological counterparts in Asia or South America.

As a measure of the physiological adaptability of primates it may be noted that various populations have at different times been subjected to relocation in alien environments by man. Some of these expatriates are the green monkeys (*Cercopithecus sabaeus*) of St. Kitts in the West Indies, the mona monkeys (*C. mona*) of Grenada, the macaques on Mauritius, and the Barbary apes (*Macaca sylvana*) of Gibraltar.

III. Form and function

DISTINGUISHING FEATURES

The basis of the success of the order Primates is the relatively unspecialized nature of their structure and the highly specialized plasticity of their behaviour. This combination has permitted the primates throughout their evolutionary history to exploit the wide variety of novel ecological opportunities that have come their way. Although there are a few highly specialized species among the lower primates (the aye-aye, the tarsier, the potto, and the lorises, for instance), the higher primates, the anthropoids, are extremely conservative in their structure; morphologically speaking, they have maintained a position in the evolutionary midstream and have avoided the potential stagnation of specialized life near the bank. Specialization is not always a liability; in times of environmental stability the specialized animal enjoys many advantages, but in a rapidly changing world it is the less specialized animals that are more likely to survive and flourish.

The importance of a large brain

The plasticity of primate behaviour is largely a function of the brain. The primate brain is distinguished by its relatively large size compared with the size of the body as a whole; it is also notable for the complexity and elaboration of the cerebral cortex, the function of which is to receive, analyze, and synthesize the incoming impulses from the sense organs and to convert them into appropriate motor actions, which in turn constitute behaviour.

It has been observed that there is no distinguishing feature which characterizes all primates except a negative one—their lack of specialization. St. George Mivart defined the order Primates in the following terms:

> . . . an unguiculate, claviculate, placental mammal with orbits encircled by bone; three kinds of teeth at least at one time of life; brain always with a posterior lobe and a calcarine fissure; the innermost digits of at least one pair of extremities opposable; hallux with a flat nail or none; a well-developed caecum; penis pendulous; testes scrotal; always two pectoral mammae.

When these criteria are examined one by one the inevitable conclusion is reached that, singly, none of these characters is unique to primates nor do all primates demonstrate every single character. Collectively, however, more primates than nonprimates show this particular constellation of characters. Perhaps a more meaningful way to look at primate characteristics is in the form of evolutionary trends. This approach is quantitative, inasmuch as it is recognized that not all adaptations are developed in primates to the same degree. A list of evolutionary trends with behavioral connotations is also a somewhat more functional inventory than Mivart's essentially morphological one and, furthermore, is more applicable to the study of fossil primates.

Primates are essentially arboreal animals whose limbs are adapted for climbing, leaping, and running in trees. Active arboreal life requires the mechanical assistance of a long tail and sensitive, grasping hands and feet with opposable thumbs and big toes, to aid in climbing and to ensure stability on slender branches high above the ground. Active arboreal locomotion also requires a much more accurate judgment of distances than life on the ground; this is facilitated by the development of stereoscopic vision, the anatomical basis of visual judgments in depth. The forward-facing eyes of primates are adaptations for this type of visual precision. A highly developed sense of smell is not nearly as important for animals leading an arboreal life as it is for those on the ground. Primates thus have a much reduced olfactory mechanism; noses are shorter, the turbinal (or scroll) bones of the nose are reduced in number and complexity compared with most nonprimate mammals.

Binocular vision

Primates, being on the whole omnivorous animals, have evolved a multipurpose dentition; incisors to cut, canines to pierce and tear, and molars and premolars to crush and grind. Such a dental arrangement equips its possessor with the power to cope with insect food, vegetable food, or flesh and provides a major contribution for the generalized way of life that is so characteristic of the order.

A basic adaptation for the primate way of life (including the unique specialization of man—upright walking) is verticality of the trunk. Above the level of the tree shrews, the ability to hold the body upright is a universal primate possession; indeed paleontological evidence indicates that truncal uprightness was a primate characteristic as far back as the early Eocene, some 54,000,000 years ago.

Above all, the principal evolutionary trend of primates has been the elaboration of the brain, particularly of that portion of the cerebral hemispheres known as the neopallium or neocortex. A neocortex is characteristic of higher vertebrates, such as mammals, which operate under the control of multiple sources of sensory input such as those from the nose, the eyes, the ears, and the skin. In many mammals the olfactory system dominates the senses and the cerebral hemispheres consist largely of paleocortex—the "smell brain"—of lower vertebrates. The arboreal habit of primates has led to a dethronement of the olfactory sense and the accession of a tactile, visually dominant sensory system. This evolutionary trend, combined with a progressive increase in the functional significance of the hands and feet in grasping and manipulation, has resulted in the dramatic expansion and differentiation of the neocortex.

GENERAL STRUCTURE

Vertebral column and posture. All primates retain a clavicle or collarbone (lost in many mammalian groups), a separate radius and ulna in the forearm, and a separate tibia and fibula in the lower leg. The single exception to this is the tarsier, in which the fibula becomes fused to the tibia.

The primate vertebral column shows a basic mammalian pattern of components including an "anticlinal" vertebra, situated in the mid-thoracic (upper-back) region of the spinal column and marking the transition between the forelimb and hindlimb segments. In a galloping greyhound the anticlinal vertebra is at the apex of the acute curve of the back. An anticlinal vertebra is characteristic of all quadrupeds and is seen in all primate families except the hominids and the pongids, whose posture is upright or semi-upright. The evolutionary trend in the vertebral column is toward shortening the lumbar, sacral, and caudal regions. Extreme shortening in the lumbar region, with complete loss of the caudal (tail) vertebrae, occurs in gibbons, the great apes, and man. Prehensility of the tail is a specialization of certain New World monkeys, but appears also for a brief period in the infants of many Old World monkey groups, in which it provides an important mechanical aid to survival.

Hands and feet. With two exceptions, all primates have retained five digits on hand and foot. The exceptions are the spider monkeys of South America and the colobus monkeys of Africa, which have lost or reduced the thumb. This appears to be an adaptation for locomotion, the rationale for which is not fully understood at present.

All (except tree shrews) possess prehensile (grasping) hands, and (except man) prehensile feet. The hands of Anthropoidea show a greater range of precise manipulative activity than those of Prosimii. Prosimians lack the functional duality of the hands of anthropoids. Duality in hand function has been described for man and nonhuman

Power versus precision grips

primates in terms of precision and power grips. The power grip of prosimians is very well developed but the precision grip is lacking. Among Anthropoidea, the New World monkeys show a considerable advance over prosimians in tactile sensitivity but they possess less functionally effective hands in prehensive terms than Old World monkeys.

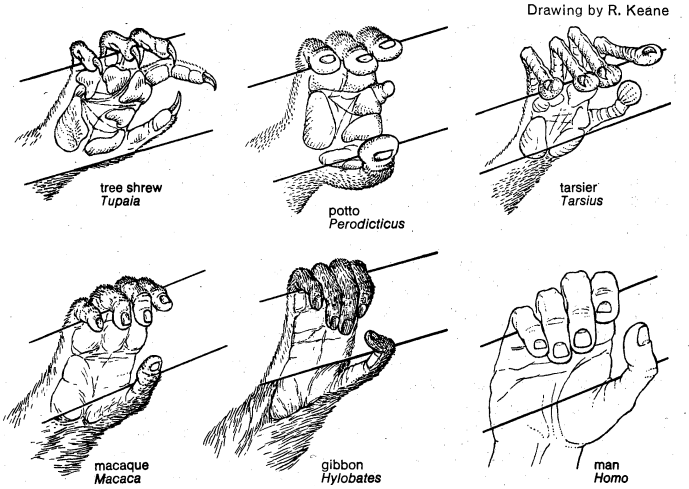


Figure 4: Shape and function of the hand in certain primates.

The critical component of the prehensile hand in terms of skilled manipulation is the opposable thumb: a thumb, that is to say, that is capable of being moved freely and independently. The movement of opposition is a rotary movement in which the thumb, swinging about its own axis, comes to face the under or palmar surface of the tips of the fingers. The opposable thumb is the basis of the precision grip which, though present to some extent in all primates, is particularly highly developed in man. Opposability is present to some degree or other in most primates but varies considerably in its functional effectiveness as an instrument of fine manipulation. The baboons and man are pre-eminent in this respect. The apes, having short thumbs and long fingers, are handicapped in relation to delicate manual dexterity but are adept in the coarser elements of hand use, particularly in relation to tree climbing and branch swinging.

Teeth. A dentition with different kinds of teeth (heterodonty)—incisors, canines, and cheek teeth—is characteristic of all primates and indeed of mammals generally. Heterodonty is a primitive mammalian characteristic and primates have evolved less far from the original pattern than most mammals. The principal changes are a reduction in the number of teeth and an elaboration of the cusp pattern of the molars.

The primitive mammalian dental formula is assumed to have been $\frac{3 \cdot 1 \cdot 4 \cdot 3}{3 \cdot 1 \cdot 4 \cdot 3} = 44$ teeth (the numbers being the numbers respectively of pairs of incisors, canines, premolars, and molars in the upper and lower jaws). No living primate has retained more than two incisors in the upper jaw and only in tree shrews is the primitive number of three incisors retained in the lower jaw. The incisors are subject to considerable variation in prosimians. Characteristically, the upper incisors are peglike, one or other pair often being absent; in the lower jaw the incisors show a peculiar conformation which has been likened structurally and functionally to a comb. The dental "comb" of prosimians is composed of the lower canines and lower incisors compressed from side to side and slanted forwards; the dental comb appears to be primarily a device for grooming the fur. Canines are present throughout the order, but show remarkable variation in size, shape, projection, and function. Characteristically the teeth of cercopithecines have a function in the maintenance of social order within the group as well as an overtly offensive role; their function as organs of digestion is relatively unimportant. They are large and subject to sexual dimorphism, being larger in males than females. Man's canines are small and are almost wholly dental in function, with no size differences between the sexes.

Evolutionary reduction of the cheek teeth

The trend in the evolution of the cheek teeth has been to increase the number of cusps and reduce the number of teeth. Both molars and premolars show this tendency. No living primate has four premolars; prosimians and New World monkeys have retained three on each side of each jaw, but at the anthropoid level two premolars are the invariable rule. The primitive premolars are uniform in shape and are unicuspid, but in primates the most posterior premolar tends to evolve either one or two extra cusps (molarization), an adaptation providing an extended tooth row for a herbivorous diet. In species with large upper canines, the most anterior lower premolar assumes a peculiar shape known as sectorial, functioning as a hone for the scythelike canine. In man, whose canines are small and unremarkable, the first and second premolars are identical in shape and two cusped.

The trend in the morphology of the molars has been to increase the primitive three cusps to four or five. The apes and man are more specialized in this feature than are the monkeys or the lower primates, having four cusps on the molar crown in the upper jaw and five cusps on the lower. A further tendency has been to reduce the molar series from three to two in both jaws, principally in the lower. Among the family Callitrichidae, which includes the marmosets, the pigmy marmosets, the tamarins and the golden marmosets, two molars are the rule, but in many other platyrrhine and catarrhine genera, as well as the genus *Homo*, the last molar (the wisdom tooth, so-called because it normally erupts in young adulthood coincidently, one might say, with the acquisition of sagacity) is frequently absent.

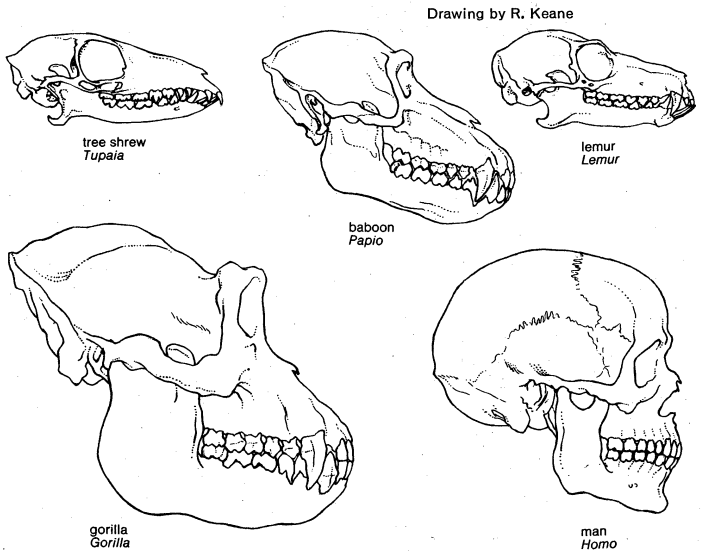


Figure 5: Skulls and teeth of primates.

Snout, muzzle and nose. The reduction of the snout in primates is a correlate of the diminution of the olfactory sense (see below *Sensory reception*). To a great extent visual acuity and manual dexterity have replaced the sensitive, inquiring nose found in so many nonprimate mammals. A marked reduction in the complexity of the "scroll" bones of the nose, the richness of the innervation of the olfactory mucous membrane, and the sensitivity of the moist tip of the nose—the rhinarium, is associated with the reduction in length of the primate snout. Although the trend in primate evolution is towards a dethronement of the primacy of the sense of smell, there are still some good snouts to be seen in those lower primates, that retain a naked moist rhinarium attached to the upper lip. The attachment of skin to gum is marked by a prominent groove bounded laterally by naked skin.

Prosimians depend for many aspects of their social and reproductive behaviour on olfactory signals which are effected by means of special scent glands distributed in different regions of the body but congregated principally in the anal and perineal regions. Marking behaviour, the placing of scent at various points in the environment, is a prominent feature of the prosimian repertoire of

communication. Marking behaviour ceases to be of much importance above the prosimian level, and in anthropoids the structures by which olfactory signals are given and received are diminished; the snout becomes abbreviated, the scroll bones less complex, and the glandular, moist rhinarium is replaced by a dry, hairy skin, and the upper lip is free and mobile. The scent glands largely disappear. Monkeys, apes and man, though they may be lower down the scale of olfactory excellence than prosimians, rely in considerable measure on the sense of smell; all higher primates, for instance, including man, sniff at unfamiliar items of food before placing them in the mouth.

The platyrrhine–catarrhine division

The shape of the nose of higher primates is one of the most reliable means of distinguishing Old World monkeys from New World monkeys at a glance. In New World monkeys or Platyrrhines (platyrrhine means "flat nosed") the nose is broad and nostrils are set wide apart, well separated by a broad septum, and point sideways; in Catarrhines (catarrhine: "downward nosed") the nostrils are set close together, point forwards or downwards, and are separated by a very narrow septum.

Some Old World monkeys, particularly those that have adopted a ground-living way of life, such as baboons, mangabeys, and the mandrills of the subfamily Cercopithecinae, appear to have re-adopted a long snout during their evolution. This structure, however, is not primarily olfactory in function but seems, rather, to be more closely related to the large size of the jaws and the prominence of the canine teeth; it should be considered a dental muzzle rather than an olfactory one.

SENSORY RECEPTION AND THE BRAIN

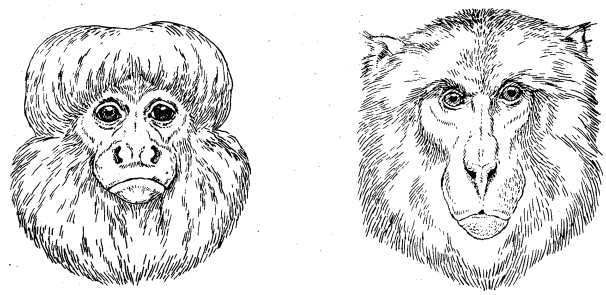
Among mammals in general, the olfactory system is the primary receptor for environmental information; consequently the brain of most mammals is dominated by the olfactory centres. In primates the sense of smell is considerably less important than the well-developed visual system and highly refined sense of touch. The primate brain is enlarged in the specific areas concerned with vision (occipital lobes) and touch (parietal lobes), thus taking a characteristic shape throughout the higher primates. Associated with the central apparatus for touch and vision are the peripheral sensory endings in the skin, particularly of the hand and face, and in the retina of the eye.

Touch. The skin of the primate hand is well adapted for tactile discrimination. Meissner's corpuscles, the principal receptors for touch in hairless skin, are best developed in the apes and man, but can be found in all primates except the tree shrew; structurally correlated with a high level of tactile sensitivity are certain anatomical features of the skin of the hands and feet, such as the absence of pads on the palms and soles, characteristic of mammals in general, and the presence of a finely ridged pattern of skin corrugations known as dermatoglyphics (the basis for fingerprints).

Eyes and vision. The evolutionary trend towards frontality of the eyes has not proceeded as far in the prosimian as in the anthropoid families. Tree shrews retain set of the eyes like that of the primitive ancestor, the central axis of each bony orbit being 140° apart. In lemurs this angle is considerably less, 60°–70°, and in the anthropoid families of apes and monkeys the divergence has been reduced to 20°. It should be noted that the axes of the eyeballs (as distinct from the bony orbits) in apes and monkeys are, in fact, parallel.

Colour vision is of considerable advantage to arboreal animals living on fruits and insects. Species that have both rod and cone receptors in their retinas include all of the Anthropoidea with the exception of some of the nocturnal species (*Aotus*, the night monkey of South America, for example), and the family Lorsidae and some of the Lemuridae among the prosimians. Rods are the more primitive retinal cells and respond to low intensity light, while cones, which respond to high light intensity, have a greater resolving power, and thus provide for a finer type of visual discrimination. Cones, moreover, are adapted for colour discrimination.

Nervous integration. The elaboration of touch and vision supplements the senses of smell, hearing, and taste,



red-backed saki
Chiropotes satanas

Japanese macaque
Macaca fuscata

Figure 6: Head of New World monkey (saki) compared with head of Old World monkey (macaque).
Drawing by R. Keane

providing the primate with a sensory armament of great range and flexibility. The primate central nervous system is sufficiently refined to deal with the elaborate bombardment of environmental information reaching it. Association areas provide the connections between the input and output centers of the brain—the motor and sensory cortex. Association areas are the memory banks where the memory of past experience is encoded in the infinitely complicated plexiform arrangement of the neurons, the brain cells, and their processes. All sensory impulses reaching the cortical centers of the central nervous system are routed through the association areas for conditioning, as it were, before reaching the effector side or output side where the appropriate response is initiated in the cells of the motor cortex and executed through the action of somatic (body) musculature. The more highly developed the association areas of the brain are, the more specific and appropriate is the behaviour and the more versatile is the animal in facing environmental demands.

The brain. As already indicated, the principal evolutionary trend in brain development has been towards elaboration. The neocortex of higher primates possesses highly developed associative functions, an aptitude for receiving, analyzing, and synthesizing the sensory input from visual, olfactory, auditory, gustatory, and tactile receptors and converting them into the appropriate motor responses.

Relative versus absolute brain size

The brain of Anthropoidea is larger, both absolutely and relatively, than that of prosimians. For instance, the weight of the simplest anthropoid brain, that of a marmoset, is three times greater than the brain weight of the bush baby, a prosimian of comparative size. This quantitative increase, in part, is attributable to the elaboration of the regions of the neocortex concerned with tactile and visual sensitivity; and in part to the elaboration of the intrinsic pathways connecting one part of the brain with another. The large brain of man is attributable not so much to an increased nerve cell content as to an increase in the size of the nerve cells and to a greater complexity of the connections linking one cell to another.

The external form of the anthropoid cerebral cortex is characterized by a complicated pattern of folds and fissures (sulci and gyri) in the brain surface. The fissural pattern is seen in its simplest form in the marmosets, but in the larger and more advanced members of the New World monkey infraorder, the capuchins (*Cebus*) for instance, the cerebrum is richly convoluted. Gyri and sulci are well marked in Old World monkeys and in the apes, the complexity of the pattern closely approximating the tortuous mazelike pattern seen in man.

REPRODUCTIVE SYSTEM

Male and female genitalia. The functions of the individual organs of reproductive systems are fairly uniform throughout the primates, but in spite of this physiological homology there is a remarkable degree of variation in minor detail of organs between groups, particularly in the external genitalia, which, by their variation, provide a morphological basis for the reproductive isolation of the species. There could be no more effective barrier to mating between different species than incompatability of the male and female sex organs.

Among the characteristics of the primate order as listed by Mivart, the penis is described as "pendulous" and the testes as "scrotal." In contrast to other mammals, the primate penis is not attached to the abdominal wall but hangs free. The testes, with a few exceptions among the prosimians, in which they are withdrawn seasonally, lie permanently in the scrotal sac, to which they migrate from their intra-abdominal position some time before birth (in man) or after birth (in nonhuman primates). In all primates except modern man (there is no evidence one way or the other regarding early representatives of the genus *Homo*), tarsiers, and some South American monkeys, the penis contains a small bone called the baculum, a typically mammalian character. The uterus of female primates shows all grades of transition between the two-horned (bicornuate) uterus, typical of most mammals, to the single-chambered (simplex) uterus of the higher primates and man. Broadly speaking, the prosimian uterus retains the primitive bicornuate pattern, while the anthropoid uterus is single-chambered.

Variations between primate taxa are demonstrated most strikingly by the glans penis, the scrotum, and perineum of the male, and the clitoris and the labial folds of the female vulva. The length and form of the clitoris, which when elongated mimics the penis, as in spider monkeys for instance, is a potent source of confusion in determining the sex of certain New World primates. The coloration of the male scrotum in forest-living primates, particularly of the genera *Cercopithecus* and *Mandrillus*, shows an infinite range of variations and provides a species-recognition signal of considerable effectiveness.

The external appearance of the genitalia undergoes seasonal variation in a number of primates. In the male, swellings of the testis and colour changes of the scrotum occur, and, in the female, swelling and coloration of the vulva and perineal region herald ovulation, sometimes most obtrusively. Turgidity and excessive vascularity of the tissues of the perineum are probably characteristic of all mammals, but there are certain primate species in which this engorgement reaches monstrous proportions, notably baboons, mangabeys, some macaque species, and chimpanzees. Regions other than the primarily genital areas may also be affected by the hormones circulating at certain periods of the reproductive cycle. For instance, in the gelada baboon, the skin on the front of the female chest, which normally bears a string of caruncles resembling the beads of a necklace, becomes engorged and brightly coloured. A German zoologist, Wolfgang Wickler, has suggested that this is a form of sexual mimicry, the chest mimicking the perineal region. The observation that geladas spend many hours a day feeding in a sitting posture provides a feasible, Darwinian explanation of this curious physiological adaptation.

Mimicry of bottom by chest colours

Placenta. The placenta, a characteristic of all Eutherian mammals, is a vascular structure that permits physiological interchange of blood and body fluids between the mother and the fetus and the breakdown products of the fetal metabolism; it also provides a two-way barrier preventing the passage of some, but not all, noxious substances and organisms such as bacteria and viruses from one individual to the other and is the source of hormones such as estrogens.

The placenta is a flat discoid-shaped "cake" in man, some of the other higher primates, and the tarsier. In many monkeys it is bidiscoidal, having two linked portions. The placenta is intimately attached on its outer surface to the endometrium, the lining of the uterus, by fingerlike processes (villi) that embed themselves in the endometrium where the vascular connections between the two circulations are achieved. The connection between fetal and maternal circulations appears as two distinct types among primates, a distinction which is believed to have had an important effect on the evolution of the order. In the first type (epitheliochorial), found in the prosimian suborder, several cellular layers separate the maternal and fetal bloodstreams, thus limiting the passage of molecules of serum proteins. In the second type, hemochorial placenta, the relationship is much more intimate, there being no fetal and maternal cellular layers separating the two circulations, and serum proteins can easily pass. A hemochorial type of placentation is found in tarsiers, apes, monkeys, and man.

The placenta is shed at birth in all primates, and, in all other than civilized man, is eaten by the mother.

IV. Evolution and paleontology

RENEWED INTEREST IN PRIMATE ORIGINS

The 1950s, 1960s, and 1970s were marked by a notable expression of interest in primate paleontology. Hardly a year has passed without some new major discovery being announced. New sites have been opened up and old discoveries redescribed and reallocated. New techniques in geological dating, in palynology (the study of fossil pollen), in paleoclimatology and paleoecology and in the archeological interpretation of fossil sites, have helped to lift primate paleontology into the forefront of the life sciences and have aroused public interest to an unprecedented level. The popularity of all aspects of the evolution of man is reflected, for instance, in the spate of books covering this field published during this period.

The African continent has contributed the greatest share of recent finds—Rusinga and Songhor in Kenya, Olduvai in Tanzania, and Sterkfontein, Kromdraai, Swartkrans, Makapan in South Africa are names with which every anthropology student is familiar. New names like Omo, Kanapoi, Lothagam, Napak and Moroto—all East African sites of hominid and pongid discoveries in the last few years—are now becoming as equally well known as Sterkfontein in the fifties and Olduvai in the sixties.

Important recent sites of primate fossils

Elsewhere, pieces of this colossal worldwide jigsaw have been discovered in Europe, notably in Hungary, France, and Italy; in the Siwalik Hills of northwest India; in the ever prolific middle Eocene Bridger Beds of North America, and in the La Venta deposits in Columbia and, recently (1969), in Bolivia. These latter discoveries are particularly welcome because of the poverty of fossils of New World monkeys.

While new discoveries have clarified the human story, older ones, that had only served to cloud it, have been repudiated. Piltdown man was shown unequivocally to be a fake in 1953; *Oreopithecus* of Monte Bamboli, Italy, has been relegated to a family remote from human ancestry, and Galley Hill Man in England has been shown to be a recent intrusion (a burial) into middle Pleistocene gravels. The questionable finds from the remoter geological periods of the Eocene and Oligocene have also been re-examined, with the result that a number of confusing fossils have been dismissed. *Anagale*, long thought to be a fossil tree shrew, is now regarded as a small carnivore; others, such as the supposed tarsioid–anthropoid annectent form, *Parapithecus fraasi*, have been reallocated within the order Primates.

Progress in constructing the phylogeny of higher primates and man has been bedevilled by a number of controversies concerning taxonomy and nomenclature, and, as a result, internal disagreements have developed which mitigate against the sort of international cooperation that is essential for so vast a project as the construction of a human phylogeny. New-school and old-school taxonomists have come into conflict. But with the rapid advances in genetics, in the new concepts of biological species and in population anthropology, a fresh equilibrium is slowly being acquired, as the pendulum swings between the reactionary "lumping" of such taxa as genera and the traditional "splitting," in which every new discovery was provided with a new generic name. The most extreme example of taxonomic lumping in primate paleontology was in a 1965 revision by E.L. Simons and D.R. Pilbeam of the Miocene–Pliocene genus *Dryopithecus*; from the 28 genera recognized prior to 1965, the authors recommended the recognition of three genera only: *Dryopithecus, Ramapithecus,* and *Gigantopithecus*. All the rest were absorbed into the three genera, a result initially satisfying but not yet subjected to the test of time.

THE PRIMATE FOSSIL RECORD

Cretaceous. The known range of the primate order was extended back from the middle Paleocene to the Late

Cretaceous (about 75,000,000 years ago) by the discovery in Montana of one premolar and four molar teeth, representing two species of insectivore-like primates that were assigned in 1965 to a new genus, *Purgatorius*. This diagnosis is based on the characters of four molar and one premolar tooth and is not by any means universally accepted.

Paleocene. The best known early primate species, of which, *inter alia*, a complete skull and a partial postcranial skeleton are available, belongs to the genus *Plesiadapis* from the Paleocene (about 60,000,000 years ago) of Cernay, France, and Colorado. The skull shows a number of dental specializations including procumbent rodent-like incisors in the upper and lower jaw and the absence of antemolar teeth, but the molar teeth show the unequivocal primate affinities of this species. Rodent affinities have been suggested for *Plesiadapis*, but most authorities agree that it is a primate, however aberrant.

Eocene. The known Eocene fossil families include the Tarsiidae (tarsiers), the Adapidae, considered probable ancestor of lemurs and lorises, and the Omomyidae, possible ancestors of the monkeys and apes.

Possible ancestors of the lemurs

The family Adapidae contains two North American genera, *Notharctus* and *Smilodectes*, which are well represented in the fossil deposits of the Bridger Basin, Wyoming. *Adapis*, a closely related genus, was widely distributed in Europe, as were *Anchomomys* and *Pronycticebus*. *Notharctus* and *Smilodectes* are not thought to be antecedent to living lemurs, principally on zoogeographical grounds, since egress from the North American continent for tropical species was effectively cut off by ecological changes soon after the beginning of the Eocene (about 54,000,000 years ago). *Notharctus* was not unlike the modern lemurs in size and general appearance. On both morphological and zoogeographical grounds, *Adapis* and *Pronycticebus* or *Anchomomys* could have provided the stem from which the living lemurs and lorises evolved.

Representatives of the Omomyidae have been found in North America, in Europe, and in Asia. According to American paleontologist Elwyn L. Simons, the Omomyidae are probably the key family for tracing the phylogeny of the higher primates, the apes and monkeys. The family was common to North America and Eurasia, a fact which reinforces the presently held view that omomyids provided the stock from which New World monkeys and Old World monkeys, apes, and man arose. No other Eocene primate family has a greater claim to be ancestral to the Anthropoidea. Their principal claim to antecedency (other than the circumstances of their zoogeography) is the dental formula: $\frac{2 \cdot 1 \cdot 3 \cdot 3}{2 \cdot 1 \cdot 3 \cdot 3} = 36$ teeth (18 pairs).

The Eocene Tarsiidae, represented by the European species *Necrolemur antiquus*, found in the Quercy deposits of France, are likely to contain the ancestor of the modern genus *Tarsius*. The tarsier is indeed a "living fossil" in the best sense of that overworked term. What is known of the skull and skeleton of *Necrolemur antiquus* indicates that it was almost identical with that of the living *Tarsius*.

Oligocene. Information on primate evolution during the Oligocene Epoch (about 38,000,000 to 26,000,000 years ago) rests principally on discoveries in three areas —Burma, Texas, and Egypt, and by far the most important is Egypt. From the Fayum, from the upper levels of the Qatrani Formation, has come the first evidence of the emerging Anthropoidea. A number of different genera have been described from the Fayum, including *Apidium, Propliopithecus, Oligopithecus, Parapithecus, Aeolopithecus*, and *Aegyptopithecus*. The Fayum, indeed, seems to be the cradle of the Old World anthropoids. Simons has indicated that the possible ancestors of Old World monkeys (*Parapithecus, Apidium*), of gibbons (*Aeolopithecus*), of apes (*Aegyptopithecus*) and even of man (*Propliopithecus*) are all represented in the fauna. Unfortunately, apart from a few fragmentary scraps of limb bones, there are no postcranial bones to help. From the evidence provided by the possible forerunners and probable descendants of the Fayum primate fauna, however, it is reasonable to assume that quadrupedalism was becoming established as the typical locomotor pattern, and that vertical clinging and leaping, the characteristic gait of the Eocene forebears of the fauna, was probably retained by at least some of the genera represented at this site.



Figure 7: Eocene fossil lemur *Notharctus*.
Drawing by R. Keane

Of unusual interest is the recent discovery of the cranium of a North American omomyid called *Rooneyia*; it is of particular note in view of a belief that primates had disappeared from North America by the late Eocene times. *Rooneyia* is also of considerable interest in itself. The skull possessed a mixture of primitive (prosimian) and advanced (anthropoid) features, precisely the combination that might be anticipated in a transitional form bridging the gap between the lower and higher primates.

Rooneyia, an important transitional form

Miocene. The Miocene Epoch is probably the most fruitful of all for paleoprimatology. During the 19 million years of its duration (beginning about 26,000,000 years ago), dramatic changes in geomorphology, climate, and vegetation took place. The Miocene was a period of volcanism and mountain building, during which the topography of the modern world was becoming established; of particular relevance to the story of primate evolution are the vegetational changes resulting from the orogenic ones. Grasses, known since the early Tertiary Era, flourished in the new conditions and in many areas that were previously forested. Grassland is known regionally by such names as savanna, llanos, prairies, etc. A new type of primate—the ground inhabitant—came into being during this period. Today ground-living primates, in terms of population numbers, include the most successful primate species, the macaques, the baboons, and man. The generalized nature of the bodily form of primates, combined with their specialized brain, made possible this critical step. It is probable that man owes his very being to this geomorphological revolution.

In the last few decades considerable additions to the knowledge of ape and human evolution have accrued from Miocene fossil beds in East Africa and Europe.

In Europe further discoveries of the gibbon-like genus *Pliopithecus* and the gorilla-chimp-like genus *Dryopithecus* have been made. *Pliopithecus* remains from Neudorf-an-der-March in Czechoslovakia have provided a remarkably complete picture of the habits of the Miocene apes, which, on this evidence, appear to have possessed bodily forms of a quadruped with morphological overtones of incipient brachiating specializations. Perhaps most interestingly of all, *Pliopithecus*, with its retention of numerous platyrrhine or New World monkey characters, emphasizes the phylogenetically conservative nature of anthropoid evolution.

In East Africa the excavations of the inshore islands and of the Kenyan shores of Lake Victoria by L.S.B. Leakey and a number of colleagues have illuminated man's knowledge of the evolution of the African great apes and have set up provocative questions regarding the origins of man. The *Proconsul* group (relegated to subgeneric status within *Dryopithecus* by the revision of Simons and Pilbeam) is known from three fossil species of ape, *Proconsul major, P. nyanzae,* and *P. africanus.* All three species are represented by a large number of jaw and facial fragments, but the only complete skull, discovered in 1948, belongs to *P. africanus.* Though complete, it was somewhat distorted by pressure of the surrounding rocks during fossilization. Subsequent reconstruction revealed a skull more monkey-like than apelike in its contours, an appearance which belied the unquestionably chimpanzee-like affinities of its teeth. The form of the skull and braincase of *P. africanus,* along with the forelimb skeleton, which is known in great detail for this species, indicates a body form that most closely resembles that of living monkeys. An American anthropologist, Sherwood L. Washburn, has described these creatures as "dental apes." The monkey-like characteristics of the *Dryopithecus* (*Proconsul*) group reflect the recency of the separation of the monkey and ape stocks of the Old World.

Proconsul africanus has been suggested as a possible common ancestor for apes and monkeys, but the general consensus is that *P. africanus* was already too specialized dentally to fit such a role. Be this as it may, there is no doubt that this Miocene fossil ape provides a very reasonable "model" for such an ancestor.

In East Africa two species of fossil gibbons, *Limnopithecus,* have also been discovered, closely related, if not generically identical, to the fossil "gibbon" of European deposits termed *Pliopithecus.* The taxonomic status of these fossils as "gibbons" is somewhat uncertain.

Appearance of hominids

It is towards the end of this turbulent epoch that the first probable ancestral hominids appear in the fossil record. *Ramapithecus punjabicus* was discovered in 1934 by an American paleontologist, G.E. Lewis, in the Siwalik Hill Miocene deposits in northwest India. Lewis tentatively expressed the view that *Ramapithecus* was a hominid, but the scientific climate of the period was not propitious for claims that seemed to imply such remote antecedents for man; as a result Lewis's perceptiveness was largely ignored. Recently, Simons and Pilbeam have resurrected Lewis's specimen and have demonstrated its hominid characteristics. At Fort Ternan in East Africa, meanwhile, Leakey found what at present appears to be another specimen of *Ramapithecus,* but of a different species and called by him *Kenyapithecus wickeri.* If and when it is proved (which seems probable) that *Kenyapithecus* was simply a *Ramapithecus* then it will, following the laws of nomenclatorial priority, be officially renamed *Ramapithecus wickeri.*

First ground-living monkey

It was also from the late Miocene that the first unequivocally ground-living monkey came to light. Attributed to the living subfamily Cercopithecinae of Old World monkeys, these specimens of *Mesopithecus pentelici* demonstrate the morphological transition between the ancestral monkey type (today represented by the leaf-eating monkeys or Colobinae of Africa, India, Ceylon, Malaya, Burma, and Southeast Asia) and the newly evolved adaptive type of ground-living monkey epitomized by *Macaca* or *Papio.* In its known features of skull and postcranial skeleton, *Mesopithecus* is broadly equivalent to the macaque monkeys. Although *Macaca* is now almost wholly restricted to Asia, the ancestral stock was at one time widely distributed throughout Europe as well.

Pliocene. The Pliocene Epoch (from about 7,000,000 to 2,500,000 years ago) was very similar to the present in terms of its geomorphology and climate. Discounting the effects of the recent influence of man on the distribution of forest and savanna in the tropics, the face of the land cannot have differed much from the aspect it presents today. Thus, one would expect that during the Pliocene (bearing in mind the effectiveness of environmental selection) essentially modern forms of primates would have made their appearance, but perhaps the most famous of the Pliocene fossils was far from "modern." *Oreopithecus* was first discovered, fragmentarily, in brown coal deposits in Italy in the middle of the 19th century. The most recent discovery of a complete, though paper-thin, skeleton (so grossly has it been compressed) in Tuscany has demonstrated the power of the environment acting upon the genetic structure, in once again reproducing the trade marks of two of the "inevitable" trends of primate evolution, brachiation, and bipedalism. *Oreopithecus* possessed a number of dental and bony characters that are typically hominid; it also possessed limb characteristics that are typically pongid. The canines were relatively short, the face was abbreviated, and the pelvis was broad. The foot showed characteristics associated with bipedal walking as did the vertebral column (all hominid characters), yet the arms were long and the fingers long and curved. The limb proportions are those of a brachiator (pongid characters). Finally, a number of characters of this confusing hominoid typically reflect those found in Old World monkeys. The present consensus of attitude towards this composite primate is that it is best segregated within the superfamily Hominoidea but in a family of its own, the Oreopithecidae. That the end of the *Oreopithecus* story has not yet been heard is certain; in the meantime the Tuscany fossil stands as a memorial to the primate order, the order which, of all of the class Mammalia, is most subject to environmental adaptation by virtue of its behavioral plasticity.

Few fossils referable to the ape family are known during the Pliocene, and monkey families are scarcely better known. *Libypithecus* and *Dolichopithecus,* both monkeys, were probably ancestral cercopithecines that still retained some features of their colobine ancestry; but neither genus can be placed in a precise ancestral relationship with modern members of this subfamily.

Pleistocene. The Pleistocene, which on faunal evidence is estimated to have started 2,000,000 to 3,000,000 years ago, is the epoch of hominid expansion (see HOMINIDAE). Knowledge of nonhuman primates is surprisingly sketchy. No ape fossils are known until relatively recent times and monkeys have been identified in only a few regions in Africa and even fewer in Asia: *e.g., Cercopithecoides* (a colobine-like form), *Gorgopithecus, Dinopithecus,* from South African deposits (forms of the living genus *Papio*); and *Simopithecus* (a giant, ancestral forerunner, according to one authority, of the present-day genus, *Theropithecus,* the gelada) from Olduvai Gorge and South Africa. It is possible that the *Papio-Theropithecus* divergence can be pushed well back into the Pliocene.

The special case of *Gigantopithecus*

One genus of the Pleistocene that is neither ape nor monkey in the sense that these taxa are interpreted today is *Gigantopithecus.* The romantic story of the discovery of the gargantuan molar teeth of *Gigantopithecus blacki* by von Koenigswald, the Dutch paleontologist, in a Chinese pharmacy has often been told. The boldness of the move that erected a new genus on such apparently slender grounds has been amply justified by the subsequent discovery of several massive jaws from Kwangsi in South China and attributed to this genus. The most recent discovery is of an enormous jaw in the Dhok Pathan deposits of the Siwalik Hills of India. Possibly dating back to the middle Pliocene, this discovery provides a respectably long period of existence for this aberrant giant-toothed hominoid genus. Clearly *Gigantopithecus bilaspurensis* was a member of the family Pongidae with divergent dental specializations that were possibly adaptive for the specialized form of herbivorousness recently termed graminovorous, an adaptation for foraging in grassland where tree products were unavailable and ground products available but hard to get. *Gigantopithecus* was a fossil ape that fell by the wayside on its way up to hominid status, a victim, possibly, of overspecialization.

V. Classification

DISTINGUISHING TAXONOMIC CHARACTERS

The identifying characters of the primate order have already been discussed above. The point has been made

that no single morphological characteristic distinguishes this group exclusively from other mammalian taxa. For instance, while it is possible to make the generalization that primates bear flat nails on the ends of their fingers and toes, it is not strictly accurate, for species such as the tree shrews, the marmosets, and the tamarins have either overt claws (tree shrews) or modified claws (marmosets and tamarins). One can also aver that primates characteristically possess two pectoral mammae, but it must be pointed out that the aye-aye (*Daubentonia madagascariensis*) has its nipples sited in the inguinal region (lower lateral region of the belly) and that mouse lemurs and tree shrews have multiple sets. Generalizations regarding opposability of the thumb, a "classic" primate character, are particularly vulnerable since some marsupials possess this trait and some primates do not. Examples of the generalized nature of primate specializations are too numerous to mention here. At first sight, big toe opposability provides a more promising diagnostic feature but, as with other features, there are exceptions to the rule, man for instance; moreover, it occurs in nonprimates.

As F. Wood Jones said in his book *Man's Place Among the Mammals*: "There is no single [characteristic] which constitutes a peculiarity of the Primates: for a primate animal may only be diagnosed by possessing an aggregate of them all." It was with this precept in mind that Sir Wilfrid Le Gros Clark introduced his list of evolutionary trends of primate evolution in 1959, which is amplified below:

1. The preservation of a generalized structure of the limbs with a primitive pentadactyly [i.e., containing five digits] and the retention of certain elements of the limb skeleton (such as the clavicle) which tend to be reduced or to disappear in some groups of mammals.
2. An enhancement of the free mobility of the digits, especially the thumb and big toe (which are used for grasping purposes).
3. The replacement of sharp compressed claws by flattened nails, associated with the development of highly sensitive tactile pads on the digits.
4. The progressive abbreviation of the snout or muzzle.
5. The elaboration and perfection of visual apparatus with the development to varying degrees of binocular vision.
6. Reduction of the apparatus of smell.
7. The loss of certain elements of the primitive mammalian dentition, and the preservation of a simple cusp pattern of the molar teeth.
8. Progressive expansion and elaboration of the brain, affecting predominantly the cerebral cortex and its dependencies.
9. Progressive and increasingly efficient development of those gestational processes concerned with the nourishment of the foetus before birth. (From W.E. Le Gros Clark, *The Antecedents of Man*; © 1959 Edinburgh University Press.)

Two further evolutionary trends may be added to the list:

10. Progressive development of truncal uprightness leading to a facultative (opportunistic) bipedalism.
11. Prolongation of postnatal life periods.

From J.R. and P.H. Napier, *A Handbook of Living Primates* (1967).

A British zoologist, R.D. Martin, has criticized many of these evolutionary trends as irrelevant as determinants of primate origins. In his view, many of them are either symplesiomorph characters (primate characters inherited from an ancestral mammal and therefore nonspecific for primates) or convergent characters (characters that have been acquired by many groups of primates in response to the environmental effects of natural selection). The disputed trends of primates, nonetheless, comprise characters that are of considerable significance for primate classification. Martin's point is that such characters are of little use in distinguishing the order Primates from other placental orders. Martin argues that the inadequate nature of the current definitions of primates is the principal reason why, when so much behavioral and structural evidence points to the opposite conclusion, the tree shrews (Tupaiidae) are presently included in the order. This may well be so. But until the true synapomorph characters (characters of independent acquisition) are identified by extensive comparative studies of all of the relevant characters of primates and nonprimates, there is little justification for arbitrary disenfranchisement of the tree shrews. "Possession is nine-tenths of the law," a principle that, as in this instance, supplies the highly desirable element of stability to zoological nomenclature.

ANNOTATED CLASSIFICATION

It is apparent that the taxonomy of primates is in a very fluid condition; but this state of affairs is wholly preferable to the virtual stagnation that existed during the first half of the present century. In view, therefore, of the likelihood of further major revisions taking place during the next decade that may well reveal the fallacies of certain current concepts, the informational content of this section is best served by a conservative, rather than a radical, approach to primate classification. Thus, the most widely accepted current classification is here followed, while fully admitting that it is not wholly adequate. In order to avoid some of the descriptive complications, derived from an imperfect classification, the following scheme outlines the distinguishing taxonomic characters of order, suborders, and families; intermediate taxa, such as infraorders and superfamilies, are omitted.

ORDER PRIMATES

Distinguished by the possession of a hallux (first toe) and a pollex (thumb), one or both opposable; eyes frontally placed on the face, orbit ringed by bone and, in higher members of the group, closed posterolaterally by bone. Brain relatively large. Most of the fingers and toes bear nails, not claws. Number of teeth 18 to 36; canines prominent but dentition as a whole not highly specialized. Diet omnivorous; body weight from about 100 grams (a few ounces) to 180 kilograms (400 pounds), more or less. Twelve Recent families, 55 genera and 197 species; 8 extinct families. Extinct groups are indicated by daggers (†).

†Suborder Plesiadapoidea

†*Families Phenacolemuridae, Carpolestidae, and Plesiadapidae*

Middle Paleocene to lower Eocene; North America and (except Phenacolemuridae) Europe; collectively, more than a dozen genera.

Suborder Prosimii

Usually quadrupedal and plantigrade, but some species show vertical clinging in trees and saltation on the ground; all have long tails (except *Indri*). Very small to medium in size; substantially insectivorous, but fruit- and leaf-eating common; activity rhythm nocturnal or crepuscular. Suborder characterized by possession of long or longish snouts, naked rhinarium, and naked tethered upper lip (except *Tarsius*); facial vibrissae (whiskers) present. Specialized sublingual fold, often serrated, below tongue. Eyes large, fully frontally facing in *Tarsius*, but somewhat divergent laterally in others; bony orbits not fully closed posterolaterally. Teeth characterized by special arrangement of lower incisors and canines to form a dental "comb," except in *Tarsius* (structure absent) and in *Tupaia* (canines take no part in comb); upper incisors usually small and peglike. Hallux and pollex more or less opposable, tips of all digits bear flat or flattish nails (except in *Tupaia*), second digit of hindlimb bears toilet (grooming) claw. Placenta epitheliochorial and non-deciduate (except *Tarsius*). One or more pairs of nipples. Single births are the rule.

Family Tupaiidae (tree shrews)

Small animals with elongated body and long bushy tail (scaly with terminal tuft in *Ptilocercus*). Short limbs, with fore longer than hind. Long pointed snout. Relatively large brain (compared with nonprimates) with diminished olfactory representation; turbinal bones of nasal region of skull not well developed. Tympanic ring within middle ear cavity. Primitive dental pattern, with 3 pairs of lower incisors, protuberant and arranged as modified "dental comb"; cheek teeth primitive in form. Digits of hands and feet bear claws; pollex is incipiently divergent, but nonopposable; hallux, nonopposable. Both arboreally and terrestrially adapted species occur within the family. Head and body length 10–24 cm (4–9½ in.), tail 12–14 cm; 5 genera, 17 species; India and Southeast Asia and Philippines.

†*Family Adapidae* (*Notharctidae*)

Lower to upper Eocene; Europe, Asia, and North America; about 10 genera.

Family Lemuridae (lemurs)

Size very small to medium; tail longer than head and body length combined; hindlimbs considerably longer than forelimbs; tail more developed in subfamily Lemurinae than in the other subfamily Cheirogalinae. Coat colours highly variable;

one species, *Lemur macaco*, shows sexual dichromatism (males black, females rufous). Tympanic bullae large, tympanic ring contained within middle ear cavity (contrasting with Lorisidae). Cutaneous glands on forearm and upper arm serving marking functions in *L. catta* and *Hapalemur griseus*; absent from other species. Dentition typical of Prosimii, including "dental comb" and reduced upper incisors (absent in *Lepilemur*). Diurnal (*Lemur, Hapalemur*) or nocturnal. Births seasonal; twins common in *Microcebus*. Head and body length 13–46 cm (5–18 in.), tail 17–56 cm; 6 Recent genera; about 14 species (some possibly extinct); restricted to Madagascar.

Family Indriidae (indri, sifaka, and avahi)

Arboreal animals with exceptionally long limbs, associated with specialized locomotor habit of vertical clinging and leaping. Skull with somewhat downwardly facing foramen magnum, short facial skeleton and very large auditory bullae. Dentition atypical of Prosimii; lower canines absent, first premolar somewhat caniniform, 2 premolars only. Anatomy of stomach and cecum, particularly, specialized in association with a bulky vegetarian diet. Activity rhythm nocturnal (*Avahi*) or diurnal (*Propithecus* and *Indri*). Fossils from Pleistocene. Three living genera, 4 species; Madagascar.

Family Daubentoniidae (aye-aye)

Arboreal, nocturnal family comprising single species *Daubentonia madagascariensis*. Hand specialized by markedly elongated and attenuated third digit. Dentition anomalous: central incisors reduced to single large pair having long roots and a rodent-like function; canines absent, premolars absent in lower jaw and reduced in upper. Symphysis between right and left halves of mandible is fibrous, a condition unique among the primates. Single inguinal pair of nipples. Madagascar.

Family Lorisidae (galagos, lorises, angwantibo, and potto)

Arboreal and nocturnal. Two major locomotor types comprise family: vertical clingers and leapers (subfamily Galaginae) and slow quadrupedal climbers (Lorisinae), the former possessing excessively long hindlimbs, and the latter subequal limbs. Tails short or absent in Lorisinae and long in Galaginae. Second digit of hand reduced, extremely so in Lorisinae (particularly *Perodicticus* and *Arctocebus*) and first digit of hand and foot disproportionately large. Skull globular in form with little protrusion of facial region. Tympanic bullae large, short external auditory canal with tympanic ring sited at junction of middle and external ear. Orbits large, with little lateral divergence. Dentition typical of suborder. Omnivorous (with large insectivorous element). Family comprises 5 Recent genera with about 12 species; Africa, Asia (Lorisinae); Africa (Galaginae).

†*Family Anaptomorphidae*

Upper Paleocene to middle Eocene; North America and Europe; about 10 genera.

†*Family Omomyidae*

Upper Paleocene to lower Oligocene: North America, Europe, and Asia; about 20 genera.

Family Tarsiidae (tarsiers)

Arboreal and nocturnal. Hindlimbs almost twice the length of forelimbs. Hands with 5 digits, fingers surmounted by circular digital pads and backed by tiny triangular nails; thumb opposable, but in a fashion distinct from other Prosimii. Head of femur cylindrical (not spherical as in all other Prosimii except *Galago*), and tibia and fibula fused in lower third (unique to *Tarsius*). Skull characterized by enormous, frontally directed orbits, incompletely closed posterolaterally; nasal space much compressed. Foramen magnum (opening for spinal cord) directed downwards. Short external auditory canal; middle ear with prominent bulla and tympanic ring, sited as in Anthropoidea. Dentition unlike typical prosimian pattern, lacking dental comb; central lower incisors absent; molars sharp-cusped tuberculo-sectorial. Placentation is hemochorial, discoidal, deciduate, as in Anthropoidea. Single young usually produced. One genus, 3 species; Southeast Asia.

†*Family Microsyopidae*

Lower to upper Eocene; North America and Europe; 4 genera.

Suborder Anthropoidea

Quadrupedal, brachiating or bipedal in gait. Diurnal in habit (except night monkey, *Aotus*). Omnivorous in diet. Tail present or absent, varying in length from species to species, prehensile in some. Size ranges from very small (*Saguinus* and *Callithrix*, the tamarins and marmosets) to very large (*Gorilla* and *Homo*.) Faces show varying degrees of protrusion, most marked in the baboons and mandrills and least developed in man. Neurocranium relatively large and with or without sagittal or nuchal crests or both; foramen magnum directed downwards and backwards. Orbits frontally facing and fully closed posterolaterally; nasal cavity narrow. Dentition comprising 30–36 teeth, canines are well developed except in man. Hands and feet prehensile; thumb semi-opposable or fully opposable (except in certain platyrrhine species). Hallux widely abductable (capable of being rotated inward) and capable of prehension in all species except man; flat nails on fingers and toes (except Callitrichidae); specialized toilet claws absent. Two pectoral mammae only. Placenta hemichorial, discoidal, and deciduate. Lengthy gestation period and extended life periods (compared with Prosimii) characterize the suborder. Suborder contains two geographical groups: New World monkeys and Old World monkeys: 6 families and 35 genera; both the tropical and subtropical zones of New and Old World.

Family Callitrichidae (tamarins and marmosets)

Small animals, diurnal, arboreal, and mainly insectivorous. Quadrupedal gait, hindlimbs being 25 percent longer than forelimbs. Thumb nonopposable, but hallux fully abductable; all digits bear modified claws (tegulae) except hallux, which has a flat nail. External nostrils face forwards and sideways (platyrrhine condition); nasal septum broad. Coat colour highly variable between species including wholly black and wholly white forms; one species (*Leontopithecus rosalia*) golden; distinguishing external features include a variety of ear tufts and a diversity of mustaches. Skull long and ovoid in shape with bulging occiput; tympanic bulla present; external auditory canal absent. Dentition consists of 32 teeth (36 in *Callimico*) with 3 premolars and only 2 molars. Tamarins are distinguished from marmosets by long lower canines which project beyond incisors; in marmosets canines and incisors are the same length. Head and body length 13–37 cm (5–15 in.), tail 20–39 cm; weight about 100–700 gm. Five genera, 15 species; South America.

Family Cebidae (New World monkeys)

Larger in size than Callitrichidae, with ears more or less naked externally. All digits bear flattened or curved nails; thumb imperfectly opposable. Gait quadrupedal. Prehensile tail present in *Ateles, Alouatta, Brachyteles*, and *Lagothrix*. Bony ear formation as in Callitrichidae. Nose typically platyrrhine. Dentition of 36 teeth; molars quadrangular and quadricuspid, and canines moderately developed in both sexes. Head and body length about 24–70 cm (tail usually of same magnitude), weight about 600 to 10,000 gm (1.3 to 22 lb); 11 genera, 29 species; southern Mexico to Brazil.

†*Family Parapithecidae*

Lower Oligocene; North Africa; 2 genera.

Family Cercopithecidae (Old World monkeys)

Medium to large in size. Arboreal or terrestrial. Characterized by naked callosities covering expanded ischial bone of pelvis. Quadrupedal, plantigrade, or digitigrade in gait; hindlimbs longer than forelimbs in most species, but in ground-adapted forms limbs are nearly equal. Thumb present in all genera (except *Colobus*) and functionally opposable. Cheek pouches present in subfamily Cercopithecinae and concerned with temporary storage of food; in subfamily Colobinae pouches are absent but stomach is sacculated in association with a predominantly leaf-eating diet. Tail length variable from very long (Colobinae) to very short or absent (some Cercopithecinae). Muzzle more or less prominent. Noses with external nasal opening facing forwards and downwards or directly downwards ("catarrhine" condition); internasal septum narrow (as contrasted with Callitrichidae and Cebidae). Skull of anthropoid type but without auditory bulla; external auditory canal present. Dentition of 32 teeth; canines large in males, moderate in females; first lower premolar sectorial, molars quadricuspid with transverse linking crests (bilophodonty). Head and body length about 32–110 cm; tail, when present, 2–100 cm; weight about 0.7–30 kg (1.5–66 lb); 13 genera, 72 species; Africa, India, China, Japan, and South East Asia.

Family Hylobatidae (gibbons and siamang)

Demonstrate highly specialized form of locomotion called brachiation. Arboreal, tailless catarrhines with forelimbs twice length of hindlimbs. Hands exceptionally long and prehensile; deep cleft between first and second digits; thumb fully opposable by means of unique mechanism at carpo-metacarpal (basal thumb) joint, associated with a specialized musculature. Ischial callosities as in Cercopithecidae but developing late in life. Well developed laryngeal (vocal) sac in *Symphalangus*. Skull of typical anthropoid form, orbits large, forwardly directed and having a well-buttressed lateral margin; strong brow ridge; sagittal and nuchal crests rare; mastoid processes absent. Long sabre-like canines equally developed in both sexes; molar cusp pattern similar to that of the Pongidae and Hominidae. Chromosome diploid number: 44. Head and body length about 40–64 cm (16–25 in.); weight 4–13 kg;

2 genera; *Hylobates* (6 species) and *Symphalangus* (1 species); Southeast Asia.

Family Pongidae (great apes: gorilla, chimpanzee, and orangutan)

Arboreal or ground living. Locomotion consists of brachiation and a modified form of quadrupedalism, called knuckle walking. Characterized by total lack of tail, relatively long forelimbs, elongated hands, and short but fully opposable thumbs. Brain relatively large compared with body size and showing well marked cerebral fissuration. Permanent dentition 32 teeth: incisors large and spatulate, particularly in upper jaw; canines large and conical, particularly in males; well-marked diastema present between upper lateral incisors and canines. Molars large; quadricuspid in upper jaw and quinquicuspid in lower; cusps arranged in so-called "*Dryopithecus*-pattern"; enamel much wrinkled in *Pongo;* third molar relatively small in *Pan*. Skull bears sagittal and nuchal crests in adult males (but rare in *Pan*). Ischial collosities usually absent (*cf.* Cercopithecidae, Hylobatidae). Thorax widest transversely (*cf.* Cercopithecidae); pelvis expanded in iliac region; lumbar vertebrae reduced in number (*cf.* Cercopithecidae). Life periods extended compared with New and Old World monkeys. Chromosome diploid number: 48. Family includes three genera: *Gorilla, Pan, Pongo. Distribution*: Africa and Southeast Asia.

Family Hominidae (man and fossil relatives)

Ground-living, hairless, omnivorous and bipedal. Hindlimbs longer than forelimbs (contrasting with Pongidae and Hylobatidae). Thumb fully opposable and hand capable of both precision grip and power grip; big toe aligned with other toes, foot nonprehensile. Brain relatively large and highly fissured. Neurocranium expanded particularly in parietal and frontal regions; Mastoid process of periotic bone present from birth. Facial skeleton nonprotrusive and strongly flexed relative to cranial base. Dentition of 32 teeth: canines small and blunt, lying flush with occlusal plane; premolars bicuspid; molars small, low, bearing rounded cusps; last molar in upper and lower jaw frequently unerupted. Postcranial skeleton shows adaptations to upright posture particularly in vertebral column, pelvis and lower limb. Chromosome diploid number: 46. Life periods (including growth period and life expectancy) extended as compared with Pongidae. Head and body length about 80 to more than 130 cm (30–50 + in.); total length (height) about 130 to over 220 cm; weight from about 40 to more than 120 kg; one Recent species, *Homo sapiens*; worldwide.

CRITICAL APPRAISAL

The classification of the primates by George Gaylord Simpson is the most widely used list today. Since it was published in 1945 a number of revisions at family level and below have been incorporated; the higher categories have remained unchanged, although they have been the subject of considerable criticism and doubt. W.C. Osman Hill's classification, which, with modifications, follows R.I. Pocock's list published in 1918, has not enjoyed a wide popularity, in spite of the success of "Primates," Hill's monumental treatise on the order. Hill excludes the tree shrews from the primates altogether; he also removes the tarsioids from among the lemurs and lorises and places them with the monkeys and apes. A compromise view between the classifications of Simpson and Hill has been proposed by American paleontologist Alfred S. Romer. Romer also omits the tree shrews from the order and gives the tarsioids equal status with lemurs and lorises, Old World monkeys, New World monkeys, and apes in a separate subordinal group. It seems likely that before very long Romer's classification, or one closely similar to it, will become generally accepted. Reference to the annotated classification given above (based on Simpson's) will indicate how ill-fitted is the living genus *Tarsius* to share subordinal status with the prosimians.

Several revisions of lower taxa (genus and species) have been proposed in recent years which are proving generally acceptable. It has been suggested that *Cynopithecus niger*, the Celebes black "ape," be included within the synonymy of *Macaca*, in which case this species therefore becomes *Macaca nigra*. Also widely accepted is inclusion of *Comopithecus* (the hamadryas baboon) in the synonymy of *Papio* (the common baboon), but opinions differ as to whether the hamadryas and the common baboon are not better considered a single species in view of the good evidence of hybridization in the wild. Synonymy of the genus *Mandrillus* with *Papio* has also been suggested, largely on the evidence of molecular data from blood protein studies, but has not found much support. Less controversial is the conclusion, based on field evidence, that the three generally accepted species of savanna monkeys (*Cercopithecus aethiops, C. sabaeus* and *C. pygerythrus*) actually comprise a single polytypic species, *C. aethiops*.

Less universally acclaimed is the suggestion by Simpson that the structural and behavioral differences between the chimpanzee and the gorilla are insufficient to justify continued generic separation. According to this view, which has some support, *Pan*, having nomenclatorial priority, should include two species, *Pan troglodytes* (chimpanzee) and *Pan gorilla* (gorilla). At a recent symposium on the Old World monkeys held in 1969, the synonymy of *Nasalis* (the proboscis monkey) with *Simias* (the Pagai Island langur) and *Rhinopithecus* with *Pygathrix* was recommended but, as yet, is by no means universally accepted. Field, museum, and laboratory work has led to numerous changes in classification of primates by subspecific level that need not concern us here.

BIBLIOGRAPHY

General works: J.R. and P.H. NAPIER, *A Handbook of Living Primates: Morphology, Ecology and Behaviour of Nonhuman Primates* (1967), a comprehensive, technical book with one or more photographs of each genus; W.C. OSMAN HILL, *Primates: Comparative Anatomy and Taxonomy*, 8 vol. (1953–70), also technical and heavily illustrated, treating the primates in greater depth; W.E. LE GROS CLARK, *The Antecedents of Man*, 3rd ed. (1971), a readable scientific account of nonhuman primates; F. WOOD JONES, *Arboreal Man* (1916), a classic work that contains much useful information; I.T. SANDERSON, *The Monkey Kingdom* (1957), an extensively illustrated popular book; I. DEVORE (ed.), *Primate Behavior* (1965); and C.R. CARPENTER, *Naturalistic Behavior of Nonhuman Primates* (1964), two technical works on primate ethology; W.W. HOWELLS, *Mankind in the Making*, rev. ed. (1967), an informative book on human evolution for the general reader; S. ZUCKERMAN, *Functional Affinities of Man, Monkeys, and Apes* (1933), a technical dissertation on primate structure.

Tree shrews: W.E. LE GROS CLARK, "The Myology of the Tree-Shrew (*Tupaia minor*)," 1924:461–497; "On the Brain of the Tree-Shrew (*Tupaia minor*)," 1924:1053–1074; "On the Skull of *Tupaia*," 1925:559–567; and "On the Anatomy of the Pen-Tailed Tree Shrew (*Ptilocercus lowii*)," 1926: 1179–1309, all in the *Proc. Zool. Soc. Lond.* (1924–26), four important papers on tree shrews; R.D. MARTIN, "Treeshrews: Unique Reproductive Mechanism of Systematic Importance," *Science*, 152:1402–1404 (1966); and M.W. SORENSON and C.H. CONAWAY, "Observations on the Social Behaviour of Tree Shrews in Captivity," *Folia Primatologica*, 4:124–145 (1966), two technical papers.

Tarsioids: W.E. LE GROS CLARK, "Notes on the Living Tarsier (*Tarsius spectrum*)," *Proc. Zool. Soc. Lond.*, 1924:217–223 (1924); and H.H. WOOLLARD, "The Anatomy of *Tarsius spectrum*," *ibid.*, 1925:1071–1184 (1925), technical papers concerned with the relationships of tarsiers; H. SPRANKEL, "Untersuchungen an *Tarsius*. I. Morphologie des Schwanzes nebst ethologischen Bermerkungen," *Folia Primatologica*, 3: 153–188 (1965), a summary, in German, of information on the taxonomy of tarsioids.

Lemuroids: J.J. PETTER, *Recherches sur l'écologie et l'éthologie des Lémuriens malgaches* (1962); "Ecological and Behavioral Studies of Madagascar Lemurs in the Field," *Ann. N.Y. Acad. Sci.*, 102:267–281 (1962); and "The Lemurs of Madagascar," in I. DEVORE (ed.), *Primate Behavior* (1965), technical works on a variety of lemurs; A. JOLLY, *Lemur Behavior* (1967), a readable, semipopular account, based on personal observations; A. WALKER, "Patterns of Extinction Among the Subfossil Madagascan Lemuroids," in P.S. MARTIN and H.E. WRIGHT (eds.), *Pleistocene Extinctions* (1967), a technical discussion of lemur evolution.

New World monkeys: R.I. POCOCK, "On the External Characters of the South American Monkeys," *Proc. Zool. Soc. Lond.*, 1920:91–113 (1920); and C.R. CARPENTER, "Behavior of Red Spider Monkeys in Panama," *J. Mammal.*, 16:171–180 (1935), two technical articles; L.A. ROSENBLUM and R.W. COOPER (eds.), *The Squirrel Monkey* (1968), a detailed account of the biology of one species; A. CABRERA, *Catalogo de los mamiferos de America del Sur* (1957), a broad treatment, in Spanish, of the taxonomy and distribution of South American monkeys; P. HERSHKOVITZ, "Mammals of Northern Co-

lombia; Preliminary Report No. 4: Monkeys (Primates), with Taxonomic Revisions of Some Forms," *Proc. U.S. Natn. Mus.*, 98:323–427 (1951), a scientific treatment of some South American forms.

Old World monkeys: R.I. POCOCK, "The External Characters of the Catarrhine Monkeys and Apes," *Proc. Zool. Soc., Lond.*, 1925:1479–1579 (1925); and "The Monkeys of the Genera *Pithecus* (or *Presbytis*) and *Pygathrix* Found to the East of the Bay of Bengal," *ibid.*, 1934:895–961 (1934), two articles providing much technical information on the langur; C. G. HARTMAN and W.L. STRAUSS (eds.), *The Anatomy of the Rhesus Monkey, Macaca mulatta* (1933 and 1961), a detailed account of the morphology of this important species; J.R. and P.H. NAPIER (eds.), *Old World Monkeys* (1970), a technical but comprehensive account of the catarrhines; N.C. TAPPEN, "Problems of Distribution and Adaptation of the African Monkeys," *Curr. Anthrop.*, 1:91–120 (1960), a scientific account of the zoogeography of the catarrhines.

Anthropoid apes: C.F.M. SONNTAG, *Morphology and Evolution of Apes and Man* (1924), a comprehensive and technical book, with many illustrations; R.M. and A.W. YERKES, *The Great Apes* (1929), a classic work, written for the general reader; C.R. CARPENTER, *A Field Study in Siam of the Behavior and Social Relations of the Gibbon (Hylobates lar)* (1941, reprinted 1967); G.B. SCHALLER, *The Mountain Gorilla* (1963); and J. VAN LAWICK-GOODALL, *My Friends, the Wild Chimpanzees* (1967), three readable accounts of the lives of individual species, based on extensive field work; V. REYNOLDS, *The Apes* (1967); and R. and D. MORRIS, *Men and Apes* (1966), well-illustrated popular books; C.P. GROVES, "Population Systematics of the Gorilla," *J. Zool.*, 161:287–300 (1970); and *Gorillas* (1970), respectively, a technical and popular work on this interesting ape; W.K. GREGORY, *The Anatomy of the Gorilla: The Studies of Henry Cushier Raven, and Contributions by William B. Atkinson and Others* (1950), a carefully illustrated, technical book.

(J.R.N.)

Primitive, Folk, and Popular Arts

The underlying unity of all man's artistic activities, at every level and in every time and place, has, during the 20th century, been widely assumed. Nevertheless, four kinds of art (each one too heterogeneous to be usefully called a style) have continued to be recognized, partly on the basis of their asserted qualities and largely on that of their supposed, intended, or actual audiences. On the one hand, with relatively small audiences, is the kind variously referred to as elite, mature, sophisticated, serious, or high—*i.e.*, the kind normally evoked when the word "art" is used without qualification. On the other hand, with relatively large audiences, are the nonelite kinds usually referred to as primitive, folk, and popular.

The four kinds of art

In this article the elite kind of art will be considered only for the sake of comparisons. The primitive, folk, and popular kinds, in that order, will be examined in terms of (1) definitions, (2) typical productions, (3) artists and patrons, (4) materials and techniques, (5) meanings and functions, (6) styles, (7) special problems, (8) external relations, and (9) appreciation and study. The grouping of the three kinds will permit comment on their occasional zones of overlap and, more important, on the way an understanding of the characteristics of one kind can shed light on the characteristics of the others. The whole discussion will be on the level of theory, introductory explanation, and general description, with specific material included only as examples.

Specific primitive arts are treated elsewhere in comprehensive articles on each of the various arts and in those on primitive cultures; single folk and popular arts are treated in similar contexts and in the separate articles FOLK VISUAL ARTS; FOLK MUSIC; POPULAR DANCE; POPULAR LITERATURE; POPULAR MUSIC; and POPULAR THEATRE.

PRIMITIVE ART

Definitions. In the literature about visual arts and about painting in particular, "primitive" has three common applications that are not wanted here and that arguably ought not to be wanted anywhere. First, with the sense "crude," the term is loosely applied to artists who do not achieve, for a variety of reasons, 19th-century European academic correctness; thus children, madmen, certain modernists, and ordinary bunglers are sometimes called primitives. Second, with the sense "naïve" and usually with the connotation "charming," the term is applied by sophisticated critics and dealers to untutored painters who make up in earnestness of ambition and in freshness of vision for what they lack in professional training; thus in the 20th century the French customs official Henri Rousseau and the U.S. housewife Anna Mary (Grandma) Moses have been called primitives and so have, to give earlier examples, a fair number of talented, ingenuous artists, many of them itinerant painters, who flourished in the United States roughly between 1800 and the Civil War. Third, with a sense that has gradually dwindled almost to zero, the term has been applied by art historians since at least the 1830s to European painters (rarely to other visual artists) of the 14th and 15th centuries; thus a refined, lyrical Sienese master such as Simone Martini may be called an Italian primitive, and a learned, courtly virtuoso of Bruges, Belgium, Jan van Eyck, may be called a Flemish primitive. In this third usage the original implication was that International Gothic and early Renaissance painting was "immature" by comparison with the art of the High Renaissance, when such aids to illusionism as mathematical perspective, the naturalistic distribution of light and shade, and the exact rendering of anatomical detail were well in hand. The growing unfashionableness of biological metaphors for the history of styles has made this implication unacceptable to most scholars but has not kept the bearer of the implication from surviving as one of the numerous fossils in the language of art analysis.

Because of the above three usages, it is necessary constantly to stress the fact that primitive art, in the sense that has been adopted by specialists in the subject, is not a product of some sort of ineptness and not a stage on the way to something presumably better. It is the fully developed art of primitive societies. And this definition is not as evasive as it may seem, for primitive societies can be described with reasonable precision: they lack written languages, cities, monumental centres, road networks, aqueducts, highly organized trade, and technology above the level of hand tools; they subsist on wild-food gathering, hunting, fishing, livestock raising, and small-scale agriculture; they are usually warlike, but without standing armies; their political organization is usually, although not quite invariably, tribal; their religions are directed more toward the successful integration of man with nature than toward the imposition of a humanly conceived order on the cosmos and on the course of universal events. Their rates of change, in all things, are normally so slow as to be practically imperceptible.

The fully developed art of primitive societies

From this description it follows that "primitive" is not at all appropriate for the art of such complex ancient societies as those of the Sumerians, the Bronze Age Chinese, the dynastic Egyptians, the pre-Classical Greeks, and the pre-Columbian Aztecs, Mayas, and Incas. The term, it would seem, may be legitimately applied to the art of prehistoric Paleolithic and Neolithic Europe and Asia, and in fact many ingenious theories try to explain prehistoric art—cave paintings, for instance—by analogy with relatively recent primitive art or relatively recent primitive art by assumptions of diffusion from prehistoric centres. But many authorities are wary of such analogies and assumptions, and in any case, of course, the amount of prehistoric art that has survived is very limited. The upshot of the whole situation is that the bulk of what professional students regard as primitive art dates from the 18th, 19th, and early 20th centuries and comes from three principal sources: African tribesmen who live south of the Sahara, especially those in the western and central parts of the continent; the Indians and Eskimo of North America; and the native inhabitants of Polynesia, Melanesia, Micronesia, and Australia. The generalization needs to be modified only slightly by reminders that diggers sometimes find much earlier material and that scattered primitive societies have been found in many places —in isolated parts of South America, for instance, in Indonesia, and in the back uplands of Asia.

Typical productions. Any attempt to grade the importance of the various primitive arts must be made against

the realizations that there are abrupt regional differences, that some elite arts have no exact primitive equivalents, that some productions are more perishable than others, and that some have been neglected by anthropologists.

Importance of sculpture in primitive visual arts

Thus, it may be safe to say that sculpture, including the carving of ritual masks, is or has been the most important of the visual arts, in terms of both quantity and quality; the statement can be supported by hundreds of brilliant examples produced by the western and central Africans, the Melanesians of the Sepik River region, and American Indians of the Northwest coast. Yet it must not be forgotten that sculpture was a minor art among most of the Polynesians (in spite of the great stone faces on Easter Island), the East Africans, and the Plains Indians. Painting and drawing rank high when one thinks of such things as the prehistoric Lascaux cave painting (if that is acceptable as primitive), Bushman and Australian rock art, Maori tattooing and house decorations, and Navajo sand designs; yet, on the whole, primitive two-dimensional arts, unlike primitive sculpture, do not have the freedom and the specific status of the major creative arts in the elite category. The importance of primitive architecture as an art has probably been underestimated, partly because most of its productions have perished and partly because the surviving ones have not been sufficiently studied. Yet there are fascinating opportunities for aesthetic analysis in the shapes, structures, and spaces of such primitive shelters as the conical-roofed hut of Africa, the lashed-together council house of the Maori in New Zealand, the barrel-vaulted longhouse of the Iroquois Indians in the U.S. Northeast, the multistoried edifice of the Pueblo peoples in the U.S. Southwest, and even the familiar tent, or tepee, of the Plains Indians.

Among the nonvisual and the performing primitive arts, all of whose productions were in theory ephemeral (although in fact quite well preserved by memory and tradition) before the arrival of the white man with his writing, notation, and eventually films and tapes, the first place should undoubtedly be assigned to music, both vocal and instrumental, and to music's practically inseparable sister art, the dance. Literature takes the form of recited prose or poetry the subject matter of which springs from the myths and legends of the tribe. The existence of primitive drama has been mildly questioned, on the grounds that what anthropologists sometimes call plays are actually ceremonies or rites, but the same point might be made about certain early forms of elite drama. Particularly elaborate examples of this sort of primitive drama have been noted in what is now Nigeria, in British Columbia among the Kwakiutl Indians, and in New Ireland in what are called the Malagan ceremonies.

Artists and patrons. Contrary to a common misconception, the primitive artist was generally a professional. (The rapid disappearance of old tribal customs makes the past tense advisable in most of this discussion.) In many places he had to earn the right to practice his art by serving an apprenticeship; in other places he could inherit the right and, with it, a religious status. The system might not prevent an ordinary tribesman or even a tribeswoman from doing some dancing, singing, and decorating.

The professional primitive artist

But important projects, especially in the visual arts, were normally reserved for the professional, and sometimes his whole creative procedure was protected by something like a sacred copyright; in Polynesia, for example, the accepted notions of taboo made it dangerous for an unauthorized person to attempt certain activities. The professional, however, was not free, in the manner of elite Western artists since the Romantic era, to produce what he liked and then look for a market. Usually he was commissioned by a knowledgeable, exigent patron to do something according to traditional rules. The patron might be a chief, a tribal elder, a person of lower rank in need of a specific work, or quite often a representative of one of the secret societies that existed in many parts of the primitive world and took charge of particularly important communal ceremonies.

Materials and techniques. The primitive visual artist was extremely dependent on local materials. If he was an Eskimo, he was not encouraged to carve in wood. If he was a Polynesian, the absence of suitable clay kept him from becoming a potter. But the dependence did not have a simple negative effect; it gave the artist a deep respect for the natural qualities of each material and made him alert to the artistic possibilities in many materials—bark, sand, feathers, shells, skins, bones, teeth, and the human body itself—that are traditionally ignored by creators in the elite category. In a somewhat analogous way the lack of efficient equipment and complex technical processes encouraged mastery of the tools and methods that were available. The widespread notion that primitive art is on the whole more crudely executed than the elite kind cannot survive an unprejudiced look at the evidence.

Meanings and functions. The fact that primitive societies are by definition nonliterate has to be taken into account in any consideration of the meaning of a work of primitive art; for, in the absence of writing as a way of storing facts and values and of communicating them across space and time, one may find these vital cultural jobs being partly done by a little wooden statue, an apparently simple piece of music, or a traditional dance. To be aware of this possibility is not to argue, as a few experts have, that the tribes of Africa, North America, and the South Pacific never indulged in sheer entertainment or pure decoration or art for art's sake. But there can be little doubt that one of the essential differences between primitive art and the other three kinds, all of which exist in literate cultures, is that the primitive kind is on the average more densely charged with symbolic meaning and, in a broadly social and religious way, more functional. The difference, although the analogy should not be pressed very far, is rather like the difference between medieval Christian and Renaissance art.

A few examples may help to clarify the point. In primitive sculpture the material itself is frequently charged with meaning; the Iroquois Indians cut some of their masks from living trees in the belief that the masks would keep some of the power residing in nature. The forms have to be read in detail and as wholes; in Polynesia a creative deity might be carved not only with a symbolically large penis and with a number of tiny offspring swarming over his body but also, to make the significance triply certain, with the overall shape of a phallus. Primitive painting shows the impulse to rate meaning as far more important than mere representing; and there are no exact primitive equivalents for such elite genres as landscape, still life, and portraiture. All of the primitive arts were at one time functional in a religious, magical, social, or economic way; they served to placate ancestors and gods, to assure food supplies, to knit the tribe together, and to accompany daily work. Often the functions were extraordinarily specific; the Tutsi of Rwanda, in East Africa, had songs to praise cows, songs to show that one ought to have cows, songs to take cows home in the evening, songs to water cows, songs for children about cows, songs to admire the royal cows, and songs to emphasize the importance of cows in history.

Functions of the primitive arts

Styles. It is perhaps worth saying again that primitive is not, properly speaking, a style. Within it, as might be expected in a kind of art produced by communities living in relative isolation, there are many sharply definable styles. Anyone with a little experience of African sculpture in a good museum soon comes to recognize the schematized, nearly abstract style of the Kota (Bakota) tribesmen of Gabon, the relative naturalism of the Fang, who are nevertheless neighbours of the Kota, and the classical, almost European naturalism of the ancient inhabitants of the town of Ife, in Nigeria. Among Pacific styles the streamlined severity of Micronesian objects is instantly distinguishable from the complex expressionism of Melanesia. In primitive literature there is a style—and sometimes several—for each of the hundreds of languages and dialects. Anthropologists and ethnomusicologists are familiar with a vast number of musical styles that range from the rhythmic polyphony of black Africa to the radical simplicity of the Sirionó Indians of Bolivia, who have no instruments and only a few two-tone or three-tone tunes, one of which is usually adopted by a singer for his entire lifetime.

It is easy to suppose that there is a relationship between style and general cultural pattern or between style and the social function of a work of art; interesting attempts to verify the supposition and systematize it have been made. Naturalistic styles, for instance, have been said to be associated with nomadic ways of life and geometric styles with settled ways. The degree of power and sacredness in primitive masks has been said to match an increasing degree of abstraction and fantasy in the styles. But not enough evidence has been sifted to make such assertions definitely acceptable, and there are always exceptions to generalizations about style.

Special problems. Considerations of style are involved in four special problems that are problems of general anthropology but that are capable of arousing particularly strong emotions among students of primitive art—partly because the solutions advanced can imply fundamental assumptions about the nature of man and of universal cultural history. The problems, inadequately but conveniently labelled, are those of evolution, diffusion, environment, and aesthetic psychology in relation to primitive art.

The problem of evolution in primitive art

The first problem has triple roots in the facts that primitive often means early, that the Darwinian theory of evolution made some injudicious converts, and that Europeans during the last half of the 19th century, when primitive art began to be seriously examined, were inclined to have a high opinion of themselves and their cultural achievements. The result, in its rawest form, was the belief in a world unilinear evolution, or progress, of art (along with other cultural elements) from an infantile, fumbling stage up to an adult, fully developed stage represented essentially by Renaissance and post-Renaissance European creations. Manful efforts were made to set forth the intervening stages and to answer such questions as which came first, naturalism or geometrism. Art that was discovered in the mostly tribal societies of Africa, America, and Oceania was classified as unevolved and as representative of some of the early stages through which Western elite art had passed. The whole theory has crumbled during the 20th century under the growing evidence of the immense variety and technical excellence of much primitive art, under the attacks of modern artists on post-Renaissance conservatism, and under the conviction that, at least in art, value judgments are relative. Anthropologists and art historians, as was pointed out earlier in this article, now agree that primitive art is not in a formative stage in relation to elite art. It is simply the art of societies that are unlike the societies that provide the audiences for elite art—and also for the folk and popular kinds.

But perhaps this conclusion does not really dispose of the problem. Granting that "evolution," with its connotation of improvement, is an undesirable term in the context and granting that 19th-century Western prejudices are no longer tolerable, many scholars nevertheless feel that the history of the various primitive arts is a legitimate field for speculation and research. One can assume (and sometimes back the assumption with archaeological evidence) that these presumably mature arts have changed, however slowly, in the past. One asks how and why they changed, how they became what they were when Europeans discovered them.

The problem of diffusion

Such questions are part of the background of the diffusion problem. Nearly all experts, of course, are willing to accept the hypothesis of diffusion from a creative centre when similar traits appear in the arts of neighbouring regions over a reasonably short period of time; no one doubts that some of the traits in the art of the natives of New Guinea derived from Indonesia. The problem arises when the traits appear in cultures that are otherwise quite different and that are separated by great expanses of time and space—sometimes by many centuries and by thousands of miles. Some ancient panpipes found in what is now Peru, for example, are tuned in the same way as some much more recent ones found in Oceania, and quite similar spiral motifs appear over a period of thousands of years in the arts of Europe, Asia, Melanesia, and North and South America. Diffusionists tend to assume that transmission, perhaps by accident, must have occurred at some unknown moment; and their assumptions are apt to be plausible in proportion to the complexity of the elements in question: the similarity of tuning systems, for instance, is harder to explain as independent invention than is the similarity of simple spirals. Antidiffusionists in such cases tend to stress whatever makes the supposed transmission improbable: the length of time involved, the difficulties of travel, and especially the lack of buttressing evidence of cultural contact. Why, they might ask, if certain stylistic motifs were diffused from the Old World to South America, did the potter's wheel fail to be diffused?

To reject, as a majority of anthropologists do, the extreme diffusionist position and to favour the idea of independent invention is to raise the environment problem, with "environment" stretched enough to mean traditional, almost natural tools as well as landscapes, climates, raw materials, and local food supplies. When primitive art is examined in the context of the environment that produced it, a critic may ask what the relationship is between a stone hatchet and a particular kind of sculpture, between the frozen landscape of the north and an Eskimo artist's conception of space, between big game and rock art, between forests and naturalism. Clearly, similar questions may be asked about any kind of art, but they are particularly important in the study of the primitive kind because of the isolation of primitive societies and their close contact with nature. And the answers are not always as obvious as they may at first seem.

The "environment" problem

The problem of aesthetic psychology may arise in connection with almost any aspect of primitive art, but it arises with a special force if it is suspected, as many investigators have suspected, that the minds of primitive artists and of their publics do not work in quite the same way as the minds of elite artists and publics. It has been argued, for instance, on the basis of incomplete but interesting evidence, that primitive societies do not possess a concept of art or that at least they do not possess anything comparable to the art concept of nonprimitive societies. One wonders then, whether they can be properly said to produce art. Again, it has been argued that the nearly abstract and schematic representations common in primitive visual art are, supposedly along with the drawings of children, "ideoplastic" in character; in other words, the artist depicts what he knows or thinks and not what he actually sees. If this is so (and, according to Gestalt psychology, it may not be), many of the criteria used in elite art are not appropriate in the primitive kind.

External relations. Evolutionism, diffusionism, environmentalism, and "ideoplasticism" are matters for the scholarly interpreter of primitive art. A problem of a different sort is posed by what happens when a primitive society undergoes modifications through contact with theoretically more advanced societies—when it undergoes, to use the technical term, acculturation. With a few temporary exceptions, the local arts in such situations have declined. They have proved to be too closely integrated with the total culture to survive. Even the forms, regarded as pure shapes and patterns, have tended to wither away, like organs deprived of function. Occasionally the forms have had a stimulating and liberating effect on elite artists, the best known example being the effect of African sculpture on European Cubists such as Pablo Picasso. Occasionally, too, the forms have been given a new content, often Christian or Western secular, in their old localities. But the effect, as a glance at the evidence can reveal, of primitive art on 20th-century elite art has been exaggerated; and the achievements of the new native workshops have been debatable. The sad rule is that primitive art, unlike the elite, folk, and popular kinds, cannot have external relations that go beyond the exchange of a few motifs.

Appreciation and study. Works of primitive art, long regarded as mere curios and then as material for anthropology, have slowly acquired the aesthetic status they deserve. Public collections of the visual productions exist in such institutions as (to name only a representative sample) the Museum of Primitive Art, the American

Museum of Natural History, and the Museum of the American Indian, Heye Foundation, all in New York City; the Field Museum of Natural History, Chicago; the British Museum, London; the Australian Museum, Sydney; the Musée de l'Homme, Paris; the Museo Preistorico Etnografico Luigi Pigorini, Rome; the Rijksmuseum voor Volkenkunde, Leiden, The Netherlands; the Museum für Völkerkunde und Vorgeschichte, Hamburg; Staatliches Museum für Völkerkunde, Munich; Museum für Völkerkunde und Schweizerisches Museum für Völkerkunde Basel, Basel; the Etnografiska Museet, Stockholm; and the museums installed in the new non-European nations. Primitive literature can be found, too rarely, in anthologies of translations; primitive music, often labelled as folk music, can be found on commercial records and in many museums on tape.

FOLK ART

Definitions. Much as a search for a usefully limited meaning for "primitive art" may be hampered by 19th-century prejudices, so a search for a merely descriptive, reasonably factual sense for "folk art" may be hampered by 19th-century sentiments. For, at about the same time that notions of "primitive" were being affected by the theory of unilinear world artistic evolution and by imperialist self-satisfaction, notions of "folk" were being affected by a fondness for the picturesque, by generous impulses toward aesthetic democracy, by the so-called arts and crafts movement, by historicism and philosophical Idealism, by nationalism and regionalism, and by an occasionally aggressive, wildly mystical sort of ethnocentrism. Thus, folk art was defined, if only by the implications of usage, sometimes as almost anything that could be considered quaint; sometimes as all non-elite art, primitive and popular included; sometimes as whatever seemed vaguely homemade; sometimes as the art of the common people, but with the latter regarded as a ghostly entity existing outside of real class structures; and sometimes as a traditional, characteristic art that preserved a cultural heritage and somehow represented the collective soul of a nation or a province.

19th-century notions of folk art

There are some more or less acceptable elements in such definitions. What is missing is a flat enough acknowledgment that folk art is the recognizable product of a nonruling, relatively unaffluent social class and that references to it except in relation to the elite art of a ruling class are meaningless. With this acknowledgment in mind, one can see that talking about a folk kind of art in primitive societies is usually not worthwhile, even though such societies have often had strongly marked class structures. Aristocracies that existed in Polynesia, particularly in Hawaii, did not have a distinct kind of art; they simply had their choice of the better artists. Royal courts in black Africa, with the exception of one at Benin, did not have court styles, although they sometimes insisted on having "noble" materials—bronze, for example, instead of wood or clay. In theory it is possible legitimately to talk about folk art in the great civilizations of the ancient world; but in practice the arts of these civilizations have usually been regarded, for good reasons, as aesthetic wholes. In modern, technologically advanced societies of the 20th century, one encounters the problem of distinguishing between folk and popular, which will be considered below in this article. For the moment it may be enough to point out that Detroit factory workers and New York secretaries are not commonly thought of as belonging to the class called the folk. Nor are city dwellers in general, including those of eras before the 20th century.

A core description of folk art

Putting all these considerations together, one arrives at what can serve as a core description. Unmistakable folk art is the art of a class of peasants, herdsmen, seamen, artisans, and small tradesmen living as a rule away from urban centres in societies that, while literate and highly civilized, are not highly industrialized. Such societies were to be found in Europe, particularly eastern Europe, until well into the 20th century, and they will still exist, of course, on other continents. Where they no longer exist, the term folk may still be appropriate for the art of social or ethnic minorities that have preserved earlier traits by living apart, culturally and often physically. More doubtfully, the term may be extended to certain traditional artistic productions—especially those associated with carnivals, national holidays, religious processions, and the like—that are the work of thoroughly unfolkish city people. But to go much further, to label as folk art everything from an American Indian buffalo robe to a comic strip, is to obliterate a valuable set of distinctions.

Typical productions. Although the folk kind of art is invariably produced, according to the only definition that makes serviceable sense, in a culture that also produces the elite kind for a ruling class, the usual elite classification of the arts into the fine, or independent, and the useful, or decorative or dependent, is pretty much ignored by folk visual artists, who are typically willing to devote their creative imaginations to the design or decoration of tools, toys, furniture, cottages, clothes, arms, banners, musical instruments, and so on. The portrait of a Spanish patron saint may appear in embroidery; narrative pictures belonging, according to the European academic classification, to the elevated "history" category are painted on Sicilian carts. Something of the same mixture of the so-called high and low is typical of folk literature, music, and dancing: tragedy takes ballad form and a noble war dance the form of a Highland fling.

Artists and patrons. In some ways the usual folk artist can be called an amateur. He lacks professional training, he normally has a second, more remunerative occupation, and he creates primarily for his own pleasure and needs or at most for those of acquaintances in his own small locality. But in other ways he is quite unlike an amateur artist in a nonfolkish social context. He is apt to work strictly within the limits of a traditional formula, he often remains anonymous, and often he is an artisan who carries over into his works of art his everyday skill. Amateur is not quite the word for a blacksmith who designs a fantastic shop sign in wrought iron nor for a baker who turns out figurines.

Materials and techniques. Works of visual folk art are rarely executed in anything except readily available, fairly inexpensive materials. A sculptor is likely to work not in marble or bronze but, rather, in wood, iron, clay, straw, ice, sugar, or whatever else is at hand; a painter may substitute an old board for the elite artist's canvas or special paper. Perhaps because of this situation and perhaps because of the already mentioned tendency to ignore the separateness of fine art, the folkish creator displays little of the elite (or, for different reasons, the primitive) concern to employ a specific technique for a specific material—to respect what a French art historian has called the vocation of each sort of artistic material. Folk wooden sculpture may be carved in a manner suitable for granite; painting may be done in an approximation of embroidery technique; pottery may bear incised patterns borrowed from metalwork or basketry; and almost any material is apt to be disguised under bright colours.

Meanings and functions. Although the folk, defined as a class, live in nonprimitive societies, they are more or less isolated from the urban centres in which cultural change may be relatively rapid; and in their predominantly rural environments they are close to nature and to the basic birth-to-death cycle of human existence. Hence folk art is strongly conservative and preservative; in fact, it is often referred to as simply traditional art (with the unfortunate implication that other kinds of art do not possess traditions). It acts as a repository for proverbial wisdom, ancient superstitions, sentimental themes, and religious beliefs that may have long since ceased to be orthodox; it accompanies and celebrates what is repetitive in the lives of individuals and of communities—baptisms, marriages, funerals, anniversaries, sowings, reapings, and the daily routine of work. Often it preserves customs the meanings of which have been largely forgotten; the colouring of Easter eggs and the decorating of Christmas trees are familiar examples. Often, too, the old meanings are given an unexpected twist by a naïve or eccentric artist. But, on the whole, folk art functions within its sphere, as does primitive art in an entire society, on the

Conservative and preservative qualities in folk art

side of continuity and social stability; and there is some justification for 19th-century talk about the expression of a collective "soul."

The "soul," however, is not so much that of a nation, a race, or a province as that of a social and economic class. The best folk art reflects the common sense, unpretentiousness, humour, and peculiar slyness of humble people in contact with small, hard realities; and not infrequently it reflects hostility toward the rich and powerful. In folk depictions of the Nativity, a gift bearer may bring not frankincense and myrrh but, instead, a jacket for the baby Jesus; in a harvest song, one may suddenly be informed that "the gentlemen," not the corn, are about to be mowed.

Styles. Encouraged by local patriotism, scholars have divided folk art into mostly national and regional styles: English, French, German, or Spanish folk music, Breton, Provençal, or Balkan costumes, Pennsylvania Dutch decorations, Mexican processional figures, and so on. This sort of classification is justified by the remarkable traditional and settled nature of much of folk art. But it should not be allowed to obscure, as it frequently does, the existence of other sorts of style, nor should it suggest that these national and regional manners are without histories. There are Jewish, Christian, Islāmic, and other religious folk styles; mountain, valley, and seaboard styles; classical, romantic, realistic, fantastic, and expressionistic styles; and naturalistic and abstract styles. There are period styles; in fact, a sizable number of vaguely ahistorical-looking provincial folk styles are arrested styles of the 18th and early 19th centuries.

Special problems. Diffusionism, environmentalism, and related theories have aroused polemicists in the field of folk art as well as those in primitive art. But perhaps stronger feelings can be, or have been, stirred among folk-art specialists by theories that attempt to answer the two rather loaded questions of how folk art is created and, second, what the difference is between a folk art and a folkcraft.

Theories about the creation of folk art

The first question has some of its roots in the facts that the adjective folk may connote "communal" and that folk art is often the work of unknown creators. Other, probably more important, roots lie in the fact that, by the beginning of the 19th century, German Romantic thinkers were committed to myths (at that time fairly harmless) about the collective spirit, or virtue, of a nation or a race. The result was the idea, set forth with varying degrees of explicitness, that folk art—in particular, poems, melodies, and dances—was a communal creation. This idea was eventually replaced by a theory of communal re-creation or communal growth in situations in which the work of an individual poet, singer, or dancer is preserved only in memories and is modified as it passes from locality to locality and from generation to generation. In the meantime another largely German idea, that of cultural sinking, was provoking argument. According to this idea, folk art—visual art in particular—is derived, after a lapse of several decades or even several centuries, from the elite art of a ruling class; thus a Crucifixion carved by a 19th-century Breton peasant is explained as a late-provincial imitation of the 17th-century Baroque of Paris and Versailles. On all these matters the opinion of a majority of modern scholars is apt to be conciliatory. Although collective improvisation is not impossible, the original creator of a work of folk art is practically always a single artist. But it is possible to grant that a good deal of communal re-creation occurs in literature, music, and the dance and that in the folk visual arts the role of the individual creator, subject as he is to the demands of tradition and function, is not as decisive as it is in elite art. Again, it can be granted that folk art is often a belated borrowing from elite. But this need not prevent recognition of the fact that the folk as a class have their own traditional themes and forms.

Difference between a folk art and a folkcraft

The question about the difference between a folk art and a folkcraft is partly philosophical (which is not a reason for neglecting it) and partly quite practical, at least for museum curators who must classify exhibits, for specialists interested in a proper division of labour, and for modern educators or animators who want to continue the 19th-century arts and crafts movement. It has been argued that, where an artist creates, a craftsman merely fabricates; that, whereas the first is a nonmaterialist who makes something for the sake of expression or of disinterested contemplation, the second is a materialist who makes something for a definite function; and that, whereas the first cannot know exactly what he is going to produce, the second has an extremely precise foreknowledge accompanied by a plan and a method of execution. On this argument many alleged works of folk art can be dismissed as works of mere craft. But the argument is challengeable, partly on the familiar ground that art is in the eye of the beholder.

External relations. Although folk art, as was pointed out above, may borrow from elite and although it may also lend motifs to everybody from classical musicians to modern dress designers, it has proved incapable of surviving in the world of 20th-century industry and communications. Like primitive art, it can live only in isolation. Unlike primitive, however, it has demonstrated, particularly in the performing arts, an extraordinary ability to transform itself into a special branch of popular—which will be discussed below.

Appreciation and study. Many of the museums of primitive art listed at the close of the previous section also have collections of folk works. Other material can be found in such institutions as the Metropolitan Museum of Art (American Wing), the New York Historical Society Museum, the Brooklyn Museum, and the Jewish Museum, all in New York City; the Library of Congress (Folklore Section) and the Smithsonian Institution, both in Washington, D.C.; the Southwest Museum, Los Angeles; various museums of American state historical societies; the Royal Ontario Museum, Toronto; the Museo Nacional de Artes e Industrias Populares, Mexico City; the Cuming, Horniman, and Geffrye museums, London; the Royal Scottish Museum and the National Museum of Antiquities of Scotland, Edinburgh; the National Museum of Wales, Cardiff; the Musée des Arts et Traditions Populaires, Paris; the Norsk Folkemuseum, Oslo; the Museu de Arte Popular, Lisbon; the Museo de Artes Populares, Barcelona; the Museo Nazionale delle Arti e Tradizioni Popolari, Rome; the Museum für Völkskunde (Staatliche Museen), Berlin; the Folk Art Museum and the National Museum of the Peoples of the U.S.S.R., Moscow; museums in all the Balkan capitals; the Crafts Museum, New Delhi; the Folkcraft Art Museum, Tokyo; and the Canton Peasants' Institute, Canton. Folk art in something like original surroundings is on view in such restored or reconstructed places as Colonial Williamsburg, Virginia; Greenfield Village, Dearborn, Michigan; the Pueblo Español in Barcelona, Spain; Pérouges, France; Skansen, Stockholm; and the "villages" that are a regular feature of world fairs. National and provincial folk music and literature have been extensively collected on records and in books. Folk dancing has been revived, or simply preserved, in many countries.

POPULAR ART

Definitions. As might be gathered from the names of some of the museums mentioned above and from countless books, articles, and conversations, it is not likely that art officials, critics, historians, and ordinary appreciators will soon be using "popular" in a very limited sense. Since the word can mean merely "commonly approved, or widely liked," it may be applied to the "Mona Lisa," the ballad "Barbara Allen," Pueblo pots, and television shows. It, or its Romance-language equivalent, has been employed frequently by English scholars, and quite regularly by French, Italian, Spanish, and Portuguese scholars, to designate what is called "folk" in the present article. In Germany the situation has been somewhat reversed, the word for "folk" being employed to embrace what many Americans are inclined to think of as "popular." Style-term coinage has added to the confusion: around 1960 the expression "Pop art" became current in the United States and Europe for what, in spite of an iconography derived from such sources as billboards and

comic strips, was distinctly a phase of elite modernism, not a whit more "commonly approved, or widely liked" than any other prominent avant-garde movement. Cultural improvers have contributed: in many places "popular" theatre does not mean an absence of elite classics but simply a lowering of admission prices. Politicians have intervened: in some totalitarian regimes "popular" means what is "commonly approved" by the government.

The examples could be multiplied. In the realm of common usage, there is not any point in even attempting to limit the sense of such a term. But there is indeed a point in the context of serious discussion of art. For there is clearly a kind of art that is neither elite nor primitive nor folkish. It is a kind whose central characteristics can be discerned, however foggy its remote frontiers may be. And the best adjective for it, in the long perspective of history and in the mind of a contemporary observer, is certainly "popular."

Description of popular art

Popular art, in this restricted sense, has probably existed ever since people began to crowd into cities and to form readily accessible mass markets; temple figurines of ancient Mesopotamia might be classified as popular, if one knew a little more about the economics of their production and distribution. But the category emerges into prominence and becomes easily distinguishable from the folk category only with the progress of reproductive technology, first with the invention of printing and then with the development of engraving, lithography, photography, the phonograph, motion pictures, radio, television, and the latest audiovisual equipment. Thus popular art can be summarily described as art intended to be popular in the usual sense—commonly approved or widely liked—among the common people in a literate, technologically advanced society dominated by urban culture. That excludes, on one ground or another, elite, primitive, and folk art. In the qualification "intended" lies the heart of the matter. The actual audience may not count at all in determining how a particular work should be classified. The sale of 1,000,000 records cannot transform a Beethoven sonata into popular art, in this restricted sense; and the box-office failure of a film concocted for millions of viewers does not transform it into elite art.

Typical productions. Although among the older surviving genres of popular art one must cite dancing, fireworks, street spectacles, circus routines, and several other sorts of almost wordless entertainment, the long-term popular trend, at least in Western countries, has been noticeably literary. This is evident not only in the success of popular fiction but also in the directions taken by nonliterary disciplines once they enter the popular category. Painting as such, for example, is relatively unimportant; what the popular audience wants are pictures that tell stories—in woodcuts, lithographs, photographs, comic strips, films, and television dramas. Music as such cannot, of course, be called unimportant; during the 20th century it has become, with the help of records, perhaps the most popular of the popular arts. But it is predominantly music with words.

Artists and patrons. There have been popular artists, including many in the 20th century, who have expressed deeply felt private convictions about the meaning of ordinary human experience; there have been not a few who have employed their talents in defense of what they believed to be justice or the cause of the common man. There have been some who have refused to grant that the customer was always right. A generalization that leaves such artists out of consideration is unfair. Nevertheless, it

Role of the popular artist

can be said that on the whole the popular artist is less committed, both to his work and to any particular set of beliefs, than are his opposite numbers in the elite, primitive, and folk categories. Unlike elite artists, he seldom starves in an attic for the sake of his art. Unlike primitive and folk artists, he is not necessarily integrated with his cultural context, and he is not encumbered with a duty to propitiate occult forces and to preserve tribal or community continuity. He is primarily a professional entertainer. He is constantly tempted to express not what he himself thinks but what he—or his immediate employer—thinks the intended audience thinks.

Materials and techniques. In the fairly remote past the materials of popular art were about the same as those of folk art, and surviving works suggest that popular artists often had a folkish lack of the primitive and elite concern to adopt specific techniques for specific materials. But the popular artist of more recent times, if he happens to be of the visual sort, is seldom in direct contact with the material of his finished creation; he is merely the first stage in a process of multiplication for a large market. He does not normally engrave the plate, run the press, or cast the statuette. Something similar to this distancing of actual execution occurs in other sorts of popular art, particularly those that have flourished in the 20th century. The singer's material and technique are partly his own and partly those of amplifying or recording equipment; the film maker's are partly people and processes beyond his firm control. The same situation may occur, naturally, in elite art, but there the creator is usually in a better position to assert himself, and so the aesthetic consequences are apt to be different.

Meanings and functions. Since popular art exists in a pluralistic society quite unlike the unified contexts of primitive and folk art and since it is normally intended as light entertainment or simply as a means of making money, all attempts to generalize about its meanings and social or cultural functions may seem futile as well as humourless. Yet it is perhaps almost as "functional" as primitive art and hence notably different, at least in this respect, from elite "contemplative" art. Popular love stories have been one of the bulwarks of monogamy in the Western world for several centuries. There can be little doubt that certain varieties of popular music, powerfully helped by modern publicity, provide young people with a satisfying conviction of belonging to an ingroup and that this conviction makes the intrinsic quality of the music a secondary matter (which is not to say that the quality is always poor). Nor can there be much doubt that a popular television show, like a Melanesian tribal ceremony, may function as a ritual to promote community cohesiveness (which is not to reflect adversely on television viewers or on Melanesians).

Styles. The "functionalism," however, of popular art differs radically from that of the primitive kind, as well as from that of folk art, by being tied, especially under modern technological conditions, to comparatively rapid changes in period style—often to short-lived fashions. In primitive Africa, America, and Oceania and in the backcountry of such a folk province as 19th-century Brittany or Pomerania, there was little that could be properly called fashion; styles might change over long periods, but they were essentially matters of tribe, race, or locality, with variations according to sex, age, class, and occupation. In popular art the situation is more or less reversed; styles are essentially matters of time. And, if 20th-century trends in mass communications, politics, sexuality, and social stratification continue, style differences other than the temporal may disappear entirely from popular art. By the 1970s the styles of popular music were already nearly universal.

Special problems. The majority of commentators would probably agree that all the special problems of popular art concern quality: how an artist can aim at a large market without aiming at the lowest common denominator; how the distancing of execution can be reconciled with the personal touch that is a usual requirement for good art; how local flavour can be preserved as globalism grows; how change itself can be kept from becoming the principal popular style as world history accelerates; and where the point is at which popular art ceases to be art and becomes an instrument for something else. By the last third of the 20th century, these perhaps too familiar, but nonetheless pressing, questions were being regularly discussed by intellectuals, by enlightened public officials, and in what was often significantly referred to as "the industry." Answers, none of which were very confidently offered, ranged from resigned proposals to let-consumers-work-things-out to exhortations to launch a cultural and hence partly economic and political revolution. Sometimes the reformers apparently overlooked the plain

fact that good popular art—the good song hit, cowboy-and-Indian film, comic strip, television series, advertisement (if that qualifies as art), social dance, magazine story, or item of house decoration—might be less rare than pessimists had alleged.

Sources of popular art

Unlike primitive and folk art, the popular kind thrives on technological advances (to which it partly owes its existence), on imported ideas, and on influences from the other aesthetic categories; and it does so not only in the sense of being able to continue an ample production. It has successfully borrowed, for instance, a large number of ideas from both traditional and modern elite painting, for use in illustrations, cartoons, magazine layouts, printed advertisements, and window displays. The latest device of an avant-garde film director is almost certain to appear, after a delay of a few years, in family movies and on the television screen. Perhaps the most effective tricks in the popular fiction of a given decade are borrowings from the elite fiction of a generation or two earlier. But the best example of this sort of aesthetic appropriation is the already mentioned incorporation and transformation of folk music; there is a vast amount of popular song and popular dance music that owes its charm and vitality to old ballads, exotic traditional tunes, and rustic rhythms. Often the content as well as the form may be adopted and adapted; thus a peasant's complaint about the Hundred Years' War may turn up as a protest song of the 20th century.

Scholarly interest in popular art is much more recent than that in the primitive and folk kinds, and much of it has been strictly sociological. Collecting has been neglected, for the evident reason that the material has seemed, until too late, to be everywhere. Many museums of folk visual art, however, have preserved examples of early popular, usualy from the 18th or 19th century. Old popular songs have been recorded or re-recorded. And of course the serious student of contemporary popular has only to look and listen.

BIBLIOGRAPHY

Primitive art: FRANZ BOAS, *Primitive Art*, new ed. (1955), stresses environment, techniques, and functions. LEONHARD ADAM, *Primitive Art*, rev. 3rd ed. (1954), defines his subject broadly enough to include both prehistoric and children's art. PAUL S. WINGERT, *Primitive Art* (1962), investigates physical, psychological, and social backgrounds and closely analyzes styles in Africa, Oceania, and North America; the result is an excellent general introduction. ANDRES LOMMEL, *Prehistoric and Primitive Man* (1966), stresses diffusion. Two useful anthologies offering a variety of viewpoints are DOUGLAS FRASER (ed.), *The Many Faces of Primitive Art* (1966); and CHARLOTTE M. OTTEN (ed.), *Anthopology and Art* (1971). CLAUDE LEVI-STRAUSS, *La Pensée sauvage* (1962; Eng. trans., *The Savage Mind*, 1966), provides fresh insights into the mentality behind primitive art. ALAN P. MERRIAM, *The Anthropology of Music* (1964), helps to correct the tendency of many anthropologists to emphasize visual material. Primitive literature is usually examined in works on folklore.

Folk art: MAMIE HARMON, GIUSEPPE COCCHIARA, and ALESSANDRO MARABOTTINI MARABOTTI, "Folk Art," in the *Encyclopedia of World Art*, vol. 5, col. 452–483 (1961), present a broad, book-length theoretical survey focussed on visual art. A comprehensive work focussed on literature is MARIA LEACH (ed.), *Funk and Wagnalls Standard Dictionary of Folklore, Mythology and Legend* (1949). BRUNO NETTL, *Folk and Traditional Music of the Western Continents* (1956), offers much general theory in the course of treating his specific subject.

Popular Art: NORMAN JACOBS (ed.), *Culture for the Millions?* (1964), presents opinions expressed in a wide-ranging symposium on the situation in the United States. MARSHALL MCLUHAN, *Understanding Media* (1964), probes in a controversial way the effects of technology. Of general interest but with a British slant are the chapters on the 20th-century best-seller and on mass communications, respectively by P.N. FURBANK and RICHARD HOGGART, in BORIS FORD (ed.), *The Modern Age* (1963). For more specific treatments, see STUART HALL and PADDY WHANNEL, *The Popular Arts* (1964); and DAVID EWEN, *History of Popular Music* (1961).

Bibliographies: Particularly useful reading lists will be found in the above studies by, or edited by, Wingert, Otten, Harmon, and Hall and Whannel.

(R.McMu.)

Primitive Law

Whereas civilized law is the law of societies whose cultures include writing and archaic law is the law of early civilizations, primitive law is the law of nonliterate peoples. Studies of civilized and archaic law are centuries old. Primitive law is a recent field of study, pursued to some extent by lawyers with an interest in comparative legal systems but primarily by anthropologists. Increasingly this field has also drawn the attention of political scientists and government administrators, especially those working in newly formed nations made up of several previously autonomous tribes or societies.

The general findings of anthropological jurisprudence show that all societies have legal systems. In all known societies disputes are governed by definitely understood, though frequently unstated, rules—with machinery and personnel to enforce them. Offenses are variously defined and handled from society to society, with a few "crimes" being almost universal. There are differences, also, in types of penalties and in the selection of persons responsible for the administration of justice. All legal systems, primitive and modern, are based on the social need to keep the group effectively and reasonably harmonious. All are dynamic and constantly changing in response to basic changes within the culture, although different societies show varying degrees of capacity for adaptation.

CULTURE, CUSTOM, AND LAW

Law distinguished from custom

Law is an aspect of culture. Culture as defined by anthropologists is the integrated sum total of learned behaviour traits characteristic of the members of a society. It constitutes a people's lifeway; it shapes their world view, modes of food getting, personal beliefs, government, religion, law, forms of housing, and technology. Through culture, each human society develops standardized modes of action on which individual behaviour is to be patterned. Such patterns of behaviour, called social norms, are popularly known as customs.

Customs or social norms are sanctioned; they are supported by rewards or penalties for conformity or nonconformity. While the potential range of human activities is enormous, organized social living requires restriction of the range of behaviour that may be allowed expression. Controls, therefore, are instituted to induce people to act in accordance with prescribed norms.

Law is a limited and special part of the system of social control, but it is never the entire system. This is an important and frequently overlooked point. Any discussion of law must take for granted a host of extralegal mechanisms such as child rearing, education, religious taboo, and other controls over behaviour. All are more fundamental than law. Law comes into play when there is disagreement among the members of the society over relations, rights, and duties or when individuals ignore the rights and expectancies of others to such a degree that formal steps are required to restore disrupted relations and to discourage repetition of the behaviour in question.

Although law differs from custom, not all students agree on what the distinctions are. Perhaps it is enough to say that law employs economic or physical sanctions in support of selected norms. The law has teeth that can bite. It consists of the network of social norms from which deviation—beyond that which may be allowed by the society—is controlled by the exaction of economic or physical penalties, according to norms of due process, by persons exercising the authority to carry out such processes. The norms to be enforced constitute the rules of substantive law, whereas those governing procedure and legitimacy constitute the rules of due process and allocation of authority. Substantive law comprises what is to be supported and enforced by the law; the rules of due process define how it is to be enforced. The allocation of authority determines the possessor of the legitimate authority to enforce such rules. It is this principle of legitimacy that separates legal force from mere violence.

SUBSTANTIVE AND PROCEDURAL LAW

Legal offenses

The content of substantive law varies tremendously among primitive societies. Two features, however, are

almost universal: homicide within the group is legally punishable everywhere; adultery, with a few possible exceptions, also comes universally within the scope of legal action. It should be noted, however, that what constitutes illegal homicide and adultery varies among different peoples. Among the Eskimo, the killing of infants, old people, and invalids, for example, is legally privileged if committed by a relative or close friend. Among these people sexual intercourse between a married woman and a man to whom she has been lent by her husband is not illegal adultery; on the other hand, it is definitely illegal if the husband has not given at least his implicit permission.

Sorcery, or detrimental witchcraft, is also a widespread object of legal control among primitive societies. Apparently, sorcery is seen as a form of vicarious homicide and aggression against the well-being of another.

Theft, interestingly enough, is not always a subject of major concern in primitive law. Among primitive peoples, access to economic resources is generaly open to all on an equal footing. Furthermore, food and implements tend to be shared freely. Substantive law among primitive food-gatherers and hunters is, therefore, largely concerned with such direct personal offenses as homicide, bodily assault, sorcery, and adultery rather than with goods. Among peoples who have achieved a developed level of farming or animal husbandry, however, the law of property and inheritance burgeons. In such economies there is the development of capital, or productive goods (such as plows or mills), which may be borrowed, lent, purchased—and misappropriated. Entry into contractual relations is common, and unlawful conversion (theft, misappropriation, fraud, etc.) and indebtedness are major causes of litigation in more complex primitive societies.

As cultures and environmental conditions change, new issues and conflicts necessitate different means of resolution. The introduction of the horse from Europe wrought great cultural transformations among the Indians of the American plains. Earlier norms among the Cheyenne had allowed the free borrowing of goods without permission, providing the borrower left an identifiable personal item as notice and security. Eventually, because horse borrowers could move great distances among the camps, this privilege led to abuses and resulted in the lodging of formal complaints by horse owners. This ended the practice of borrowing of horses without asking. An old general norm permitting borrowing was legally modified to accommodate new cultural circumstances.

Sanctions

Primitive legal systems focus much more heavily on ameliorating ruptured social relationships than on exacting vengeance or punishing wrongdoers. Imprisonment is an invention of civilization and has no place in any known primitive society. The death penalty, it is true, is often imposed for homicide, sorcery, and a few other extremely "dangerous" acts, such as incest. Among North American Indians, the Comanches could kill their wives or cut off their noses for adultery or desertion. Cheyennes could expose adulterous wives to mass rape. Adultery with the wife of the paramount chief of the African Ashanti was punished by an excrutiatingly long, drawn-out surgical death. Pueblo Indians severely beat minor offenders and executed serious deviants. The henchmen of Trobriand Island chieftains executed upstarts who insulted the chief. Exile is used in some tribes and some situations as a substitute for death, but rituals may be used to bring an exile back into the community.

Most legal actions in the majority of primitive societies end in a public restatement of norms and obligations and a symbolic reaffirmation of such obligations through the payment of economic damages. Valuables—be they cattle in Africa; pigs in Melanesia; fine mats in Polynesia; horses and guns among Plains Indians; shell money in Micronesia, Melanesia, and California; or yams in the Trobriands—are tokens of social esteem and rewards for social achievement. When one is deprived of "valuables" to pay damages or fines, he loses social esteem.

Thus, although physical coercion may be the ultimate legal sanction, the more usual sanctions are the exaction of damages in private law (between individuals) and of fines in public law (between the individual and society).

It has often been written that one attribute of law is its universal applicability. This is not entirely true of primitive legal systems, in which the total situation influences the outcome much more than is ideally allowed in Western law systems. Legal norms are rarely absolute; the range of deviation from prescribed norms is often called the "range of permissible leeway." Within this range, some combinations of circumstances and personalities will result in legal sanctioning whereas others will not. In private law, the esteem in which a kinship group holds each individual member is often crucial. Has he fulfilled his kinship duties well and willingly? Does he lend strength to the group? Or is he a shirker and troublemaker? The group will go all out for the valued member; it may not lift a hand for the black sheep. Permissible leeway maintains an elment of uncertainty in law and allows for adaptability in effecting justice and maintaining society.

Permissible leeway

ALLOCATION OF AUTHORITY

In primitive legal systems the authority to initiate and carry through legal actions is commonly vested in the injured party, supported by his kinship group. The private prosecutor is, for the particular occasion and for that offense, an official representative of social authority, even though he wears no constable's badge, policeman's uniform, or judicial robe. A Comanche husband whose wife had run off with another man could himself, for example, take formal legal steps to collect damages from the absconder; if he was too weak a person, a kinsman or other brave warrior could act as prosecutor for him. Although there might be much pushing and hauling in the negotiating process, every recorded Comanche case of adultery and wife absconding in the 19th century resulted in the payment of damages by the offender; and no chiefs entered into these cases.

Such social allocation of authority to private individuals and kinship groups is a primitive form of private law. In its extreme form, it exists in a number of hunting and gathering tribes as a "self-help" process without the involvement of mediators. Usually, however, it operates with a "third party," a person or group of persons who enter the dispute-settlement process as negotiators, mediators, arbitrators, or judges. The negotiator, or mediator, like the Ifugao *monkalun* in Luzon in the Philippines, facilitates settlement through persuasion. He may, however, threaten to throw the whole force of his powerful family against a disputant who violates the norms of procedure. Or he may, like the priest of the earth among the African Nuer, threaten to curse the whole kinship group that refuses to accept a reasonable settlement of a serious case. Failure to come to terms in a dispute involving a just claim that has been legitimately prosecuted may finally lead to physical compulsion. It has been reported that among the Ifugao, if the dispute is not resolved, the culprit may be slain, especially if a *monkalun* has been involved; in such a case, public opinion will usually support the action and the slain person's family usually will not retaliate.

The role of judge

The role of judge among primitive societies is usually implicit in chieftainship; indeed, decision-making authority, in both the executive and the judicial capacity, is one major attribute of "chief" and "king" as opposed to the "headman" of systems that are not politically centralized. Chiefs are infused with the symbolic identity of the corporate wholeness of their societies and as such their authority is endowed with legitimacy. The whole society may be so symbolized in the person of the chief that offenses against the person of the chief no longer are seen as private wrongs but as crimes against the society. Offenses increasingly are looked upon as breaches of the king's peace. The scope of royal courts may widen. Centralized political and judicial authority may develop, based on factors other than the dominance of a kinship group. Among the Cheyenne Indians, for example, such authority was vested in a tribal council of chieftains. The heads of military fraternities also directed legal actions on behalf of the entire tribe. Among the Pueblo Indians in the southwestern U.S., the central legal authority resides in a

high priest, cacique (chief), and in a council of ranking priests of the religious fraternities.

The foregoing discussion has dealt with legal systems that are self-determining. Each of them has been developed over time by a primitive people for its own social purposes in consonance with its own culturally defined goals. The law of such systems may be called *organic*.

There have been, however, a number of primitive states in Africa, Asia, and Middle America that have lived by conquest. Conquered tribes were incorporated into a sociopolitical structure dominated by the conquerors, who often imposed legal restraints upon the subject peoples that were not of the latter's making. Such law is *tyrannic*. By definition, tyrannic law consists of a network of regulations imposed by a social group possessing the raw power to establish norms and to enforce them. Power is seized; authority is not allocated as a socially recognized privilege or right.

Feud and law

Although primitive legal systems have well-developed concepts of due process, these do not function very well in the area of private law. Private law involves direct negotiation by parties in disagreement, and in primitive law the defendant must agree to some form of punishment or restitution or there can be no settlement. If the defendant is recalcitrant, perhaps to unreasonable demands, procedure breaks down. The defendant may deny further suit and possibly incur physical attack by the plaintiff. If the defendant has indicated that he rejects the legitimacy of the claim, he will deny the plaintiff's right to resort to force. Feud or internecine war may result.

Feud is a state of conflict between two kinship groups within a society, typified by a series of illegal killings and counterkillings between the kinship groups, usually initiated in response to an original homicide or other grievous injury. Legal historians traditionally have seen primitive society as marked by a constant state of feud, as rent by violent retaliation and blood revenge, and, in general, as an arena of violence ruled by the law of the jungle and the retaliative lex talionis ("eye for eye, tooth for tooth").

There is very limited evidence for this. Legal procedures or ritual devices such as regulated combat that are used as a means of avoiding or terminating feuds have been universally found to exist in such cultures. Even when feud is a reported institution in primitive cultures, there is usually little or no empirical evidence that it actually takes place. An example would be the case of the Trobriand Islanders, for whom it was reported that although honour made "vendetta" obligatory in cases of homicide, it was evaded by the substitution of blood money (compensation). There was much talk of feud but no evidence of specific cases.

If there are strong grounds for doubting that feud occurs with anything like the frequency so often supposed, there is no question but that the folk belief in feud is widespread. The belief is a folk phenomenon in its own right. Real feud is socially destructive and dangerous. When it does occur, it represents a breakdown of the technique of peaceful resolution of conflict. Among the Northwest Coast Indians of North America, the threat of violence was not always enough to insure the success of negotiations; often feud was the result.

Law exists everywhere to preclude the recourse to bloody battles within societies. Folk belief in the expected requirements of honour to engage in feud is often countered by a stark recognition of the effects of feud. It seems quite likely that this very fear of feud is one of the factors that nourishes and sustains primitive legal systems.

THE EVOLUTION OF LAW

The greatest contrast between primitive and civilized law is in the area of procedure and is based largely on differences in social structure. The simpler primitive societies are characterized by decentralized units of organization lacking overall political authority. Each kindred, lineage, or clan has considerable autonomy, and headmen and elders function as leaders and guides to the group's activities. When several kinship groups exist in a village, the elders may form a village council, and one of them may also serve as village headman or chief. The scope of such village or tribal authority is, however, usually limited when compared to the autonomy of the kinship groups.

In primitive law the primary authority for most legal actions is the kinship group working in support of its individual members. Hence, private law dominates most primitive law. At the same time, almost every primitive society recognizes some acts as legally punishable by the society at large or through explicitly delegated officials. Illegal acts punished by public prosecution are crimes, as against the torts of private law.

The major trend in the development of law from the most primitive societies to the most civilized is in the direction of a shift of the right to prosecute and impose sanctions; this right is transferred from the wronged individual and the kinship group to officials representing the general society. As cities emerged in the Mediterranean area, urban migrants representing a huge variety of tribal legal systems became detached from their tribal moral orders and kinship groups. The need for law generated elaborate legal systems promulgated and managed by the ruling elite of priests and warriors. The legal person became *citizen* rather than *kinsman*.

Four functions of law

All legal systems perform four functions essential to the maintenance of societies and their cultures. First, they determine obligatory relations among the members of a society, making clear which activities are permitted and which are ruled out. Second, they allocate authority and designate who may initiate and carry through legitimate corrective actions leading to economic or physical sanctions when norms are violated. Third, they dispose of troublesome cases as they arise so that social harmony may be re-established. Fourth, they continually redefine relations among the members of the society as the conditions of life change, in order to maintain adaptability in the substantive and procedural law.

Primitive law and modernization

Primitive legal systems have been strongly affected by the impact of modern legal systems. As of 1970 only a handful of primitive societies in the Amazon Basin and perhaps a hundred or more in the interior of New Guinea had not been brought under the nominal administrative control of some national state. Colonial and national administration of law with respect to tribal societies has varied considerably in policy and application.

In Africa, the British ruled by "dual mandate" through the first half of the 20th century. In native affairs indigenous headmen and chiefs were permitted to administer their own courts in accordance with tribal law and custom (except for the death sentence) and in cases that were not contrary to public policy or to "natural law and justice." Their decisions were subject to appeal and review by British authority. In later years, when people like the Nuer and Tiv had no indigenous courts, the British provided for their organization. Systems of folk law were on the whole retained and kept in reasonable consonance with the values and forms of tribal social organization.

In the Dutch East Indies, adat (customary) law was likewise continued for Indonesian village communities.

In India, following the Mutiny of 1857, law was codified for British India upon the model of English common law, modified somewhat to accommodate the social conditions of India. Religious and family law was reserved to Hindu and Muslim law, according to the religion of the disputants.

In the French colonies, the French civil and criminal codes were said to be universally applied.

In the United States, from 1832 to 1953, all Indian tribes were legally recognized as "domestic dependent nations," not subject to the laws of the states in which they were located. Except for 11 major crimes brought under federal jurisdiction by acts of Congress, each tribe retained its own legal jurisdiction over its members as long as they were within reservation boundaries. In 1968, the Civil Rights Act for Indians introduced the Bill of Rights into Indian tribal procedure for the first time.

The importance of anthropological study of tribal law systems for administrative and legislative guidance of national lawmakers should be self-evident. Much damage to family and community life could be spared if national law were adapted to the living law of the various peoples that make up the nation. The urge to use law as a unifying

force in newly developing nations raises the problem anew: it may be seen in every postcolonial nation of Africa and Asia, and it reaches bizarre proportions in the Territory of Papua and New Guinea, where a population of 2,300,000, composed of several thousand indigenously independent peoples, speaks an estimated 700 distinct languages and dialects.

World law as primitive law

International law has been termed primitive law on the level of world society. What passes as international law consists of a considerable body of substantive rules for the conduct of affairs between nations, recognized by tacit concensus or agreed upon in conventions, pacts, and treaties or verbalized in arbitration awards, rare World Court awards, and United Nations resolutions. But this body of norms has as yet no supporting sanctions other than withdrawal of reciprocal services, damage to national reputation, the threat of feud (war), and the waging of war. Force and the threat of force have not yet been brought under control by and for organized world society.

BIBLIOGRAPHY. H.J.S. MAINE, *Ancient Law: Its Connection with the Early History of Society and its Relation to Modern Ideas* (1861, paperback ed. 1963), the classic English-language work on law and society; L. KRADER (ed.), *Anthropology and Early Law: Selections from the Writings of Paul Vinogradoff and Others* (1967), late-19th-century selections in historical jurisprudence; BRONISLAW MALINOWSKI, *Crime and Custom in Savage Society* (1926), on reciprocity as the basis of legal obligation in the Trobriand Islands; E.A. HOEBEL, *The Law of Primitive Man: A Study in Comparative Legal Dynamics* (1954), a theory of law in terms of culture and function, with detailed analyses of six primitive sociolegal systems ranging from very simple to complex, and with K.N. LLEWELLYN, *The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence* (1941), an analysis of Cheyenne Indian law by a lawyer and anthropologist; L. NADER (ed.), *Law in Culture and Society* (1969), compendium of ten modern case studies of non-Western legal systems and three of Western legal systems; M. GLUCKMAN, *The Judicial Process Among the Barotse of Northern Rhodesia* (1955), case materials and stimulating theoretical discussions of the comparative implications of the law of the Lozi tribe, and *The Ideas in Barotse Jurisprudence* (1965), a wide-ranging comparative analysis of Barotse legal concepts and their relation to Barotse legal action; P.J. BOHANNAN, *Justice and Judgment Among the Tiv* (1957), case studies among a northern Nigerian people as observed in tribal courts; L.J. POSPISIL, *Kapauku Papuans and Their Law* (1958), a field study of an isolated tribe of New Guinea gardeners; P.H. GULLIVER, *Social Control in an African Society* (1963), on dispute settlement outside the courts among the Arusha tribe of northern Tanzania.

(E.A.H.)

Primitive Religion

Because of the misleading connotations of the term primitive, most present-day scholars hesitate to classify any religious phenomena as primitive. Nevertheless, for lack of a satisfactory substitute, "primitive religion" still appears in the title of many books and articles. In a study by E.E. Evans-Pritchard, an outstanding contemporary anthropologist, *Theories of Primitive Religion* (1965), primitive religion means the religion of peoples "living in small scale societies with a simple material culture and lacking literature." The term is generally used in this article in that sense, but other meanings that have been attached to it are also considered, particularly those that refer to the elements and basic forms of religious experience and behaviour, whether historical or psychological. Those beliefs and practices that have been labelled "primitive" are still of central interest to students of religion in general and, in many cases, are still very much alive today.

NATURE AND SIGNIFICANCE OF "PRIMITIVE" RELIGION

Importance and influence of the study of primitive religion. The religions considered "nonprimitive"—*e.g.*, Judaism, Islām, Christianity, and Buddhism—are characterized by literate theological and philosophical traditions, even though their adherents are often neither literate nor educated in the literature of their faiths. These religions, however, represent only a fraction of the total religious experience and expression of mankind, which have been mainly "primitive." "Primitive" religion then claims attention as representing the religions of most human beings for most of human history; and, though followers of supposedly "advanced" religions have often looked down on these "primitive" faiths, the more philosophically and sociologically minded have also sought to learn something essential about the nature of man and divinity from them.

The basic form of human religion

During the 18th- and 19th-century confrontations between a religious and a scientific view of the world, when the validity of all religious belief became a matter of general dispute, the nature of the primitive religion that supposedly underlay and preceded all formal creeds became a subject of central interest. Primitive religious belief and practice could be invoked either to demonstrate that all religion was based on delusion or that since men had always been religious, religion must relate to the deepest realities of human nature and of the universe.

Thus, many writers whose influence is still widely diffused in intellectual life made use of information about primitive religion. In theology it was notably represented in the voluminous writings of a German priest and anthropologist, Wilhelm Schmidt (1868-1954), and his school, in which it was ultimately held to corroborate orthodox Christian doctrine. In political and social theory, Friederich Engels drew upon the religions of North American Indians in *The Origins of the Family, Private Property and the State* (1902), and Karl Marx must have been acquainted with such information. His view of the essentially social nature of religious sentiment has been widely shared by students of primitive religion. Sigmund Freud and Carl Jung, the outstanding innovators in psychology, brought primitive myth and symbolism into the mainstream of psychoanalytic theory, Freud's *Totem and Taboo* (1913) in particular basing its argument upon what was then known of Australian Aboriginal religion. Through classical scholars like Jane Harrison and Gilbert Murray, the living religions of then remote societies were brought to bear on the interpretation of Greek and Roman texts; and the distinguished Sanskrit scholar and philologist Max Müller referred to African and other beliefs in his works on mythology and language. In creative literature the characteristically modern writers of this century—*e.g.*, T.S. Eliot, Ezra Pound, D.H. Lawrence, and W.B. Yeats—received some inspiration from accounts of primitive religious experience; and in more popular literature, writers of adventure stories—*e.g.*, Henry Rider Haggard, perhaps the best informed—exploited anthropological sources with varying degrees of sensationalism, as they still continue to do. Behind the use all these writers made of anthropological knowledge were scholars who collected, organized, and sifted information, especially in the English-speaking world (see below *Development of the study of primitive religion*).

Effect on imaginative literature

The problematic meanings of the term primitive and suggested substitutes. The word primitive has had pejorative overtones, intensified by 19th-century reconstructions of human evolution that placed primitive man low in a scale of human development from prehuman ancestry to contemporary civilization. Thus used, the term constantly confuses what is historically primary with what is psychologically immature, the early with the crude or simple, and for this reason various substitutes are often used in modern writings. Thus "ethnic," tribal," "preliterate," "preindustrial," "small-scale," and now "underdeveloped" and "developing" appear where once "primitive" would have unhesitatingly been accepted. Any one of these may well be somewhat more suitable in a particular study of a particular people. In a short general survey one must retain the older term, but it may still be necessary to state that, for serious students of the subject, primitive religion, far from connoting "savage" superstitions, exhibits the very basis of man's religious constitution, to which the professional theologians and philosophers have but added their own systematic rationalizations. Thus, although William James in *The Varieties of Religious Experience* (1902) has nothing to say about the religions considered below, he shares something of the spirit in which they are now studied in pleading for more attention to be paid to the religious experi-

ence of the common man, as against the highly intellectualized theological constructs of professional thinkers. Attempts to rediscover in primitive religion forms of experience that seemed to have been superseded by scientific materialism and dogmatic religious orthodoxies have also characterized the middle years of this century, two examples being the writings of Henri Frankfort and Mircea Eliade.

Observations of travellers

Development of the study of primitive religion. From classical times observations on the customs, including the religions, of the "barbarians" and "pagans" occur in anthropo-geographical writings, the most famous being those of Herodotus; but these classical and later and fuller medieval references reveal more about the outlook of their authors than about primitive religion. By the Renaissance, travellers' observations on cults and beliefs from distant parts of the world had been further assimilated, if still without much real understanding. What was known of primitive religion was fitted into the biblical and Christian context of belief, where it appeared either as a special category of religion, idolatry, or, as in Sir Thomas More's *Utopia* (1516), was introduced more sympathetically to suggest what "natural" religion without the Christian revelation would be like. The idea that primitive religion started with idolatry, or more precisely with the worship of animals and material objects called by Portuguese voyagers "fetiches," was widespread into the 18th and well into the 19th centuries.

In the 18th century also, with such work as the French Jesuit missionary Joseph-François Lafitau's comparison of the North American Indians with the ancient Hebrews and Egyptians (*Moeurs des sauvages amériquains comparées aux moeurs des premiers temps*, 1724), debate arose as to whether extant primitive religions had emerged independently or represented degenerate and fragmentary survivals of the religion of Genesis. Montesquieu, an 18th-century French philosopher, took primitive religion into account in his sociological history, and the British moral philosopher David Hume (in *The Natural History of Religion*, 1757) argued that monotheism had grown with the growth of human reason out of an original polytheism. Hume's account of the origin of religion in the personification of the unknown causes of natural events has lived on in one or another form until the present day.

But it is to the 19th-century scholars, with their vastly greater knowledge of the peoples of the world, that present-day students more particularly look back. They were obsessed with positing one or another origin for religion, with evolutionary reconstructions of human society and psychology that now have a merely historical interest; but their theories at least encouraged argument and the later effort to test them by accumulating reliable empirical evidence. Two writers who made very substantial contributions to the study were the great English scholars Sir Edward Burnett Tylor and Sir James George Frazer. Tylor, regarding the basis of religion as "a belief in spiritual beings" (a more acceptable view than earlier suppositions that primitive men worshipped material idols), coined the term animism for this primitive faith and argued that man had arrived at the conception of spirits by reflecting on the experience of incorporeality found in his own dreams and the withdrawal of a vital principle, the soul at death. Herbert Spencer, who, especially in the first volume of *The Principles of Sociology* (1896), compiled a great deal of information on primitive religion, came apparently independently to the same conclusion. In *Golden Bough* (1890), probably the most famous single treatise on primitive religion, Frazer proposed a vast scheme of psychological and theological evolution, beginning with magic, for him a primitive attempt to coerce nature by technical acts and formulas, which was replaced by religion, involving the propitiation of unknown forces controlling human destiny, and which was in turn being superseded by modern experimental science.

The great English and French scholars

Such were in principle psychological theories of the primitive basis of religion, but in France Émile Durkheim and his followers (as also W. Robertson Smith in Britain) rather emphasized its social and collective nature. In *The Elementary Forms of the Religious Life* (1915), a still indispensable work, Durkheim saw a totemic belief of the Australian Aborigines as those most elementary forms and suggested that gods, like the totemic clan animals revered by the Aborigines, symbolized and expressed social cohesion. These streams of thought, along with others concerned with the interpretation of primitive mentality, myth, and ritual, meet in the writings of later anthropologists. From early in the 20th century and increasingly during it, direct contact with "primitive" peoples by trained observers with a thorough knowledge of their languages has opened up far deeper channels of enquiry into the refinements and complexities of their thought. Among the immediate forerunners of contemporary studies may be mentioned Bronisław Malinowski, A.R. Radcliffe-Brown, Paul Radin, and Robert H. Lowie, of whom the last two wrote general works on primitive religion.

Although his work is probably less often referred to than theirs, perhaps a more pervasive influence in this century was Lucien Lévy-Bruhl, who in several books argued that primitive thought was prelogical and exhibited an interest in mystical participation with the life of the natural world quite unlike the analytic habit of thought of scientific Western man. It would now be admitted that Lévy-Bruhl underestimated the rational and empirical elements in the thought of his "primitives." He drew attention, nevertheless, to the possibility of approaching their religious propositions and symbolism in their own right as forms of mystical thought and not merely seeing them as a failure to develop the mental procedures of European natural scientists. Work done in the field since the 1940s at least has been more and more concerned with elucidating religious symbolism and (under the impact partly of the widely acclaimed contemporary French anthropologist Claude Lévi-Strauss) with the elaborate analysis of belief and myth, in an attempt, ultimately, to deepen understanding of human thought and imagination through study of its spontaneous expressions in primitive life.

RELIGION IN PRIMITIVE CULTURE

Specific characteristics and functions in primitive communities, life, and thought. The societies that most students of primitive religion have been concerned with are, or were, similar in three principal respects. First, as compared with urban and more particularly industrial cultures, they are small in scale, with a corresponding emphasis on face-to-face, personal relations between their members. Local groups of kin and neighbours play a more significant part in national or tribal life than is the case in modern states, and much of primitive religious belief and practice is bound up with the needs and structure of these local communities.

The three common characteristics

Second, with some important exceptions, they have been economically poor, without great distinctions of wealth, and technologically unable to dominate their natural environment. The close economic interdependence of their members and their ultimate dependence on the forces of nature have important consequences for their religions, which emphasize the communal aspects of human life and common subjection of all men to forces beyond their control.

Third, though often rich in oral tradition, they have lacked written traditions and, with them, the accumulative records of private introspection that eventually produce systematic and formalized philosophy and theology. It is often said, therefore, that they are religions of action rather than of reflection, though this distinction can be over-stressed.

Following from the first point, primitive religion has been said to be largely polytheistic, for there are family gods, clan gods, gods of the local community, and other gods associated with features of the locality and with the needs and shared experience of the local community. The effect of their cults is to bind believers in communions based upon kinship and neighbourhood. To take one example, the highland people of New Guinea present a pattern of many local communities, each with its different

gods; but the gods are of a basically similar kind, and belief and practice are generally uniform over the whole territory.

But to say that primitive religion is polytheistic is to disregard other significant and widespread characteristics. Belief in some kind of high god, a transcendent being in whom the nature of divinity in general is conceptually unified, is widely reported, and Wilhelm Schmidt devoted much attention to investigating this aspect of primitive belief. It is often reported that these high gods are too distant from men to concern themselves with local affairs, though a single inclusive concept of divinity emerges when local needs and interests give way to broader national, tribal, and intertribal concerns. It is less usual, however, to find a pantheon with gods or gods and spirits arranged in a fixed hierarchy. The god (or gods) dominant in a particular situation has first place, corresponding to a mode of belief which Max Müller called henotheism (adherence to one god, while acknowledging the existence of others).

The moral sanctions of the gods

The relative lack of centralized coercive secular power in primitive cultures leaves to the gods the important sanctions for correct behaviour. It is not that gods announce moral rules; rather, they support the moral principles traditionally taught by the ancestors with whom they tend to be closely connected. So if a man fails to carry out binding obligations to his kin, his immorality may be corrected by illness, interpreted as divine retribution, more than by other social and merely secular pressures. When the relations between kin form the fabric of local community, this idea of the retributive justice of the gods is a powerful sanction for approved behaviour. Similarly, when comparative and general poverty make for great local interdependence, the belief that the generous man has the blessing of the gods encourages the mutual economic support actually essential for communal survival.

A simple technology and the absence of extensive scientific knowledge mean that many results brought about on the modern world by purely technical acts and events explained in the modern world by tested scientific generalizations are in the primitive world part of the province of religion. Thus, rainmaking ceremonies and rites to promote fertility—*i.e.*, to promote life—are central social concerns. At the individual psychological level, too, religious action gives the reassurance of being able to deal symbolically with suffering and misfortune, of being able to divine truths that men need to know for their own well-being, and of making direct contact, often through spirit possession, with the suprahuman forces believed to affect human health and happiness. Hence the growing study of ethno-psychiatry examines primitive religious belief and ritual seriously with regard to their psychotherapeutic effects.

Various types of religiocultural configurations. In the past many attempts were made to construct typologies of religion, but with increased knowledge it has now become clear that any attempt to divide forms of belief and ritual into mutually exclusive types is bound to do violence to the facts. Thus, such terms as fetishism, animism, totemism, ancestor worship, and shamanism can no longer be regarded as denoting well-defined and distinct modes of religious belief. Rather, certain dominant themes appear in the religions of different societies. So for all peoples contemplation of the dead and, more particularly, of dead ancestors excites some sentiment of piety and often of fear, but the extent to which the ancestral dead are placed at the centre of a religious cult varies greatly from one society to another. An example of a people whose religious outlook is concentrated on the power of ancestral ghosts is that of the Manus of Melanesia, who regard the father's or sometimes some other kinsman's ghost as supervising the affairs of the living household, sometimes protecting and sometimes punishing its members. This Sir Ghost, concretely present in the skull of the person from whom it derives and which his family preserves in the home, is thought in particular to punish sexual transgressions by causing illness. When illness occurs diviners are consulted to discover which Sir Ghost or other ghost is angry. The Manus are fishers and traders, and Sir Ghost also requires the correct fulfillment of economic obligations.

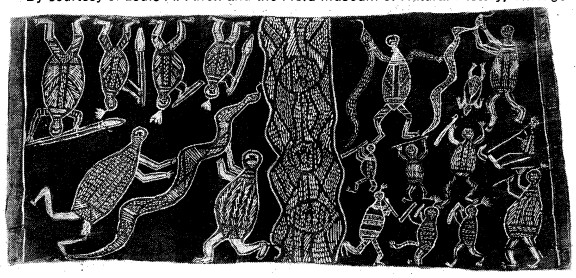
Relation to the dead: ancestral ghosts

A concern with the power of ancestral ghosts, however, is also found among many African peoples whose way of life and systems of morality are very different from those of the Manus and among whom religious service to the dead is combined with cults of nonhuman spirits and gods. The Nuer of the southern Sudan, whose political system is based upon patrilineal descent and who therefore trace their ancestry much further back than do the Manus, have in common with many peoples of the Nile region a belief in the ghostly vengeance of the embittered dead. But here, despite the importance of genealogy, there is no cult of the dead, and graves are not for long remembered. Instead there is a concentration upon an entirely spiritual high god, Kwoth, and many nonhuman spirits who partake of his nature. To these prayers and sacrifice are offered, and diviners are employed to suggest which spirit has caused any particular misfortune. Here the dominant theme of the religion is belief in spirit in numerous manifestations, or "refractions," as they have been called by E.E. Evans-Pritchard. Nuer have little or no interest in the material representation of spiritual forces, in great contrast to the proliferation of images and shrines of the gods found among the Yoruba, the Akan, and many other West African peoples.

Where kingdoms have developed, the cult of kingship, as represented in the king and his ancestors, has often become the dominant theme of religious practice. It was so in ancient Egypt and remains so in many African kingdoms where, just as the kingship itself must not die, the king in person must not appear to die like other men. This sacralization of the king, sometimes as high priest of the nation as, for example, in Ashanti (now in Ghana) and other West African states and sometimes as himself divine, is one of the main interests of Sir James Frazer's *Golden Bough*, which gives examples from many parts of the world and not only from peoples who have been called "primitive."

Among many peoples the commonest cultic practices are centred round certain animals, plants or other species, and sometimes human artifacts. These species are regarded as emblems or material symbols of the spiritual vitality and integrity of particular social groups within the society, and in some cases elaborate myths explain their association with those groups, while ceremonies are performed to further their well-being. This aspect of primitive religion, called totemism (from a North American Indian word) by both past and present writers, was more particularly identified among the Australian Aborigines. There a wide range of species, the totems, are thought to link those who revere them with what they call "the dream-time," which was defined by Adolphus Peter Elkin, an eminent Australian anthropologist, as "the ancient time of the heroes . . . the eternal dream-time of spiritual reality. . . ."

By courtesy of Louis A. Allen and the Field Museum of Natural History, Chicago



Depiction of the Australian Aboriginal "dream-time," showing (left) the rainbow serpent creating the ancestral spirits, and (right) modern clansmen performing a ceremony in honour of the ancestral spirits. Bark painting by the Aboriginal artist Mangangina. Groote Eylandt, Australia.

Individual religious experience

But it is not only through totemic observances that this contact with spiritual reality is sought. In some primitive religions (and again, not only the religions of peoples who have been called primitive) the quest for an indi-

vidual spiritual experience that will provide guidance through the dangers of life might be said to be the dominant theme. This is particularly so among some North American Indian groups, where men will prepare themselves for this significant vision by solitude, fasting, and self-mortification. Sometimes also, as in American Peyote cults, drugs are used to produce this vital experience of the ultimate situation of the individual psyche. Widespread also are cults of spirit possession, in which the worshipper invites ultrahuman, spiritual forces to enter his body and speak through him, both to exorcise injurious spiritual influences from himself and from others and to announce truths beyond rational investigation. Emphasis on possession, ecstasy, and exorcism is characteristic of what has been called shamanism (after the Tungus word for "medicine man" in East Siberia) amongst Amerindians, Eskimos and Mongolians, but it is a widespread and apparently increasing feature of many African religions. Shamanism, however, involves special experiences and techniques in some respects different from those of African mediums (see SHAMANISM).

Salient distinctions from or similarities with the advanced religions. Not only is it impossible to differentiate mutually exclusive types of primitive religion (and still less to place them in an evolutionary hierarchy) but also what has been called "primitive" religious practice and belief has far more in common with "advanced" religions than it has apart from them. The literary heritage of the latter does indeed permit the formation of longer and more varied traditions and the incorporation into them of a wider range of individual opinions, intuitions, and experiences, but the oral traditions of nonliterate peoples (which tape recording makes more fully available than ever before) also display a greater depth, range, and consistency within the particular community than had previously been supposed. To take only two examples, *The Myth of the Bagre* from northern Ghana (Jack Goody, ed.; 1972) runs into some 12,000 lines of verse that are memorized in a fixed form as a recapitulation of man's historical relation to the gods, while the priesthood of the Maori of New Zealand has long been known as the repository of a rich historical and religious tradition. Priesthood, prayer, sacrifice, and asceticism, as well as elaborate mystical symbolism and mythological validation are found equally in those religions which have been called primitive and those which are described as advanced. Doctrinal dispute following from wider confrontations of opinion in religious matters is probably much more characteristic of literate religious traditions than of the rest, and it is widely reported of primitive faiths that they make no claim for the exclusive truth of their own teachings. Nor, on the whole, do they have any doctrinal certainty in asserting what happens beyond this life. Their emphasis lies rather on communal well-being than on the cultivation of individual virtue, and their ethical teachings are often less centrally incorporated in religious lore than those of the more moralistic "religions of the book." And as compared with the literary form in which the spiritual beings of those religions often tend to be presented, primitive religions seek and appear to provide direct experience of the power of the gods. It is, of course, a return to some such primal spirituality that revivalists of the "advanced" religions have often sought.

Tradition, institutions, and doctrine

PRIMITIVE WORLD VIEWS

Distinction between sacred and profane. It is commonly accepted in modern western society that religion is concerned with the supernatural; however, as Durkheim recognized, the notion of the supernatural depends upon a notion of the natural as being governed by scientifically established laws, an intellectual construct of a specialized and historically recent kind. Hence, like Durkheim, many students of primitive religion would still hold that religion is not concerned with the supernatural so much as with the sacred—with sacred periods in time, sacred places, and sacred persons set apart from the ordinary flow of common life for special religious attentions. Primitive (and not only primitive) religion, then, would be concerned with the knowledge of sacredness and its protection, and with distinguishing the sacred from its complementary opposite, the profane. For Durkheim, the sacred represented the collective strength and aspirations of the community, experienced in some special periods of time (as in the religious cycle of a church calendar), in some persons (such as priests and kings), and in some places (such as shrines or centres of pilgrimage). All these are connected with rituals and religious prescriptions and proscriptions that do not apply to the daily sequence of ordinary life, to ordinary men, or to ordinary places, which represent the realm of the profane. Also, of course, "to profane" may have the meaning of actively opposing the sacred and, thus, desecrating or polluting. For a comprehensive discussion of the sacred and profane, see SACRED OR HOLY.

Dynamistic, daemonistic, and theistic views of the sacred. A main point of departure for students of the sacred was the Melanesian and Polynesian concept, mana, first noted by Robert Henry Codrington, an English missionary and anthropologist, in a book published in 1891. Like the words totem and tabu (*tapu*), that indigenous term became part of the technical vocabulary of anthropological and religious studies. Mana signifies the intensity of an impersonal force or power attributed to some persons, things, and situations. Mana, Codrington stated,

> is what works to effect everything which is beyond the ordinary power of men, outside the common processes of nature; it is present in the atmosphere of life, attaches itself to persons and things, and is manifested by results which can only be ascribed to its operation.

Mana, animism, and high gods

The apprehension of this ethically neutral sacredness, a spiritually dynamic power, which in some contexts might readily be translated simply as "luck" and in others perhaps as "strong effectiveness," was regarded as the psychological basis of religion.

Closely connected with mana in earlier discussion was the idea of taboo, indicating in its original Polynesian meaning something set apart or forbidden. R.R. Marett, another British anthropologist, considered at length whether mana was a magical or a religious conception and whether taboo was "negative magic" in the sense that "to break a taboo is to set in motion against oneself mystic wonder-making power in one form or another." In any case, for him a belief in mana, a generalized impersonal power, was historically and psychologically prior to animism, a belief in spiritual beings, asserted by Tylor to be the original form of religion. Preanimism (belief in mana) thus gave way to animism, and animism and polytheism, with more advanced thought, were superseded by monotheism. As against this it was argued by some, particularly Wilhelm Schmidt, that many peoples then called "primitive," the Bushmen of the Kalahari Desert and the pygmies of the Congo, for example, had a conception of a supreme god, often associated with the firmament and, hence, a high god. Whether or not primitive religions incorporated the conception of a supreme divinity now appears more a matter of theological and missionary controversy than a scientific problem. The evidence now available shows that some peoples do emphasize the unity and transcendence of divinity more than others and that the same people will stress divine unity and transcendence in some circumstances and divine multiplicity and immanence in others, as, indeed, in the so-called higher religions.

These early attempts to define the fundamental features of "primitive," as distinct from any other religion, have little relevance to contemporary discussion, except that it is from them and from the vocabulary they adopted that the general reader is bound to take his bearings. What they were concerned with, however, were three very profound and perennial themes in religious experience, exposed, as we now see, not only in the "advanced" religions but also in a rich variety of ethnographic literature from all over the world. The first is that of divine unity and divine multiplicity. The second is what may be called the ambivalence of the quality of sacredness. Not only do sacred things, persons, and places both attract men in veneration and keep them at a distance in avoidance and

respect but also the gods themselves often have, from the human point of view, both benevolent and malign aspects. Hence they are sometimes called to help men and others to leave them untroubled. In much of what has been reported of "primitive" religion, the gods, though they may sanction morality, are themselves ethically neutral, uniting the goodness and badness that are recognized as categories of merely human experience, so that what kills men in one context cures them in another. And, finally, there is the theme of man's ultimate passiveness in relation to a universe animated by powers beyond his technical or rational control.

Primitive cosmogonies and cosmologies; sacred time and times, sacred space and places. In the absence of modern astronomy and physics, explanations of the structure of the universe and the origin of the world are incorporated in "primitive" religious teaching. All peoples have some form of cosmology, though not all have the elaborate cosmogenic myths of the evolution of the universe that are characteristic of the religious traditions of ancient Egypt and Mesopotamia. But that the Australian Aborigines, regarded by earlier anthropologists as the most primitive of living men, should have particularly complex cosmological myths that are periodically dramatized in liturgical performance is sufficient to show that primitive religion is far from impoverished (or "primitive") in this respect. Indeed, the myth of Genesis is an unusually simple account of creation, a fact that may well explain how widely men have found it possible to accept it.

Creation myths

By courtesy of the Australian News and Information Bureau



Australian Aborigines dancing a tribal myth.

That myth, however, does provide what all cosmological myths incorporate as a central teaching—an account of how the palpable world of a particular population in a particular geographical location emerged from an earlier state regarded as more formless, more constricted, and less diversified. So, for example, in Maori cosmology from New Zealand, as also in much African cosmology, sky and earth are represented as having been originally conjoined. Only when sky and earth are separated (often through some act of the first human beings) does the world as it is now known emerge, with a clear distinction between what is above and what is below, between light and dark, and so on.

From this separation follows one of the major polarities of religious thought and symbolism—that of higher and lower powers and orders of being. With the complementary opposition of sky and earth, high and low, often go other symbolic oppositions: that between male, usually associated with the higher, and female, usually associated with the lower; the earth, often represented indeed as a mother; that between superior strength, particularly intellectual and spiritual strength, and human weakness; and that between light and dark, knowledge and ignorance. Similarly, the complementary oppositions of east and west or of what is on the right hand and what is on the left are commonly associated, respectively, with the auspicious and the inauspicious, the strong and the weak, and life and death. Consequently, in many rituals the east and the right are directions symbolizing birth and vitality, and the west and the left are reminders of death and grief.

As with space, so with time. There are not only the seasons of innumerable religious calendars, often running parallel to and sacralizing the seasonal cycle upon which the well-being of those who live at a subsistence level depends. There are also times of crisis at which special precautions of abstinence must be taken—for example, the exclusion of sexual intercourse before hunting, as among the Lele of the Congo, or in the various periods and forms of abstinence kept by the Crow and other Plains Indians of North America in preparation for their guiding visions. In a sense, as primitive religion seems at least implicitly to acknowledge, all human suffering comes from the limitations imposed upon fulfilling, in the material world dominated for human beings by time and space, the freedom of action that can be present to the imagination.

The cause of suffering

Hence, in many rituals a sacred time is represented. In the individual's experience of time, life ends with death. In the collective and religious representation of time, life and strength are recreated regularly in a repetition of the same cycle of ceremonies. And this sacred time of ritual is accompanied by the sacred spaces and places to which men attribute what values they will: the altar, the shrine, the marking out of the particular area in which rituals are to be held, and the recognized abodes of gods and spirits —all represent the imposition of a special form of order, a sacred order, upon the flux of ordinary spatial and temporal experience. Thus, in "primitive" religions of which the historical provenance is known and recent, such as the Ghost Dance religions of North America or the Christian separatist churches of Africa, the deliberate recreation of time and space in religious terms is found, so that, for example, the return of the ancestors or the coming of a messiah are imminently expected. In general it may be said, however, that an elaborate eschatology (doctrine of the final state of the soul and world) is not a common feature of primitive religions, which fix the attention on the sufferings of this life and usually pretend to little knowledge of any other.

Primitive views of man's nature, origin, vocation, and destiny. In many myths, including the biblical ones, man is represented as having been originally nearer to God than he now is. In some myths he is made by God, as a potter makes a pot, for example. Among the Nilotic Dinka and other African peoples, man at one time had easy access to God, and suffering and death were unknown. But having imagined such a state, which reflects the good fortune and well-being that men even now associate with the necessary help of the gods, many primitive religions concentrate on man's condition as it is now, rather than as it was in the past or on what it might be in some ideal future. When men die, they return to or are merged with the ancestors in conditions of "another world," which is rarely elaborately defined. Man's original destiny, myths frequently suggest, was not to die, but what happened to bring death to him in the beginning cannot be reversed. His own vocation is to live his life according to ancestral teachings and to reproduce himself so that a line of descendants will secure for him the only certain immortality. Consequently, though some peoples have quite complex representations of the composition of the human personality—the Akan of Ghana, for example, recognize in it three constituents—speculation about the qualities and fate of anything that could be called the individual human soul is minimal. Only when (as it seems from a modern standpoint) the guilty conscience of the living is disturbed by the dead, particularly when illness is attributed to their dissatisfaction, are attempts made to identify its cause. So with all the rich imagination of primitive religious exegesis, the basic view of man's place in the universe probably held by most people for most of human history has been little different

Views of the soul and afterlife

from that reported by Knud Rasmussen, explorer and ethnologist, from an Eskimo:

> We fear! . . . . And therefore our fathers, taught by their fathers before them, guarded themselves about with all these old rules and customs, which are built upon the experience and knowledge of generations. We do not know how or why, but we obey them that we may be suffered to live in peace.

PRIMITIVE MYTHOLOGY AND SYMBOLISM

The role of myth in primitive cult and culture. The form in which the knowledge of previous generations is handed on is, of course, partly in practical teaching of practical knowledge. But metaphysical truths are apprehended through myth. As H. and H.A. Frankfort wrote,

> Myth is a form of poetry which transcends poetry in that it proclaims a truth; a form of reasoning which transcends reasoning in that it wants to bring about the truth it proclaims; a form of action, of ritual behaviour, which does not find its fulfillment in the act but must proclaim and elaborate a poetic form of truth.

Malinowski and many followers have rightly pointed out that myths provide a "charter" for contemporary belief, action, and social relationships. More particularly they often explain the principles, if not all the details, of contemporary rituals. Here what modern historians would distinguish as myth, legend, saga, fable, or history shade into one another, especially in an oral tradition. But myth, which reveals and perpetuates a body of truths to live by, is a way of interpreting the essential structure of the little moral and physical universe in which many people live and, hence, is recognized as a mode of thinking about man's place in nature and society. The cosmological myths earlier mentioned, which explain how the world as a whole came to be arranged as it is, are accompanied by others establishing and validating the innumerable local forms of social and religious order.

Symbolism in primitive thought and action. With mythical thought, which connects the past and the present to representations of the human condition that transcend them both, go symbolic thought and action, which reorder the given world according to principles both intellectual and affective. Sir James Frazer had already identified two elementary features of symbolic thought and action in what he called "homoeopathic or imitative magic" and "contagious magic," though his interpretation of them as misunderstandings of the relation between cause and effect would no longer be accepted. In the former, objects and experiences that were in one sense similar were related by the mind so that one could stand for (or as it would now be said, symbolize) the other. The Arunta, a group of Australian Aborigines, have as their totem the witchetty grub and perform ceremonies to multiply it. "A long narrow structure of branches," wrote Frazer, "is set up to imitate the chrysalis case of the grub. In this structure a number of men, who have the grub for their totem, sit and sing of the creature in its various stages. Then they shuffle out of it in a squatting posture, and as they do so they sing of the insect emerging from the chrysalis." In "contagious magic" things which have been in intimate contact and thus are united in thought are treated as though one could stand in belief and ritual for the other. Thus, in sacrifices a part of the sacrificial beast is often offered to the gods as substitute for the whole, or a relic of a holy man is venerated (for example, the Buddha's tooth) in commemoration of all the man has stood for. This use of a part to symbolize the whole is a very common feature of symbolic thought and behaviour. More complex are whole systems of symbolic classifications that increasingly command attention. Here the basic features of the spatial ambience and of human social life—what is "higher" and "lower" in nature, for example, and what is socially or spiritually superior and inferior are assimilated to one another in thought. Hence, to take the simplest example, priests, kings, and temples are often associated with high places. Of great interest to students of primitive religion are also the variety and permutations of colour symbolism: white often representing enlightenment, peace, and life (with its association sometimes also with sap and sperm); black representing danger and misfortune; and red representing blood and physical vitality. The general principle behind all this symbolic activity is to bring experiences of the natural and material world together with the "inner" moral and imaginative experience of man and, thus, to express, in a form accessible to the senses, truths accepted on the moral and mental plane.

Symbolism in imitative and contagious magic

Primitive art and iconography. The extent to which spiritual realities are represented in man-made material forms (whether "art" forms or not) varies greatly from one culture to another. Peoples whose sacred objects are few, simple, and produced with little artistic or craftsmanlike effort, such as the Nilotes and the Bushmen of Africa, often have a wealth of verbal art at the service of their religions; therefore, the richness of a religious tradition is not to be judged by the complexity of its iconography. Further, where images of the gods are numerous and deploy the talents of specialized craftsmen, they are often produced less for aesthetic effect than to preserve a material impression of divinity that may be so placed in the home or the temple that it can rarely be fully seen, like the supposedly "magical" pictures of prehistoric cave art. Primitive iconography is thus, from the primitive point of view, less a form of art than an effort to give physical form to realities of the religious imagination. Since the evocation of divinity, whether in words or in materials, is a form of creation, those who are skilled in the activity are often accorded a special religious status under the auspices of the divinities they reveal. The forms in which divinities are depicted vary from schematic and abstract diagrams and pictograms to the fully rounded forms most familiar in West African or Melanesian carving, but both abstract and representative forms are found in the religious art of both these areas; their purpose, decorative as they may be, is less one of decoration than of bringing vividly to the mind and the senses the direct impact of the divine on men.

Non-aesthetic functions of primitive art

PRACTICES, INSTITUTIONS, PERSONAGES

Primitive religious experience and its expressions. In earlier European writings of primitive religion, much was made of a supposed contrast between the fear that inspired primitive religious practice and the hope, love, and confidence preached by the "revealed" religions, Christianity in particular. But even in biblical teaching "the fear of the Lord is the beginning of wisdom," and it is now accepted that the societies formerly called "primitive" are no more obsessed with fear of the unknown than are the advanced societies of today. Nevertheless there can be no doubt that anxieties, which in scientifically and technologically advanced societies are dispelled for some by a sense of scientific and technological competence to deal with their causes, are the therapeutic province of primitive religion. Only when the spiritual cause of suffering is diagnosed and named can proper action be taken to remove it by prescribed ritual action. As a result, in much of the world today, where both "primitive" religion and modern medical science are available, it is common to find that patients turn for relief both to medical science, which deals with their symptoms, and to their own religious practitioners, who deal with the cause of those symptoms. For them, the phenomena of their world are not, as theoretically in some modern science, simply "objects" to be controlled by technical and scientific knowledge: the world is also a living subject, and its life is to be respected, as when, for example, North American Indians objected to the injury done to the earth itself by the metal implements of rationalized agriculture.

Basic religious actions. Thus, a sacramental view of man's relation to his world is implied in primitive belief and ritual, in that the nonhuman creatures and phenomena with which he shares the world (and especially those he regards as vitally affecting his own life) are regarded as signs of a further spiritual reality. They are to be understood and propitiated, even bargained with, rather than dominated as in rationalistic scientific tradition. A central act in man's relationship with the divine is sacrifice, often blood sacrifice, when the life of the sacrificial

Sacrament, sacrifice, and prayer

victim, be it a beast, a vegetable, or even a human being, is offered in exchange for the lives of those who sacrifice and, often, also to augment their own fertility and vitality. Formerly, the essence of primitive sacrifice was thought to lie in the notion of communion with the god through eating the flesh of the sacrificial victim, but it is clear now that this interpretation was strongly influenced by the Christian doctrine of the Eucharist. The commonest sacrificial rituals are probably piacular (expiatory), involving giving satisfaction to the ancestors and the gods so that men may live in peace and health. Thus, sacrifice may be seen as a form of exchange relationship, in which, basically, a superior vitality is received from the gods in exchange for an inferior form of life. In hymns and prayers, the suprahuman powers thought to determine human well-being are praised, exhorted, and sometimes cajoled to come near to men to help them or, as frequently, to accept the offerings men make to them and leave the human sufferers in peace.

Salient rites. Since what is sought from primitive religious practice is "life," ritual develops primarily around critical periods in the life of the individual or the community. These critical periods may be times in the natural cycle when rain is needed for the crops and herds, when game is migrating and may be more or less accessible to the hunter, or when the harvest, with its promise of future plenty or its threat of future deprivation, is gathered in. The local communal cycle of yearly rituals is often complemented by other cycles covering several years, which draw together and affect the welfare of many local communities normally living in comparative isolation from one another. Also there are the "rites of passage," performed when the persons are passing from one social role or status to another. Such major crises are birth and the purification of the mother when the child seems likely to live, initiation when children and youths are recognized as having grown to the status of adults, marriage, and death, sometimes accompanied by exhumation and reburial when the dead person is finally accepted in human memory as having left the community of the living. Similarly, there are rites to mark the passage of an heir to the kingship itself or of a neophyte to the assumption of the full priestly role.

Worship or veneration centred on natural objects or forces. Modern students of primitive religion, not unlike its own practitioners, are concerned with the different forms in which "Nature" is sacralized and brought into contact with human life. Therefore, the innumerable forms in which sacredness is thought to manifest itself, itemized at length in Frazer's *Golden Bough,* are matters for detailed symbological investigation in each particular culture. It helps little, however, to learn that people "worship" animals, plants, and celestial phenomena; what counts is how the religious imagination apprehends the relationship of these phenomena to human affairs. So, for example, totemism, the intimate association of natural beings and forces both with a transcendental world of spirit and with specific social groups, may be seen as the systems of ideas and metaphors in which particular peoples conceive the relation of the natural to the social and mental world. The totemic species, be they species of animals, plants, or trees, are symbols of invisible, intellectual, and affective realities. Hence, in an oft-quoted aphorism, Claude Lévi-Strauss wrote "that natural species are chosen not because they are 'good to eat' but because they are 'good to think'."

Totemism

Worship or veneration centred on human or ex-human beings and qualities. Many writers, following Durkheim, have seen some analogy between the material emblems of totemic spirits and heraldic devices, national flags, and the like that also symbolize communal solidarity of different kinds. All such symbols call to mind a community embracing past, present, and future members, and, where the more strictly totemic representations are found, they are usually closely associated with the ancestors. But, as with the Manus of Melanesia, the ancestral dead may themselves provide the main focus of religious attention. Among the Tallensi and LoDagaa of northern Ghana, the main emphasis seems to be upon the spiritual authority of the dead, a reflection of the moral and legal authority of living parents. On the other hand, Maurice Freedman, a contemporary British anthropologist, writing of the filial piety for which the Chinese are particularly well known, refers to "the relative ineffectiveness of Chinese ancestors, their general air of benevolence, and . . . the lack of strong feelings of hatred or guilt towards them on the part of their descendants."

But in either case the dead command religious veneration and are commemorated by sacrifice. To attend properly to their burial, their graves, and their supposed postmortem wishes is an important religious duty. A single Chinese text quoted by Freedman serves to suggest the spirit and intention of such piety:

> I, Ling Kwang . . . presume to come before the grave of my ancestor . . . Ling Kung. . . . Cherishing sentiments of veneration, I look up and sweep your tomb. Prostrated I pray that you will come and be present, and that you will grant to your posterity that they may be prosperous and illustrious. . . . Always grant your safe protection. My trust is in your divine spirit. . . . Reverently I present the fivefold sacrifice . . . earnestly entreating that you will come and view them.

The special quality of dead kings

In general, the wider the community for which the dead are significant, the more elaborate the ritual and ceremonial surrounding them and the more they are assimilated in idea to a transcendental divinity. Thus, the kings of the Shilluk of the Sudan have the title *jwok piny* ("god on earth") in their lifetime, and, like many other kings in whom divinity itself is held to be manifest they must not be thought to die like ordinary mortals. Their spirits, like the spirits of national or tribal heroes, do in fact continue to inspire the communities which they serve and which serve them, and their immortality suggests to their followers that they themselves, not as individuals but as generation after generation of the same people, are immortal.

Special disciplines and practices to attain religious states. The assumption of important social roles, such as those of king or priest, is usually accompanied by special forms of self-discipline, as are also initiation and mortuary rites. Commonest among these are abstentions from certain foods and from sexual intercourse, fasting, and periods of seclusion or withdrawal from the normal round of social life. The theme of "rebirth" is frequently stressed. Similarly, the officiants at important religious ceremonies may prepare themselves by various forms of purification involving self-mortification in order to achieve a special religious condition. In addition to this, there are the deliberate and positive efforts of individuals to make personal contact with divinities. These are most strikingly exemplified in spirit mediumship and shamanism, when the medium or medicine man sets out, as it were, to relinquish his own personality so that he may become the vehicle for divine activity. Mediums and shamans get their "call" very commonly through illness, during which they receive divine instructions about the particular forms of self-discipline they are to practice. This very prominent feature of primitive religious experience continues to assert itself in syncretic religions, so that leaders of Melanesian Cargo Cults, African separatist churches, and other millenarian movements have revealed to them forms of self-abnegation considered necessary for a new and more satisfying way of life. In Cargo Cults, this has often taken the extreme form of the wholesale destruction of traditional forms of wealth. The objective of such disciplines is to acquire a deeper insight into spiritual reality and to lift the practitioner, if only temporarily, to a state of being that partakes in the divine transcendence. Many American Indian tribes are specially noted for this deliberate visionary quest, undertaken with fasting and other forms of self-inflicted suffering and accompanied by the use of drugs. Such practices, like the persistent dancing and drumming of voodoo rituals, appear to produce a state of dissociation in which the ordinary human personality subordinates itself to the gods.

Magical, divinatory, and spiritualistic practices. The religious practices summarized above are concerned basi-

cally with an effort to perceive the true nature of human vitality and human suffering, so that the former may be augmented and the latter diminished. Consequently, the occasional arbitrariness of the taboos, rules, and procedures by which this supra-rational truth is approached may itself appear as a guarantee of its divine origin. In this respect it is no longer easy to draw a firm line, as was common in the past, between magical and religious practices; for if "magic" involves, as was once maintained, the attempt to influence events by a quasi-scientific knowledge and "religion" involves the attempt to do the same through the supplication of higher powers, both are commonly represented in the very same rituals. This is most clearly seen perhaps in the immense range of divinatory practices throughout the world, one famous example being the complex system of divination called Ifa among the Yoruba of Nigeria, where spiritual guidance is sought through the systematic manipulation of 16 palm nuts. Here the truth, usually about a client's illness or misfortune, is arrived at by precise, quasi-scientific procedures, but it is a truth which is thought to have divine sanction and authority: "Truth is the character of Olodumare [God]. Truth is the word that cannot fall. Ifa is truth. Truth is the word that cannot spoil. . . . 'Those who speak the truth are those whom the gods will help.' " Similarly, in the voice of the spirit speaking through its human medium and manifested commonly by violent trembling and shaking of the body followed by entry into trance, the true nature of the condition about which it is being consulted is thought to be announced. It should be recognized, however, that such primitive religious practices leave room for a good deal of skepticism about individual mediums, whose possible fraudulence is well-known to their clients. Nevertheless, as Raymond Firth has noted, "scepticism about the validity of the performance of some spirit mediums [does] not extend to the performances of all." E.E. Evans-Pritchard made the same point about the Sudanese Azande's consultation of oracles and "magicians" about witches.

The divinatory and spiritualistic paths to truth

Religious personages; types of religious authority. Contact with truths hidden from other men thus confers religious authority, and knowledge of those truths permits the religious leader to restore and maintain health and harmony, whether in the body of the individual patient consulting a diviner or in the body social in which a sacred king reigns. The Mughal emperor Akbar (1542–1605) claimed that ". . . a king possesses, independent of men, the ray of Divine wisdom which banishes from his heart everything that is conflicting" (quoted by Max Müller, *Introduction to the Science of Religion*, 1873), and it is essential in many religious ceremonies that the congregation and officiants should be at peace with each other. This peace may encompass more than a single tribe or nation where, as commonly in reaction to foreign rule, prophets have arisen who claim to speak for God on behalf of all their followers. From the family elders serving their own ancestors, to the minor priests officiating for local gods at local shrines, to the national or tribal cult, and finally to intertribal unions under the auspices of an inspired prophet, religious authority embraces larger and larger communities. As it does so, it claims to define truths that have more and more general application. The teachings of primitive religions, profoundly relevant in their specialized social and natural environments, here begin to give way to intuitions of more universal significance.

BIBLIOGRAPHY. *The Encyclopedia of Religion and Ethics*, ed. by JAMES HASTINGS, 13 vol. (1908–26, reprinted 1955); and JAMES G. FRAZER, *Golden Bough*, 3rd ed. rev., 12 vol. (1907–15; abridged ed., 1922 and 1959), though out of date, remain comprehensive works of reference. *Man, Myth and Magic* (published in parts from 1970), is a popular encyclopaedia. E.E. EVANS-PRITCHARD, *Theories of Primitive Religion* (1965), critically surveys most earlier contributions. E.B. TYLOR, *Primitive Culture*, 2 vol. (1871; paperback ed., 1958), is a historical landmark. PAUL RADIN, *Primitive Religion: Its Nature and Origin* (1938); and R.H. LOWIE, *Primitive Religion* (1924, reprinted 1970), represent later, primarily psychological, approaches. EMILE DURKHEIM, *Les Formes élémentaires de la vie religieuse* (1912; Eng. trans., *The Elementary Forms of the Religious Life*, 1915, reprinted 1965); and HENRI HUBERT and MARCEL MAUSS, *Essai sur la nature et le fonction du sacrifice* (1898; Eng. trans., *Sacrifice: Its Nature and Function*, 1964), are influential sociological analyses, the latter, however, drawing mainly on Semitic and Vedic studies. LUCIEN LEVY-BRUHL in *Les Fonctions mentales dans les sociétés inférieures* (1910; Eng. trans., *How Natives Think*, 1926), and other works, puts forward his own anti-intellectualist view of "primitive mentality." Earlier theoretical works based upon field research and representing different aspects of the "functionalist" interpretation of the subject are A.R. RADCLIFFE-BROWN, *The Andaman Islanders* (1922, reprinted 1964); and the writings of BRONISLAW MALINOWSKI, whose basic ideas appear in *Magic, Science and Religion, and Other Essays*, ed. by ROBERT REDFIELD (1948). The best modern writings appear in monographs on specific peoples, though CLAUDE LEVI-STRAUSS, *La Pensée sauvage* (1962; Eng. trans., *The Savage Mind*, 1966), and other works by him are influential structuralist re-analyses of an encyclopaedic range of information. JOHN MIDDLETON (ed.), *Myth and Cosmos* (1967), *Magic, Witchcraft and Curing* (1967), and *Gods and Rituals* (1967), are useful selected readings. From a wide range of modern monographs, there may be mentioned the following, grouped according to area. (*Africa*): E.E. EVANS-PRITCHARD, *Nuer Religion* (1956); GODFREY LIENHARDT, *Divinity and Experience: The Religion of the Dinka* (1961); JOHN MIDDLETON, *Lugbara Religion: Ritual and Authority Among an East African People* (1960); V.W. TURNER, *The Forest of Symbols* (1967). (*Asia and Southeast Asia*): HANS SCHAERER, *Die Gottesidee der Ngadju Dajak in Süd-Borneo* (1946; Eng. trans., *Ngaju Religion: The Conception of God Among a South Borneo People*, 1963); MELFORD E. SPIRO, *Burmese Supernaturalism: A Study in the Explanation and Reduction of Suffering* (1967); S.J. TAMBIAH, *Buddhism and the Spirit Cults in North-East Thailand* (1970); ARTHUR WALEY, *The Way and Its Power: A Study of the Tao tê ching and Its Place in Chinese Thought* (1934, reprinted 1958); C.K. YANG, *Religion in Chinese Society* (1961). (*Americas*): L.C. FARON, *Hawks of the Sun: Mapuche Morality and Its Ritual Attributes* (1964); NIELS FOCK, *Waiwai: Religion and Society of an Amazonian Tribe* (1963); E.Z. VOGT, *Zinacantán: A Maya Community in the Highlands of Chiapas* (1969); E.C. PARSONS, *Pueblo Indian Religion*, 2 vol. (1939). (*India*): WILHELM KOPPERS, *Der Urmensch und sein Weltbild* (1949; Eng. trans., *Primitive Man and His World Picture*, 1952); E.B. HARPER (ed.), *Religion in South Asia* (1964); S.C. ROY, *Oraon Religion and Customs* (1928); M.N. SRINIVAS, *Religion and Society Among the Coorgs of South India* (1952); CHRISTOPH VON FUERER-HAIMENDORF, *Morals and Merit: A Study of Values and Social Controls in South Asian Societies* (1967); VERRIER ELWIN, *The Religion of an Indian Tribe* (1955). (*Pacific*): ELSDON BEST, *Maori Religion and Mythology* (1924); RAYMOND FIRTH, *The Work of the Gods in Tikopia*, 2nd ed. (1967); KENELM BURRIDGE, *Mambu: A Melanesian Millenium* (1960); R.F. FORTUNE, *Manus Religion: An Ethnological Study of the Manus Natives of the Admiralty Islands* (1965); PETER LAWRENCE and M.J. MEGGITT (eds.), *Gods, Ghosts, and Men in Melanesia: Some Religions of Australian New Guinea and the New Hebrides* (1965); F.E. WILLIAMS, *Drama of Orokolo: The Social and Ceremonial Life of the Elema* (1940). The *International Bibliography of Social and Cultural Anthropology*, published regularly since 1955 by UNESCO, gives a wide perspective on current writings.

(R.G.Li.)

Primulales

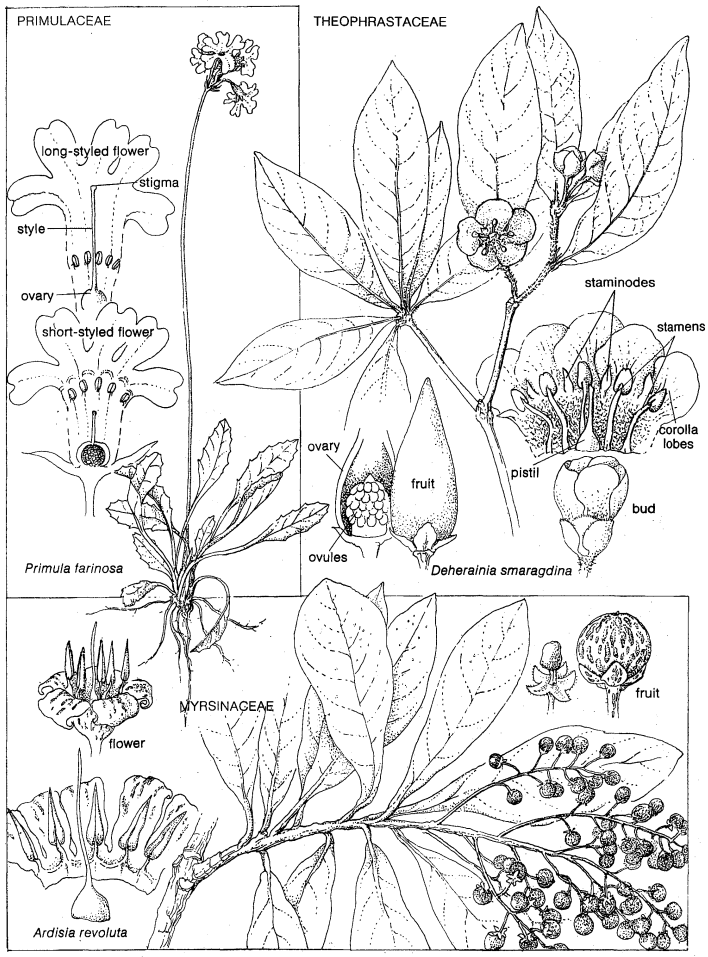
The primrose order of flowering plants (Primulales) contains three families, about 64 genera, and nearly 2,000 species. The group has long been of popular interest because of the attractiveness of many of the wild plants, such as shooting star (*Dodecatheon*) and loosestrife (*Lysimachia*), and the beauty of the cultivated members, such as cyclamen (*Cyclamen*) and primrose (*Primula*).

Two of the families (Theophrastaceae and Myrsinaceae), which contain about half of the species in the order, are composed entirely of trees or shrubs. A few species of these families are climbers or epiphytes (plants that grow upon other plants instead of being rooted in soil). The order is worldwide in distribution, although the two woody families have a much more restricted range than the third, the primrose family (Primulaceae), which is composed of herbaceous (nonwoody) plants.

The economic importance of the order lies almost entirely in its ornamental value. Very few of the woody

representatives are in cultivation and then only infrequently. In the tropics some of the woody species are used to poison fish and as a tapeworm remedy (*e.g.*, the pea-sized fruits of *Myrsine africana*).

Drawing by M. Pahl



Representative plants from the three families of the order Primulales.

Cross-pollination ensured by flower structure

Natural history. In the wild the plants reproduce mostly by seed, and in cultivation this method is also common. Cross-pollination in the woody families is attained by the occurrence of species with separate male and female flowers. In some species of the family Primulaceae, however, cross-pollination is effected by heterostyly, a condition in which some flowers have differently sized or positioned sexual structures (male stamens and female pistils) than others of the same species. Heterostylous flowers appear superficially identical, but some have short styles (the narrow, upper part of the pistil) and elevated anthers (the saclike pollen-producing part of the stamen); others have long styles and anthers near the inner base of the tubular corolla (fused petals). Self-pollination is almost entirely prevented by the different lengths of styles and stamens. Cross-pollination is practically ensured when an insect visits the flowers because it transfers pollen only between male and female structures of similar length, which are in different flowers. An insect (usually a bee) visiting a short-styled flower, for example, could encounter long stamens and pick up a load of pollen from them. This pollen would then be deposited on the stigma (the sticky, pollen-receiving region at the tip of the style) of a long-styled flower on a subsequent flower visit, but not on a short-styled flower because the pollen is not carried in an appropriate place on the insect to reach the stigma of a short-styled flower. A similar restriction operates when an insect picks up pollen from the short stamens of a long-styled flower. The pollen can only be deposited on the stigma of a short-styled plant. The centre, or "eye," of the corolla is often a different and brighter colour than the rest of the flower and is thought to function as a guide for the insect.

The family Primulaceae occurs primarily in cool climates and damp, usually acidic soils. A large number occur in mountainous regions, and many grow in the arctic tundra or in alpine communities at high elevations. The two woody families are subtropical and tropical.

Form and function. Few morphological characters other than the flowers and fruits make the order readily distinguishable. The leaves are simple, lack the small basal appendages called stipules, and are generally alternately arranged along the stems; but this combination of traits is by no means unique. The location of the stamens in front of the corolla lobes, rather than in the usual position of being alternate with them, and the unusual free-central placentation—the attachment of the ovules to a short column that arises from the interior base of the ovary—make the order completely distinctive. The pistil is composed of five carpels, or ovule-bearing segments, which is not unique but is sufficiently unusual to make it useful in recognizing members of the order.

The two woody familes differ morphologically from each other in two characters. The Theophrastaceae has, in addition to the five stamens, five petallike staminodes, or sterile stamens, that lie alternately with the corolla lobes. The seeds of this family are large and yellowish. The family Myrsinaceae lacks staminodes—at least petallike ones—and the seeds are small and dark coloured. In both families the fruits are more or less fleshy. The family Primulaceae, being herbaceous (rarely a little woody at the base), is readily differentiated from the other two families. The fruit, a capsule, is not fleshy.

Evolution and classification. The two woody families, which are included in the order chiefly because of the structure of the flower and fruit, may have had a different evolutionary origin from that of the herbaceous family. The form and structure of the flowers in the three families, however, are so striking and similar that their placement in one order is widely accepted.

There is strong evidence that the more or less immediate ancestors of the order had ten stamens and a five-celled pistil. Loss of the five outer stamens and of the septa forming the five cells readily explain the present type of flower. The carpel number of five contrasts with the usual two-carpelled condition of other plants with petals fused into a tubular corolla (sympetalous plants). On the widely held supposition that the larger the number of parts a flower has the less evolved it is, this order is considered primitive within the sympetalous group. Sympetalous plants are, however, considered to be evolutionarily advanced over most plants with free-petalled flowers.

Similarities to other plant orders

The primrose order seems to be related to the ebony order (Ebenales), but neither is considered to have been derived from the other, and the two orders may have had a common origin. The origin has been sought among flowering plant groups with separate, free petals. The Caryophyllaceae (pink family), consisting primarily of herbaceous plants, has several characters remarkably similar to the Primulaceae, including the unusual free-central placentation. For this reason there is widespread opinion that this group is ancestral to the Primulaceae. Because woody plants are considered more primitive than herbs, however, and therefore the woody families of the order would not likely be derived from herbaceous families, it has been considered logical to look for a woody ancestor. Some authorities have recently suggested relatives of the woody, free-petalled tea or camellia order (Theales) as likely ancestors. A classification has been proposed that derives the families Theophrastaceae and Myrsinaceae from the order Theales, and removes them from the Primulales. This leaves the family Primulaceae in the primrose order; it is considered to be derived from the Caryophyllaceae.

Evidence from chemical studies of plants is not yet definitive. Leuco-anthocyanin pigments occur in the order Theales and the family Myrsinaceae but not in the Theophrastaceae. Although infrequent in sympetalous families, they do occur in the Primulaceae and in the possibly

related family Plumbaginaceae (leadwort family). But in the putative ancestor, the Caryophyllaceae, they are lacking.

Fossil leaves and flowers have been found representing the family Myrsinaceae in the Northern Hemisphere from Tertiary (2,500,000 to 65,000,000 years ago) and upper Oligocene (about 26,000,000 years ago) deposits. These are from regions in which living species of the family no longer occur.

Annotated classification. The following descriptions are generalizations; some are true in all instances and some have exceptions, usually rare ones. The superior-fruit (*i.e.*, an ovary positioned above the attachment point of the other flower parts) character finds an exception in the genus *Samolus* (family Primulaceae) in which the capsule is half-inferior, or partly enclosed by the basal parts of the sepals, petals, and stamens. The genus *Coris*, of the same family, with a single species in the Mediterranean region and northeast Africa, has a bilaterally symmetrical corolla. It has been suggested that this aberrant species does not belong in the order Primulales.

Some exceptions to general characteristics of order

ORDER PRIMULALES

Trees, shrubs, or herbs with leaves usually alternate, simple, and lacking stipules. Flowers usually radially symmetrical; with the basic parts (sepals, petals, stamens, and carpels) 5 in number (rarely 4 to 8). The 5 petals united into a radially symmetrical corolla, the 5 stamens arising from the inner surface of the corolla tube and opposite the corolla lobes. Ovary superior, of 5 fused carpels, but only 1 interior chamber, with few to numerous ovules attached to a centre axis rising from the base of the ovary. Three families, about 64 genera, and more than 2,000 species.

Family Theophrastaceae

Trees or shrubs with leaves characteristically crowded at ends of branches. Flowers often large and showy; with a whorl of 5 petallike staminodes (sterile stamens) alternating in position with the corolla lobes in addition to the 5 fertile stamens that are opposite the corolla lobes. Fruit a fleshy drupe (stony-pitted fruit) or berry with few to many seeds. Seeds large and yellowish or orange red in colour. Four genera and 60 to 110 species, depending on the authority consulted, distributed in the American tropics and Hawaii.

Family Myrsinaceae

Evergreen trees or shrubs, often with leathery leaves. Leaves glandular-punctate (with glands, translucent or coloured dots, depressions, or pits dotting the surface) and with resin ducts, which are also present in the stems. Calyx (sepals) and corolla of 4 to 6 parts or lobes; staminodes absent or small. Fruit a fleshy berry or drupe; ovules numerous; seeds small and dark-coloured. Thirty-two genera and 1,000 species widely distributed in the tropics and subtropics of the world; they are especially abundant in Africa, Malaysia, and Australia.

Family Primulaceae (primrose family)

Herbaceous plants (rarely woody at the base) with leaves often in a basal rosette or, when on an aerial stem, frequently opposite or whorled. Flowers often on leafless stalks (scapes). Calyx persistent (remaining attached into fruiting season). Fruit a several-seeded capsule, opening by 5 apical valves or teeth. Seeds angular. Twenty-eight genera and 800 to 1,000 species (half of which are in the genus *Primula*) distributed mostly in mountainous, northern temperate regions.

Critical appraisal. There are serious differences of opinion concerning what families ought to constitute the order Primulales. The most widely accepted viewpoint is that presented here, which recognizes the three families described in the annotated classification section above.

Another taxonomic system, gaining in acceptance, recognizes the families Myrsinaceae, Theophrastaceae, and Aegicerataceae (consisting of the single genus *Aegiceras*, here placed in the Myrsinaceae) to constitute a separate order, the Myrsinales. In that system the primrose order contains the two families Primulaceae and Plumbaginaceae. The main division between the two orders is based on woodiness versus the herbaceous habit, the Myrsinales containing the woody plants, and the Primulales containing the herbaceous ones.

Still other systems accept all the above mentioned families (and sometimes also the family Plantaginaceae) into a more broadly defined primrose order.

The order is thus distinctive morphologically; its evolutionary origins and whether or not its families came from one or more than one ancestral stock still are not clear. Continued anatomical and phytochemical studies may provide the answers.

BIBLIOGRAPHY. WALTER C. BLASDALE, *Cultivated Species of Primula* (1948), a richly illustrated book that gives the history of the genus, descriptions and ranges of the species, and information on horticultural methods; ROY GENDERS, *Primroses* (1959), a short, popular, illustrated account on the propagation of these plants, together with instructions for hybridizing and displaying; DORETTA KLABER, *Primroses and Spring* (1966), a brief, illustrated, popularly written book with information on how to grow primroses; R.B. CHANNELL and C.E. WOOD, JR., "The Genera of the Primulales of the Southeastern United States," *J. Arnold Arbor.*, 40:268–288 (1959), a semitechnical treatment, with keys, descriptions, and illustrations of the 10 genera of the primrose order found in the southwestern United States; CYRUS R. LUNDELL, *The Genus Parathesis of the Myrsinaceae* (1966), a fully illustrated technical taxonomic treatment of this tropical American genus of over 70 species; SIR W. WRIGHT SMITH and H.R. FLETCHER, "The Genus *Primula:* Section *Nivales*," *Trans. R. Soc. Edinb.*, 60:563–627 (1942); "Section *Farinosae*," 61:1–69 (1943); "Section *Petiolares*," 61:271–314 (1944); "Sections *Obconica, Sinenses, Reinii, Pinnatae, Malacoides, Bullatae, Carolinella, Grandis,* and *Denticulata*," 61:415–478 (1946); "Sections *Cuneifolia, Floribundae, Parryi,* and *Auricula*," 61:631–686 (1948), a definitive, highly technical taxonomic treatment of the species of primrose; E.H. WALKER, "A Revision of the Eastern Asiatic Myrsinaceae," *Philipp. J. Sci.*, 73:1–258 (1940), a technical taxonomic treatment of the six genera and more than 100 species of the family occurring in the area; JAMES DAVIS RAY, *The Genus Lysimachia in the New World* (1956), a technical taxonomic treatment with distribution maps and illustrations.

(G.J.G.)

Prince Edward Island

A thin crescent of land nestling on the southern shores of the Gulf of St. Lawrence, Prince Edward Island is Canada's smallest province, its serenity attracting hundreds of thousands of summer tourists. A mild climate and fertile soil favour agriculture, while its location makes the island a base for lobster and groundfish operations. Despite these natural assets, the provincial economy has lagged appreciably behind national averages for productivity, employment, and per capita income, and the emigration of its young people has been a perennial concern. As a result, a 15-year development plan—the only one of its kind in the political subdivisions of North America—was instituted in 1969. The province's small size and limited resources have precluded its having a major role in the political and economic life of the Canadian nation, though many island natives have become distinguished citizens of other parts of Canada and the United States.

From North Cape to East Point, the curving island is about 140 miles (225 kilometres) long, and varies from four to 40 miles in breadth. It lies between 46° and 47° north latitude, and 62° and 64° west longitude. On the south, the Northumberland Strait separates the island by a minimum of nine miles from the mainland of the provinces of Nova Scotia and New Brunswick. Narrowings of the island roughly mark off the three counties—Prince in the west, Queens in the centre, and Kings in the east. The total land area of 2,184 square miles (5,657 square kilometres) was inhabited by 110,200 people at the 1971 census, making the island the most densely populated of Canadian provinces (for related information, see CANADA; CANADA, HISTORY OF).

History. Before the European explorations, Micmac Indians from the mainland used the island for fishing, hunting, and some planting in the warmer seasons. They called the place Abegweit, which meant "Cradled on the Waves." By the 1970s, only a small island in Malpeque Bay, on the north shore, was reserved for their descendants.

Legend suggests that John Cabot, the English-sponsored Venetian explorer, may have seen the island in 1497, although historians credit its discovery to Jacques Cartier, the French navigator, in June 1534. Claimed for France in 1603 by Samuel de Champlain, the first gover-

nor of French Canada, and designated on his maps as Île St. Jean, it was not colonized until 1720 when 300 settlers from France, sponsored by the commercial company of the Count de St. Pierre, established Port la Joie at the entrance to the harbour of Charlottetown, the present capital. In addition, fishermen and trappers from the French-speaking mainland colony of Acadia established several other small communities on the island. The French regime lasted only 38 years, and then the British occupied the island in 1758, dispersing over 3,500 of the settlers.

British administration

The island was formally ceded to Great Britain in 1763, and six years later the colony was separated from Cape Breton and its first governor appointed. In 1799 St. John's Island was renamed in honour of Edward, duke of Kent, who was then commander of the British forces in North America.

Under British administration the island was surveyed and divided into three counties, each with a townsite and "royalty," and 67 lots or townships of about 20,000 acres each. In 1767 these lots were awarded by ballot to grantees or proprietors who were expected to promote settlement, and for the next century absentee-landlord problems beset the colony. Representative government reached the island in 1851, and the Land Purchase Act of 1875 brought to an end this unique system of land tenure. By 1900 the population reached 100,000. It has fluctuated within 15 percent of that figure for the past 70 years.

In 1864 a conference intended to discuss the possible union of the three Maritime Provinces fortuitously prepared the way for the confederation of all the provinces in the Canadian nation. This Charlottetown Conference was the historic forerunner of the Quebec Conference of 1864, which actually resulted in the founding of the Dominion of Canada. Prince Edward Island has thus been known as the "Cradle of Confederation," even though it did not finally join the union, as Canada's seventh province, until 1873, when forced to do so by severe financial troubles. Resourceful politicians then persuaded the federal government to make concessions that enabled the province to purchase the lands still held by foreign proprietors and resell them to resident farmers; to transfer to the federal authority financial responsibility for the heavily-indebted island railway; and to obtain assurance of continuous communication with the mainland.

By 1970, after a century of chronic economic depression in the area, relieved occasionally by flourishing markets for wooden sailing vessels, silver fox furs, potatoes, or lobsters, Prince Edward Island had joined with Nova Scotia and New Brunswick in a renewed study of Maritime union. At the same time, the island had embarked on its own comprehensive development plan.

The landscape. Numerous streams, bays, and tidal estuaries indent the irregular coastline. On the north side, the bays are generally blocked by dunes; but on the east and south there are good natural harbours. Along the 1,100 miles of pink, sandy beaches, red sandstone cliffs average about 20 feet in height but sporadically exceed 50 feet. Most of the island is within five miles of the sea or tidal inlet and no place is more than ten miles from salt water. Water pollution is not extensive, but effluents from processing plants, pesticides, erosion, and sedimentation have seriously degraded some water resources.

With an average mean temperature of 42° F (5.6° C), and an annual mean precipitation of 39.5 inches (1,003.3 millimetres), the island climate is conducive to forests. The original stands were mainly deciduous, but coniferous trees are as common as they are elsewhere in the region. Besides the native birch, maple, pine, and spruce, foreign species include linden, horse chestnut, black walnut, and European mountain ash. Tall trees line the streets of the towns, as well as the country lanes.

Though a gentle undulation is most characteristic, the landscape ranges from rolling hills in central Queens County to level stretches in western Prince County. The highest elevation is nearly 500 feet above sea level. The soils are classed as podzols, comparatively low in plant nutrients and high in acidity. In over half the island, the red, sandy loam is suitable for clover and timothy hay, oats and barley, and turnips and potatoes. Dairying, mixed farming, and crops for canning and freezing are found in the more central parts of the island, and tobacco is grown in southern Kings County. The average family farm is about 137.4 acres, and the growth period is from mid-May through mid-October.

The people. Most residents are descendants of early settlers from the British Isles: Highland Scots, Englishmen, and southern Irish. Other ancestral strains include several hundred Loyalists who settled here after the American Revolution, and about thirty Acadian families, recorded in the census of 1765, who are the progenitors of over 16,000 island residents today. Many original settlement patterns can still be traced in the place names and family names of the island. People of Scottish ancestry, constituting one-third of the population, are found everywhere, but particularly along the north shore from East Point to Malpeque Bay, throughout Kings County and eastern Queens County; many are descendants of the Glenaladale settlers who arrived at Scotchfort and Tracadie in 1771–72, and the Selkirk settlers who came to the Orwell District in 1803. Somewhat less than one-third of the people claim Irish ancestry and are found mainly in areas of later settlement, notably western Prince County and the interiors of Kings and Queens. English and Loyalist descendants are found mostly in villages and towns in Queens and in eastern Prince County. The Acadians are found in large numbers south of Rustico, near Cape Egmont, and west and north of Cascumpeque Bay. Since 1945, about 120 Dutch families, as well as a sprinkling of Americans, Belgians, Lebanese, and Germans, have settled in this province.

Population

In 1971 the population of Kings County was 18,100; Prince County (including Summerside's 9,300), 41,700; and Queens County (including Charlottetown's 18,600), 50,400. Apart from Charlottetown, the one city, the province has six towns and seventeen incorporated villages. In 1966 the urban population stood at 39,747, and the rural population at 68,788, including 30,841 farm residents. In the 1961 census, 95,564 called English their mother tongue; 7,958, French; and 1,107, another language. At that time, Roman Catholics numbered 48,256; United Church members, 27,395; Presbyterians, 12,744; Anglicans, 6,085; and Baptists, 5,942.

Economy and social conditions. In 1966, the labour force was estimated at 35,400, out of a total population of just over 109,000. Unemployment rates were three to seven percent above the national average and output per worker was below the average—close to 50 percent in agriculture. Per capita income amounted to between 60 and 70 percent of national averages.

During the late 1960s, the government proposed a series of reforms, aimed at enabling the people to create viable economic enterprises for themselves. These included full exploitation of agriculture; development of tourist facilities; better use of forest assets; improvement of fisheries; expansion of manufacturing; public investment in housing, health, and welfare services; and extension of programs of education and training. Federal authorities subsequently designated the province a Special Rural Development Area under the Fund for Rural Economic Development Act, and the two governments entered into a 15-year agreement starting April 1, 1969. The agreement covers the entire province in a comprehensive and coordinated plan intended to promote economic and social development, to increase income and employment opportunities, and to raise standards of living.

Development plan

This plan will cost approximately $745,000,000, of which the federal contribution will be approximately $225,000,000. A Federal-Provincial Advisory Board approves all projects, recommends all programs and budgets, evaluates progress, and suggests amendments. The plan is administered by the province and the federal office of Plan Co-ordinator, but the responsibility for the progress of the plan belongs to the provincial Department of Development headed by the premier.

Agriculture will continue to be the base of the island's economy. Through the Land Development Corporation, low-income farms may be purchased and resold or

leased, one objective being consolidation of the area into a resource base of about 2,500 commercial farm units. Such transactions may involve upward of 400,000 acres. This strategy, along with improved farming and marketing methods, is expected to double the 1966 net agricultural value added to $48,000,000 by 1976 and to $68,000,000 by 1983.

The tourist industry in 1968 had a net value of $14,618,491, but the growth rate has been 10 percent annually since that date, and the objective of the plan is a net value of $18,000,000 by 1979. In 1967 there were 375,000 visitors, and gross revenue was slightly over $13,000,000; by 1970 the visitor count had risen to 557,000, and the gross revenue was estimated at $18,000,000.

The fishing potential is to be developed by consolidating processing plants, reducing both the number of inshore fishermen and the number of ports, and by maintaining the normal growth rate in offshore fisheries. The object is to increase efficiency, quality, and productivity, and thereby effect some increase in total returns and substantial increase in individual incomes. Island fisheries were valued at $9,300,000 in 1970, and the Irish moss industry added another $1,300,000.

Economic growth potential

The net provincial product in 1966 stood at a total of $133,000,000 and was expected to increase to $205,000,000 by 1978 and to $270,000,000 by 1983; under the development plan, the net provincial product is projected to reach $330,000,000 by 1978 and $440,000,000 by 1983. Per capita income, averaged from 1963 to 1965, was $1,236 and could normally be expected to rise to $2,140 by 1983; but in the same period, the plan aimed at boosting per capita income to $3,070.

At the start of the 1970s, the Prince Edward Island Lending Authority was established; the Health Services Commission was appointed to administer the provincial plan for socialized medicine; agencies were set up for housing and pollution control; welfare services were being reorganized; and the Department of Education had embarked on reducing 308 administrative units to five jurisdictional districts by 1972. In addition, the 17 regional high schools were being changed to several comprehensive high schools that would offer academic, general, commercial, and industrial arts courses.

Transportation. As over 90 percent of the travel between the province and the mainland is by automobile, proposals for a causeway over the Northumberland Strait are frequently reviewed. Meanwhile, the Northumberland ferries sail, except in the winter months, between the province and Nova Scotia, while Canadian National Railways icebreakers provide year-round service for both automotive vehicles and freight trains between the island and New Brunswick. The Canadian National Railways, extending the entire length of the island, remains the leading carrier of bulk commodities, though passenger service has been discontinued. By 1970, over 1,400 miles of paved and 1,900 miles of unpaved roads laced the island. Potatoes and pulpwood were shipped out from ports on the island's south and east coasts. Eastern Provincial Airways provides daily jet plane services between Charlottetown, the other Atlantic Provinces, and Montreal.

Administration. The province is represented in the Federal Parliament by one member from each of four electoral districts as constituted in 1968: Cardigan, Hillsborough, Malpeque, and Egmont. Four Senate appointments also are allotted to the province.

The provincial government consists of the lieutenant governor, and an Executive Council chosen from the 32 members of the Legislative Assembly. Two assembly members are elected from each of the sixteen districts, as constituted in 1966. The premier heads the Executive Council of ten ministers, whose responsibilities range from education, agriculture, and fisheries to tourism, industry, natural resources, and development.

Cultural life and institutions. Both amateur and professional productions are fostered at the Fathers of Confederation Memorial, a complex containing a large auditorium, art gallery and museum, children's theatre, and various studios. This centre, which adjoins historic Province Building in Charlottetown, also houses a public library and legislative library and archives.

Holland College, an institute of applied arts and technology, and the University of Prince Edward Island were both chartered in 1968, replacing two institutions that originated over a century earlier, Prince of Wales College and St. Dunstan's University. At the start of the 1970s, Holland College enrolled about 250 students yearly, while the new university enrolled almost 4,000, including part-time and summer-session students. The campuses are located in Charlottetown, but extension courses are offered in several other island communities.

Throughout the province, such handicrafts as weaving, pottery-making, and woodworking are encouraged by government agency and the Women's Institutes, and the annual music festival attracts thousands of participants from the schools of the province. There are radio stations and daily newspapers in Charlottetown and Summerside and a television studio in Charlottetown.

As Prince Edward Island continues its efforts to achieve prosperity while maintaining independence, the programmed economy recently adopted is expected to transform many of the accustomed ways of doing many things. The people, on the other hand, seem determined to preserve their rural and insular heritage, their traditional values, and their humane mode of life.

BIBLIOGRAPHY. J.H. BLANCHARD, *The Acadians of Prince Edward Island, 1720–1964* (1964), a particularly rich source of familial and parochial history; F.W.P. BOLGER, *Prince Edward Island and Confederation, 1863–1873* (1964), the definitive work on this subject; A.H. CLARK, *Three Centuries and the Island* (1959), a fascinating account of the historical geography of settlement and agriculture; *Development Plan for Prince Edward Island: A 15-Year Federal Provincial Program for Social and Economic Advancement* (1969), the official document of this comprehensive development plan; F. MACKINNON, *The Government of Prince Edward Island* (1951), the definitive work on this subject.

(B.O'G.)

Printing

Printing traditionally has been defined as a technique for applying under pressure a certain quantity of colouring agent onto a specified surface to form a body of text or an illustration.

Certain recently developed processes for reproducing texts and illustrations, however, are no longer dependent on the mechanical concept of pressure or even on the material concept of colouring agent. Because these processes represent an important development that may ultimately replace the other processes, printing should probably now be defined as any of several techniques for reproducing texts and illustrations, in black and in colour, on a durable surface and in a desired number of identical copies. There is no reason why this broad definition should not be retained, for the whole history of printing is a progression away from those things that originally characterized it: lead, ink, and the press.

It is also true that, after five centuries during which printing has maintained a quasi-monopoly of the transmission or storage of information, this role is being seriously challenged by new audiovisual and information media. Printing, by the very magnitude of its contribution to the multiplication of knowledge, has helped engender radio, television, film, microfilm, tape recording, and other rival techniques. Nevertheless, its own field remains immense. Printing is used not merely for books and newspapers but also for textiles, plates, wallpaper, packaging, and billboards. It has even been used to manufacture miniature electronic circuits.

The invention of printing at the dawn of the age of the great discoveries was in part a response and in part a stimulus to the movement that, by transforming the economic, social, and ideological relations of civilization, would usher in the modern world. The economic world was marked by the high level of production and exchange attained by the Italian republics, as well as by the commercial upsurge of the Hanseatic League and the Flemish cities; social relations were marked by the decline of the

landed aristocracy and the rise of the urban mercantile bourgeoisie; and the world of ideas reflected the aspirations of this bourgeoisie for a political role that would allow it to fulfill its economic ambitions. Ideas were affected by the religious crisis that was due primarily to the links between the Catholic Church and the older feudal society, links that would lead to the Protestant Reformation.

The first major role of the printed book was to spread literacy and then general knowledge among the new economic powers of society. In the beginning it was scorned by the princes. It is significant that the contents of the first books were often devoted to literary and scientific works as well as to religious texts, though printing was used to ensure the broad dissemination of religious material, first Catholic and, shortly, Protestant.

The significance of the Western alphabet to the use of movable type

There is a material explanation for the fact that printing developed in Europe in the 15th century rather than in the Far East, even though the principle on which it is based had been known in the Orient long before. European writing was based on an alphabet composed of a limited number of abstract symbols. This simplifies the problems involved in developing techniques for the use of movable type manufactured in series. Chinese handwriting, with its vast number of ideograms requiring some 80,000 symbols, lends itself only poorly to the requirements of a typography. Partly for this reason, the unquestionably advanced Oriental civilization, of which the richness of their writing was evidence, underwent a slowing down of its evolution in comparison with the formerly more backward Western civilizations.

Printing participated in and gave impetus to the growth and accumulation of knowledge. In each succeeding era there were more and more men who were able to assimilate the knowledge handed to them and to augment it with their own contribution. From Diderot's encyclopaedia to the present profusion of publications printed throughout the world, there has been a constant acceleration of the rate of change, a process highlighted by the Industrial Revolution at the beginning of the 19th century and the scientific and technical revolution of the 20th.

At the same time, printing has facilitated the spread of ideas that have helped to shape alterations in social relations made possible by industrial development and economic transformations: the works of the 18th-century philosophers, the United States Constitution, the Declaration of the Rights of Man and of the Citizen of the French Revolution, and 19th-century social doctrines.

By means of books, pamphlets, and the press, information of all kinds has reached all levels of society in most countries.

The advantages of print

In view of the contemporary competition over some of its traditional functions, it has been suggested by some observers that printing is condemned to disappear. On the other hand, this point of view has been condemned as unrealistic by those who argue that information in printed form offers particular advantages different from those of other audio or visual media. Radio scripts and television pictures report facts immediately but only fleetingly, while printed texts and documents, though they require a longer time to be produced, are permanently available and so permit reflection. Though films, microfilms, punch cards, punch tapes, tape recordings, holograms, and other devices preserve a large volume of information in small space, the information on them is available to human senses only through apparatus such as enlargers, readers, and amplifiers. Print, on the other hand, is directly accessible, a fact that may explain why the most common accessory to electronic calculators is a mechanism to print out the results of their operations in plain language. Far from being fated to disappear, printing seems more likely to experience an evolution marked by its increasingly close association with these various other means by which information is at man's disposal.

This article is divided into the following sections:

I. History
  Origins in China
    Invention of movable type (11th century)
    Transmission of paper to Europe (12th century)
  The invention of printing
    Xylography
    Metallographic printing (1430?)
    The invention of typography—Gutenberg (1450?)
    The Gutenberg press
  Improvements after Gutenberg
    The metal press (1795)
    Stereotypy and stereography (end of the 18th century)
    Koenig's mechanical press (early 19th century)
    Attempts to mechanize composition (mid-19th century)
    Typecasting compositions (1880)
  19th-century innovations
    Reproduction of illustrations
    Lithography: Senefelder (1796)
    Photosensitivity: Niepce (1820s)
    Gravure and rotogravure (1890s)
  The 20th century
    Discovery of offset (early 20th century)
    Dry offset (1920)
    Colour printing
    Automation of composition (after 1929)
    Programmed composition (1950s)
    Photocomposition
    Toward direct impression
    Serigraphy and collotype: a renaissance
    Three-dimensional printing (1960s)
    Office printing
II. Modern printing techniques
  Composition and typesetting
    Mechanical composition and typesetting
    Phototypesetting
    Makeup of letterpress copy
  Printing (press operation)
    Colour printing
    Letterpress printing
    Rotogravure
    Offset
    Other printing processes
    Printing inks
  The future of printing

I. History

ORIGINS IN CHINA

By the end of the 2nd century, the Chinese apparently had discovered, empirically, a means of printing texts; certainly they then had at their disposal the three elements necessary for printing: (1) paper, the techniques for the manufacture of which they had known for several decades; (2) ink, whose basic formula they had known for 25 centuries; and (3) surfaces bearing texts carved in relief. Some of the texts were on classics of Buddhist thought inscribed on marble pillars, to which pilgrims doubtless applied sheets of damp paper, daubing the surface with ink so that the parts that stood out in relief showed up; some were religious seals evidently used to transfer pictures and texts of prayers to paper. It was probably this use of seals that led in the 4th or 5th century to the development of ink of a good consistency for printing.

A substitute for these two kinds of surfaces, the marble pillars and the seals, that was more practical with regard both to manageability and to size appeared perhaps by the 6th century in the wood block. First, the text was written in ink on a sheet of fine paper; then the written side of the sheet was applied to the smooth surface of a block of wood, coated with a rice paste that retained the ink of the text; third, an engraver cut away the uninked areas so that the text stood out in relief and in reverse.

To make a print, the wood block was inked with a paintbrush, a sheet of paper spread on it, and its back rubbed with a brush. Only one side of the sheet could be printed.

Early wood-block works

The oldest known printed works were made by this technique: in Japan about 764–770, Buddhist incantations ordered by the Empress Shotoku; in China in 868, the first known book, "The Diamond Sutra"; and, beginning in 932, a collection of Chinese classics in 130 volumes, at the initiative of Fong Tao, a Chinese minister.

Invention of movable type (11th century). About 1041–48 a Chinese alchemist named Pi Sheng appears to have conceived of movable type made of an amalgam of clay and glue hardened by baking. He com-

posed texts by placing the type side by side on an iron plate coated with a mixture of resin, wax, and paper ash. Gently heating this plate and then letting the plate cool solidified the type. Once the impression had been made, the type could be detached by reheating the plate. It would thus appear that Pi Sheng had found an overall solution to the many problems of typography: the manufacture, the assembling, and the recovery of indefinitely reusable type.

In about 1313 a magistrate named Wang Chen seems to have had a craftsman carve more than 60,000 characters on movable wooden blocks so that a treatise on the history of technology could be published. To him is also attributed the invention of horizontal compartmented cases that revolved about a vertical axis to permit easier handling of the type. But Wang Chen's innovation, like that of Pi Sheng, was not followed up in China.

In Korea, on the contrary, typography, which had appeared by the first half of the 13th century, was extensively developed under the stimulus of King Htai Tjong, who, in 1403, ordered the first set of 100,000 pieces of type to be made of cast bronze. Nine other fonts followed from then to 1516; two of them, in 1420 and 1434, were made before Europe in its turn discovered typography.

Transmission of paper to Europe (12th century). Paper, the production of which was known only to the Chinese, followed the caravan routes of Central Asia to the markets at Samarkand, whence it was distributed as a commodity across the entire Arab world.

The transmission of the techniques of papermaking appears to have followed the same route; Chinese taken prisoner at the Battle of Talas, near Samarkand, in 751 gave the secret to the Arabs. Paper mills proliferated from the end of the 8th century to the 13th century, from Baghdad and then on to Spain, then under Arab domination. Paper first penetrated Europe as a commodity from the 12th century onward through Italian ports that had active commercial relations with the Arab world and also, doubtless, by the overland route from Spain to France. Papermaking techniques apparently were rediscovered by Europeans through an examination of the material from which the imported commodity was made; possibly the secret was brought back in the mid-13th century by returning crusaders or merchants in the Eastern trade. Papermaking centres grew up in Italy after 1275 and in France and Germany in the course of the 14th century.

But knowledge of the typographic process does not seem to have succeeded, as papermaking techniques had, in reaching Europe from China. It would seem that typography was assimilated by the Uighurs who lived on the borders of Mongolia and Turkistan, since a set of Uighur typefaces has been found that date from the early 14th century and are carved on wooden cubes. It would be surprising if the Uighurs, a nomadic people usually considered to have been the educators of other Turco-Mongolian peoples, had not spread the knowledge of typography as far as Egypt. There it may have encountered an obstacle to its progress toward Europe, namely, that, even though the Islāmic religion had accepted paper in order to record the word of Allah, it may have refused to permit the word of Allah to be reproduced by artificial means.

Contribution of the Uighurs

THE INVENTION OF PRINTING

Thus, the essential elements of the printing process collected slowly in western Europe, where a favourable cultural and economic climate had formed.

Xylography. Xylography, the art of printing from wood carving, the existence and importance of which in China was never suspected by Marco Polo, appeared in Europe no earlier than the last quarter of the 14th century, spontaneously and presumably as a result of the use of paper. It had been observed that paper was better suited than rough-surfaced parchment for making the impressions from wood reliefs that manuscript copyists used to reproduce the outline of ornamental initial capital letters.

The process was extended to the making of religious pictures. These at first appeared alone and later were accompanied by a brief text. As engravers became more skillful, the text finally became more important than the illustration, and in the first half of the 15th century small, genuine books of several pages, religious works or compendiums of Latin grammar by Aelius Donatus and called *donats*, were published by a method identical to that of the Chinese. Given the Western alphabet, it would seem reasonable that the next step taken might have been to carve blocks of writing that, instead of texts, would simply contain a large number of letters of the alphabet; such blocks could then be cut up into type, usable and reusable.

It is possible that experiments were in fact made along these lines, perhaps in 1423 or 1437 by a Dutchman from Haarlem, Laurens Janszoon, known as Coster. The encouraging results obtained with large type demonstrated the validity of the idea of typographic composition.

But the results were disappointing with regard to type destined for use for text of the usual size. The letters of the roman alphabet were smaller than Chinese ideograms, and cutting them from wood was a delicate operation. Moreover, type made in this way was fragile, and it wore out at least as quickly as blocks carved with a whole text. Further, since the letters were individually carved, no two copies of the same letter were identical any more than when the text was engraved directly on a wood block. The process, thus, represented no advance in ease of production, durability, or quality.

Metallographic printing (1430?). Metallographic impression is more likely to turn out to be the direct ancestor of typography, although the record is far from clear. Several medieval craft guilds, notably the metal founders, the die-cutters, and goldsmiths and silversmiths, were familiar with the technique of using dies. Masters of this technique apparently realized that it could be applied to a process that would enable texts to be set in relief more quickly than by carving wood blocks, probably in three steps: (1) a set of dies, each bearing a letter of the alphabet, was engraved in brass or bronze; (2) using these dies, the text was struck letter by letter to form a mold on the surface of a matrix of clay or of a soft metal such as lead; (3) lead was then poured over the surface to form a small plate that, once hardened, would bear the text in relief.

Steps in metal relief printing

The theoretical advantages of this process were that only one engraving per letter, that of the die, was required to make the letter as often as desired, and any two examples of the same letter would be identical, since they came from a single die; sinking the matrix and casting the lead were rapid operations; the lead had better durability than wood; and by casting several plates from the same matrix the number of copies printed could be rapidly increased.

Metallographic printing appears to have been practiced in Holland around 1430 and next in the Rhineland. Gutenberg used it in Strassburg (now Strasbourg, France) between 1434 and 1439.

But the experiments were not followed up because of problems created by the cast plates. It was difficult to strike each letter die with the same force and to keep a regular alignment, and, worse, each strike tended to deform the adjacent letter. It may well be that the major value of metallographic printing was that it associated the idea of the die, the matrix, and cast lead.

The invention of typography—Gutenberg (1450?). This association of die, matrix, and lead in the production of durable typefaces in large numbers and with each letter strictly identical, was one of the two necessary elements in the invention of typographic printing in Europe. The second necessary element was the concept of the printing press itself, an idea that had never been conceived in the Far East.

Johannes Gutenberg is generally credited with the simultaneous discovery of both these elements, though there is some uncertainty about it, and disputes arose early to cloud the honour.

It is true that his signature does not appear on any printed work. If masterpieces such as the 42-line Bible of

1455 rather than the imperfect products of a nascent typography such as the *donats* of 1445 or the "Astronomic Calendar" of 1447–48 are attributed to him, this is because of deduction and historical and technical cross-checking. The basic assumption is that, since Gutenberg was by profession a silversmith, he would have retained the role of designer in an association set up at Mainz, Germany, with the businessman Johann Fust and Fust's future son-in-law the calligrapher Peter Schöffer. The assumption is based solely on the interpretation of obscure aspects of a lawsuit that Gutenberg lost against his associates in 1455.

Evidence supporting Gutenberg's claim

Apart from chronicles, all published after his death, that attributed the invention of printing to him, probably the most convincing argument in favour of Gutenberg comes from his chief detractor, Johann Schöffer, the son of Peter Schöffer and grandson of Johann Fust. Though Schöffer claimed from 1509 on that the invention was solely his father's and grandfather's, the fact is that in 1505 he had written in a preface to an edition of Livy that "the admirable art of typography was invented by the ingenious Johan Gutenberg at Mainz in 1450." It is assumed that he had inherited this certainty from his father, and it is hard to see how a new element could have persuaded him to the contrary after 1505, since Johann Fust died in 1466 and Peter Schöffer in 1502.

The first pieces of type appear to have been made in the following steps: a letter die was carved in a soft metal such as brass or bronze; lead was poured around the die to form a matrix and a mold into which an alloy, which was to form the type itself, was poured.

Spectroscopic analyses of early type pieces reveal that the alloy used was a mix of lead, tin, and antimony—the same components used today: tin, because lead alone would have oxidized rapidly and in casting would have deteriorated the lead mold matrices; antimony, because lead and tin alone would have lacked durability.

It was probably Peter Schöffer who, around 1475, thought of replacing the soft-metal dies with steel dies, in order to produce copper letter matrices that would be reliably identical.[1] Until the middle of the 19th century, type generally continued to be made by craftsmen in this way.

The typographer's work was from the beginning characterized by four operations: (1) taking the type pieces letter by letter from a typecase; (2) arranging them side by side in a composing "stick," a strip of wood with corners, held in the hand; (3) justifying the line; that is to say, spacing the letters in each line out to a uniform length by using little blank pieces of lead between words; and (4), after printing, distributing the type, letter by letter, back in the compartments of the typecase.

The Gutenberg press. Documents of the period, including those relating to a 1439 lawsuit in connection with Gutenberg's activities at Strassburg, leave scarcely any doubt that the press has been used since the beginning of printing.

Perhaps the printing press was first just a simple adaptation of the binding press, with a fixed, level lower surface (the bed) and a movable, level upper surface (the platen), moved vertically by means of a small bar on a worm screw. The composed type, after being locked by ligatures or screwed tight into a right metal frame (the form), was inked, covered with a sheet of paper to be printed, and then the whole pressed in the vice formed by the two surfaces.

This process was superior to the brushing technique used in wood-block printing in Europe and China because it was possible to obtain a sharp impression and to print both sides of a sheet. Nevertheless, there were deficiencies: it was difficult to pass the leather pad used for inking between the platen and the form; and, since several turns of the screw were necessary to exert the required pressure, the bar had to be removed and replaced several times to raise the platen sufficiently to insert the sheet of paper.

It is generally thought that the printing press acquired its principal functional characteristics very early, probably before 1470. The first of these may have been the mobile bed, either on runners or on a sliding mechanism, that permitted the form to be withdrawn and inked after each sheet was printed.

Next, the single thread of the worm screw was replaced with three or four parallel threads with a sharply inclined pitch so that the platen could be raised by a slight movement of the bar. This resulted in a decrease in the pressure exerted by the platen, which was corrected by breaking up the printing operation so that the form was pushed under the press by the movable bed so that first one half and then the other half of the form was utilized. This was the principle of printing "in two turns," which would remain in use for three centuries.

IMPROVEMENTS AFTER GUTENBERG

Several of the many improvements in the screw printing press over the next 350 years were of significance. About 1550 the wooden screw was replaced by iron. Twenty years later, innovators added a double-hinged chase consisting of a frisket, a piece of parchment cut out to expose only the actual text itself and so to prevent ink spotting the nonprinted areas of the paper, and a tympan, a layer of a soft, thick fabric to improve the regularity of the pressure despite irregularities in the height of the type. An engraving of an early printing shop is shown in Figure 1.

By courtesy of Archiv fur Buchgewerbe

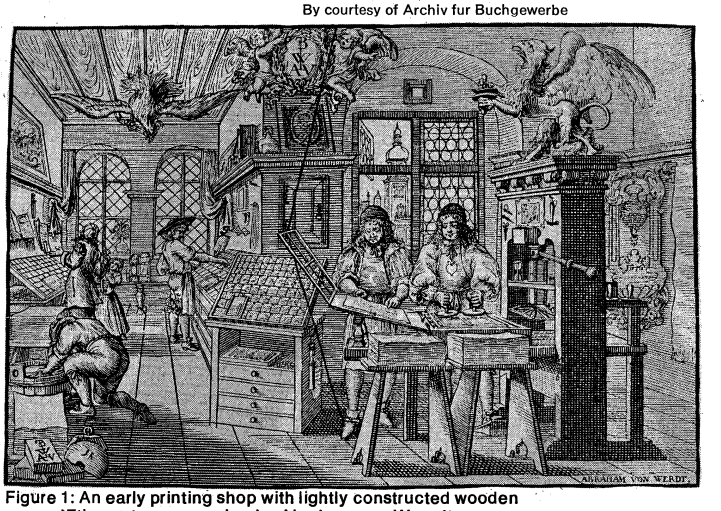


Figure 1: An early printing shop with lightly constructed wooden press; 17th-century engraving by Abraham van Weerdt.

The Dutch press

About 1620 Willem Janszoon Blaeu in Amsterdam added a counterweight to the pressure bar in order to make the platen rise automatically; this was the so-called Dutch press, a copy of which was to be the first press introduced into North America, by Stephen Daye at Cambridge, Massachusetts in 1638.

About 1790 an English scientist and inventor, William Nicholson, devised a method of inking using a leather-covered cylinder (later replaced by a compound of gelatin glue and molasses), the first introduction of rotary movement into the printing process.

The metal press (1795). The first all-metal press was constructed in England in about 1795. Some years later a mechanic in the United States built a metal press in which the action of the screw was replaced by that of a series of metal joints. This was the "Columbian," which was followed by the "Washington" of Samuel Rust, the apogee of the screw press inherited from Gutenberg; its printing capacity was about 250 copies an hour.

Stereotypy and stereography (end of 18th century). An increasing demand for printed matter stimulated the search for greater speed and volume. The concepts of stereotypy and stereography were explored. Stereotypy, used with notable success around 1790 in Paris, consisted in making an impression on text blocks of type in clay or soft metal in order to make lead molds of the whole. The stereotyped plates thus obtained made it economically possible to print the same text on several presses at the same time. The plates left the pieces of type

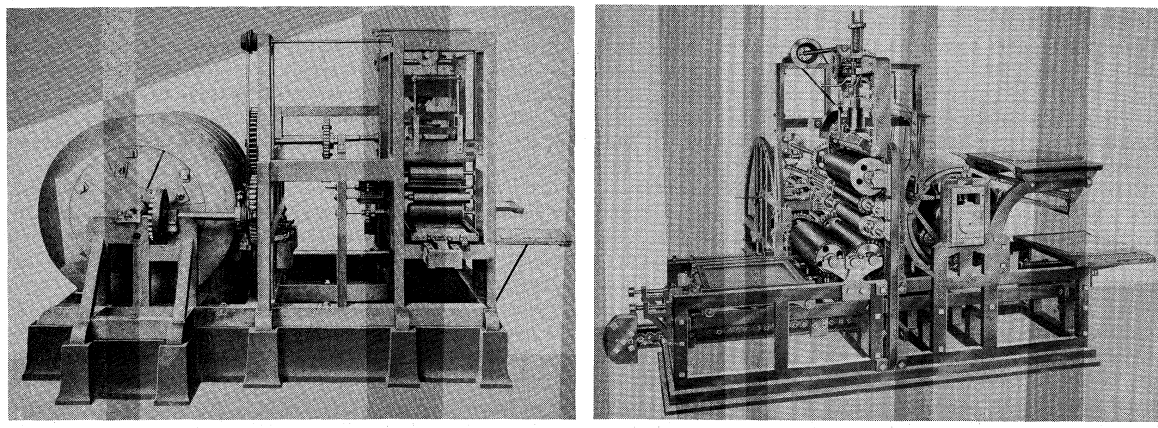


Figure 2: (Left) Friedrich Koenig's mechanical platen press, 1811, and (right) the first stop-cylinder printing machine, 1811, built by Koenig and Andreas Bauer.
By courtesy of Koenig and Bauer AG

in the form immediately available for further use and thus accelerated the rate at which they rotated.

A variation of stereotypy was the application, after 1848, of galvanoplastic metallization, in which process plates of thin metal lined with a base of lead alloy were made by placing a coat of copper on a wax mold of the typeform.

Stereography aimed at bypassing the composition of the type in making the mold. Attempts to perfect the old metallographic method of preparing a clay matrix by stamping with dies brought no better results. In 1797 a variation was tried in which sets of copper matrices of each letter were made in large numbers. The matrices were then assembled according to the wording of the text, so that they covered the whole surface of the bottom of a mold in which the lead plate was then cast. Once the cast had been made, the matrices were available for further use.

Koenig's mechanical press (early 19th century). The prospect of using steam power in printing prompted research into means by which the different operations of the printing process could be joined together in a single cycle.

In 1803, in Germany, Friedrich Koenig envisaged a press in which the raising and lowering of the platen, the to-and-fro movement of the bed, and the inking of the form by a series of rollers were controlled by a system of gear wheels (see Figure 2, left). Early trials in London in 1811 were unsuccessful.

Presses with a mechanized platen produced satisfactory results after the perfection, in the United States, of the "Liberty" (1857), in which the action of a pedal caused the platen to be held against the bed by the arms of a clamp.

Though Nicholson very early took out patents for a printing process using a cylinder to which the composed type pieces were attached, he was never able to develop the necessary technology involved.

The cylinder was in fact the most logical geometric form to use in a cyclical process. It was also the one capable of providing the greatest output. Given an equal amount of energy, the pressure exerted by a platen had to be spread over the whole of the surface to be printed, whereas the pressure exerted by a cylinder could be concentrated on the strip of surface actually in contact with the cylinder at any one instant.

A limited demonstration of the efficiency of the cylinder had been made as early as 1784 on a French press for books for the blind.

In 1811 Koenig and an associate, Andreas Bauer, in another approach to the rotary principle, designed a cylinder as a platen bearing the sheet of paper and pressing it against the typeform placed on a flatbed that moved to and fro. The rotation of the cylinder was linked to the forward movement of the bed but was disengaged when the bed moved back to go under the inking rollers (see Figure 2, right).

In 1814 the first stop-cylinder press of this kind to be driven by a steam engine was put into service at the *Times* of London. It had two cylinders, which revolved one after the other according to the to-and-fro motion of the bed so as to double the number of copies printed; a speed of 1,100 sheets per hour was achieved.

The London *Times* press of 1814

In 1818 Koenig and Bauer designed a double press in which a sheet of paper printed on one side under one of the cylinders passed to the other cylinder, to be printed on the other side. This was called a perfecting machine. In 1824 William Church added grippers to the cylinder to pick up, hold, and then automatically release the sheet of paper.

The to-and-fro movement of the bed that was retained in these early cylinder presses constituted an element of discontinuity; to make the cycle completely continuous, not only would the platen have to be cylindrical but the typeform also. In 1844 Richard Hoe in the United States patented his type revolving press, the first rotary to be based on this principle. It consisted of a cylinder of large diameter, bearing columns of type bracketed together on its outer surface; pressure was provided by several (six, eight, or ten) small cylinders, each of which was fed sheets of paper by hand. This system gave speeds of more than 8,000 copies per hour; its only drawback was its fragility; faulty locking up of the forms caused the type to fall out of the cylinder.

This defect was remedied by applying stereotypy to the production; that is, forming curved plates by making an impression of the typeform on strong pasteboard, the flong, or mat, which was fixed against the inside surface of a rounded mold, which was injected with lead alloy. In France, from 1849 onward, experiments were conducted with this process; it was regularly used in London by the *Times* from 1856 onward and after 1858 was in general use.

But feeding the press with paper still remained outside the mechanized cycle. Mechanization of this step was accomplished by the use of a continuous roll of paper supplied on reels instead of sheets. Techniques for producing paper in a continuous roll had been known since the beginning of the century. The first roll-fed rotary press was made by William Bullock of the United States in 1865. It included a device for cutting the paper after printing and produced 12,000 complete newspapers per hour. Automatic folding devices, the first of which were designed by Bullock and Hoe, were incorporated into rotaries after 1870.

Later, numerous other types of curved stereotype plates were used on rotary presses. These included electrotype plates that are curved before being backed; rubber or plastic plates made by molding or by a photomechanical process; and metal wraparound plates made by photoengraving or electronic engraving.

Attempts to mechanize composition (mid-19th century). Unlike the mechanization of the printing process, mechanization of the composition process was difficult to achieve in the 19th century. The invention of a compression mold in 1806 opened prospects for the mechanization

of the production of type. In 1822 in Boston, William Church took out a patent for a typesetting machine consisting of a keyboard on which each key released a piece of type of the corresponding letter stored in channels in a magazine. The pieces of type thus obtained had to be assembled by hand and the line justified. Church had avoided the problem of distribution and shown an intuition as to its solution by annexing to the magazine a device for constantly casting new pieces of type.

Numerous machines based on the same principle and with the further addition of a mechanism that placed the type pieces selected the right way round appeared in the course of the next 50 years. On one of these the more than 10,000 pages of the ninth edition of *Encyclopædia Britannica* were composed. These machines produced type at the rate of 5,000 to 12,000 pieces per hour, as opposed to about 1,500 by hand composition. But in all of them the type was simply delivered in a continuous row, which had to be divided into lines and justified.

These machines were completed by the introduction of a mechanical distributor, which was a sort of reverse compositor: pieces of type from lines that had been used passed before the operator, who pressed the corresponding key on his keyboard for the appropriate channel in the magazine to be opened up. The speed of mechanized distribution did not exceed 5,000 pieces of type per hour and was, thus, no faster than hand distribution.

Mechanization of letterpress composition faced two difficulties: first, justification, which required intelligent estimation of the size of spaces to be provided between words; and, second, the time taken during which the pieces of type were used for printing, which delay kept composition and distribution from being integrated into one cycle.

Mergenthaler's Linotype machine

Typecasting compositions (1880s). Finally, in the 1880s in the United States, German-born Ottmar Mergenthaler invented the Linotype, a typecasting compositor that cast a solid one-piece line, or slug, from movable matrices of each letter. Each of the matrices was individually notched so that it could return only to its proper slot in the magazine after use. Justification was carried out by inserting wedged spacebands between groups of matrices immediately after making up the words of a given line. Here the matrices rather than type pieces went through the four basic operations of letterpress composition; cast lead was used for printing. The Linotype can produce the equivalent of 5,000 to 7,000 pieces of type per hour.

In 1885, also in the United States, Tolbert Lanston invented the Monotype, which casts individual pieces of type for a line and justifies each line by a system of counting in units the width of the spaces taken up by the pieces of type. The matrices are indefinitely reusable, and the pieces of type, which are used only for the impressions, are returned to the caster. The contemporary Monotype typecaster is controlled by a ribbon of paper perforated on a separate keyboard. It can produce 10,000 to 12,000 pieces of type per hour.

In 1911 another American, Washington I. Ludlow, perfected a typecasting machine for the large display type that bears his name, in which matrices are assembled by hand in a composing stick, which is then inserted above the opening of a mold; the matrices are also distributed by hand.

19TH-CENTURY INNOVATIONS

In the course of the 19th century several important innovations laid the foundation for a number of printing techniques which were not directly related to Gutenberg's invention.

Reproduction of illustrations. The first process for reproducing illustrations was xylography, woodcuts that printed in relief and that therefore could be used together with letterpress, the picture blocks and the pieces of type for texts being locked into the same form. As early as the second half of the 15th century, xylography faced competition from engraving on metal that printed by intaglio; the metal plate (copper, sometimes brass, zinc, and even steel after 1806), engraved with a tool (burin) or etched with acid, was inked and carefully wiped so that ink remained only in the incisions and was transferred to paper under pressure in a cylinder press derived from the rolling mill. Since the intaglio method of printing was not compatible with woodcut printing, sheets of text and of illustrations for the same book had to be printed separately.

Presses for printing curved intaglio-engraved plates were perfected during the 19th century with mechanized inking with the use of rollers and wiping with the use of revolving cloth bands or rotating disks covered with calico. Their printing capacity was limited.

As early as the end of the 18th century, however, intaglio printing had inspired a method for continuous printing of textiles by passing them under an engraved and inked cylinder from which excess ink had been removed by a scraper. In France in 1860 this technique was applied to printing paper for school-book covers. A solid copper cylinder was engraved not with continuous lines but with a multiplicity of tiny cavities in such a way that they retained the ink uniformly despite gravity, centrifugal force, and the action of the scraper. The process was suitable only for simple graphics.

Lithography: Senefelder (1796). A third printing process that had undergone significant development was lithography, neither relief nor intaglio printing but based on the principle that water and grease will not mix. In 1796 Aloys Senefelder of Prague investigated the properties of a stone with a calcium carbonate base and a fine, homogeneous, porous surface. A design drawn on its surface with greasy ink, wetted with water and then brushed with ordinary ink, retained the ink only on the design. This could consequently be reproduced on a sheet of paper pressed against the stone. Senefelder also established that a design drawn on such a stone and printed on paper could be transferred to another stone in as many identical copies as desired, side by side, which made it possible to obtain several copies at a time by printing a single large sheet. He further established that a metal such as zinc had the same properties.

Senefelder envisaged a press in which the stone, secured to an undercarriage, was inked, covered with the sheet of paper with a sheet of pasteboard above it, and submitted to pressure. By 1850 the first mechanized lithographic press with a cylinder, flannel-covered rollers for wetting, and rollers for inking was perfected.

The fact that it was possible to replace the stone by a zinc plate that could be curved made it possible to build rotary presses (the first in 1868) in which the paper passed between the plate-bearing cylinder and the impression cylinder.

Photosensitivity: Niepce (1820s). It was while searching for a means of automatically inscribing an image on a lithographic stone, then on a tin plate, in order to engrave it in intaglio, that Joseph-Nicephore Niepce in the 1820s established that certain chemical compounds are sensitive to light. This discovery marked the origins of photogravure (see also PHOTOENGRAVING) and led to both the invention of photography (between 1829 and 1838) and the use of photographic processes for the printed reproduction of photographs.

In 1852 Fox Talbot, a British scientist and inventor, placed a piece of black cloth (tulle) between the object he wanted to reproduce (the leaf of a tree) and the photosensitive coating spread on a steel plate and obtained a picture that retained the fine mesh of the tulle. Consequently, etching with acid resulted not in an extensive and uniform erosion of an area but in tiny juxtaposed pits all over the photosensitive coating and varying in depth according to the degree of exposure. Talbot simultaneously had invented the screen and also had opened the way for a new development in intaglio printing: rotogravure.

The screen was perfected in the 1880s by substituting for the cloth two sheets of glass with uniform parallel lines that crossed perpendicularly. The screen made possible letterpress and lithographic reproduction of the full range of tones of a photographic document by using the effect of the diffusion of light through the mesh of its grid and

converting the different intensity of tones into the different thicknesses of the printing surface.

Gravure and rotogravure (1890s). The circular mechanization of intaglio engraving, meanwhile, came up against two associated difficulties: the need to engrave an infinite number of tiny cells and the need to engrave them directly onto a cylinder. There were problems, because the rubbing of the squeegee to remove excess ink excluded the use of a curved plate that would not have provided a uniform surface in the area in which it was attached, and it was not possible to get photosensitive solutions to adhere to a cylinder.

In 1862–64 J.W. Swan of Britain invented carbon tissue, which was coated with a photosensitive solution and could therefore be exposed to light before being applied to a metal surface of any shape.

The Klietsch process

In about 1890 a Czech, Karl Klič (also spelled Klietsch), thought of copying a grid screen directly onto carbon tissue, which could be used to transfer the cells necessary for intaglio printing to a cylinder at the same time as the image to be reproduced. In 1895 Klič, with English colleagues, founded the Rembrandt Intaglio Printing Company, which published reproductions of pictures, on paper, by rotogravure. They kept their process a secret.

In a parallel way, patents for a slightly different process, in which the image to be reproduced was screened before making the impression on the carbon tissue, were taken out in Germany and the United States. But a workman from the Rembrandt Intaglio Printing Company emigrated to the United States in 1903 and there revealed Klič's secret, and rotogravure, using his method, became widespread.

THE 20TH CENTURY

Beginning with the invention of the offset technique, the 20th century saw the steady development of innovations in the direction of mass production, speed, and economy.

Discovery of offset (early 20th century). At the same time, lithography was undergoing a new evolution. After the first mechanical presses had been perfected, this process had developed along two lines: (1) printing on thin sheets of metal (for example, tinplate for packaging canned foods) using a transfer process (1878) in which the impression cylinder carrying the metal sheet to be printed did not come in contact with the stone but did with an intermediary cylinder covered with rubber, the blanket, which transferred the image from the stone to the metal; and (2) printing on paper, which was done only comparatively infrequently in the last years of the 19th century, on cylinder or rotary presses.

In 1904 at Nutley, New Jersey, an American printer, Ira W. Rubel, discovered that an image accidentally transferred from the plate cylinder of his rotary to the rubber blanket of the impression cylinder during a paper-feed stoppage could itself be used for printing and in fact produced a superior impression. Rubel and an associate constructed a three-cylinder press, the first offset press, the term since used to describe this constantly expanding printing process.

Dry offset (1920). A few years later a problem arose in connection with printing the background of checks with a water-soluble ink to prevent forgeries. It was proposed that the lithographic plate of the plate cylinder be replaced with a stereotype plate or with a letterpress wraparound plate. This process, which combines the relief of letterpress, which does not require wetting, with the transfer of offset, is known as dry offset, or letterset. Its area of application is not limited to check backgrounds but is used in all areas of conventional printing.

Since 1950 another process has been developed, particularly in the United States, which combines rotogravure with the transfer of offset for printing wallpapers, plastic floor coverings, paper plates, and other products.

Colour printing. As early as 1457 a psalter, which Peter Schöffer signed but which some researchers now attribute to Gutenberg, included, in imitation of the contemporary illuminated manuscripts, paragraphs beginning with ornamental capital letters printed in two colours. This was accomplished by the use of two wood blocks that fitted one inside the other and could be separately inked.

Early colour experiments

Experiments to reproduce pictures in several colours on wood blocks were made in Germany in the 16th century. In the 17th century, different inks were applied to the different parts of the same engraved metal plate in such a way that all the inks were transferred to the paper in a single pressing. In 1719 a painter, Jacques-Christophe Le Blond, took out a patent in England for a process that used the three primary colours, blue, yellow, and red, and black for outlining shapes. Using a dense grid, he engraved four metal plates, bringing out on each plate the relative importance of the colour involved. The same sheet of paper then went through four successive impressions, each in a different colour.

In the 19th century the scientific definition of the principles of trichromatism, the enunciation of the fundamental theories of three-colour analysis and synthesis of colours by photography, the perfecting of coatings selectively sensitive to colours, and finally the use of the screen, instead of Le Blond's hand-drawn grid, established the modern trichromatic technique (which is called quadrichromatic when black is also used).

Automation of composition (after 1929). The search for maximum efficiency had from the very beginning posed the problem of both the mechanization and the automation of composition. The Monotype system, with its separation of keyboard and caster, had constituted one approach to the solution, since the same caster could work at full speed when fed with perforated tapes produced on several keyboards.

The perfection of teletypesetter remote-control composing equipment in the United States by about 1929 permitted widespread application of the principle of separation of human function on the one hand and mechanized function on the other. The operator produces a tape on which each letter, symbol, and space is represented by a combination of perforations. A translator device reads the tape and, according to each combination of holes, orders the release of the necessary matrices for letters, signs, and justifying spaces. Machines casting one-piece fully spaced lines or slugs are able to produce more than 20,000 characters per hour.

Programmed composition (1950s). The production of the perforated tape remained relatively slow, however, because of the time taken by the operator to decide where to make the division in a word at the end of a line. Eventually, in the second half of the 20th century, electronics provided the means of automatically making this decision.

In the 1950s the BBR system, named by the initials of three inventors in France, introduced programmed composition. Starting with a perforated tape continuously produced by the operator, a computer takes over the task of determining the length of lines, the places where words are to be divided according to grammatical rules and typographic usage, the integration of corrections, and even the presentation of the text according to the layout. The speed at which a final tape bearing this information can be produced is limited only by the performance of the perforator, which is the outlet device of the computer. Operating speeds have exceeded 300,000 characters per hour, or ten times the capacity of the most modern slug-casting machines.

Introduction of magnetic tape

During the 1960s, perforated tape began to be replaced by magnetic tape, which is even more rapidly made, at a rate of about 1,000 characters per second, or 3,600,000 per hour. Although magnetic tape is useless for mechanical composers casting pieces of type or lines in lead, such speed is practical for other types of machines not burdened with the weight of lead and the inertia of their mechanical devices.

Photocomposition. In preparing cylinders for rotogravure, offset plates, or letterpress wraparound plates, it is illogical to use the vast weight of lead in letterpress composition to produce a reproduction proof that will then be photographed. Before the end of the 19th century this consideration led to consideration of machines for composing headings by photographing the images of the let-

ters in succession. In 1915 the Photoline, a photographic equivalent of the Ludlow, assembled matrices of transparent letters in a composing stick in order to film each line of the heading.

First generation of phototypesetters: mechanical. The next idea to be tried involved the adaptation of existing typesetters by replacing the metal matrices with matrices carrying the image of the letters and replacing the caster with a photographic unit. The industrial application of this idea resulted in the Fotosetter (1947), a phototypesetter, and its variant the Fotomatic (1963), controlled by a perforated tape, both derived from the Intertype slugcasting machine; the Linofilm (1950), derived from the Linotype; and the Monophoto (1957), derived from the Monotype. Retaining the mechanical limitations of machines intended to shape lead, they could not achieve appreciably higher rates of performance. Photocomposition had to be rethought in functional terms. This approach was explored as early as the 1920s with the Uher typesetter, built in Germany, whose photographic matrices were attached to a rotating disk.

Second generation of phototypesetters: functional. The second generation of phototypesetters is characterized by the maximum elimination of factors of inertia, the number of moving parts having generally been reduced to only two: a disk or drum carrying the photographic matrices and revolving permanently and an optical device of prisms or mirrors allowing directional action on the stream of light provided by an electronic flashtube.

The first revolutionary application of this notion was the Lumitype, invented as the Lithomat in 1949 by two Frenchmen, René Higonnet and Louis Moyroud, who applied for patents in the United States. Executed entirely by phototypesetting, *The Marvelous World of Insects* was done on their machine in 1953. The first model had an attached keyboard. On later models with a separate keyboard production speed exceeded 28,000 characters per hour.

A new Linofilm (1954), functional and electronic, was fitted with a device for selecting matrices by the action of the blades on the photographic shutter, producing 12 characters per second or 43,200 per hour. The model that succeeded it (1965) is equipped with a drum and is capable of double the output.

The Photon-Lumitype 713 (1957) also performs at the rate of 70,000 to 80,000 characters per hour. But at this speed the technique of using a rotary matrix case reaches its limit because of the problems posted by centrifugal force.

The Lumizip 900 (1959) introduced a further revolutionary change by retaining as moving parts only the lens, which scans in a single movement the fixed series of light matrices so as to photograph at one time the whole line of 20 to 60 letters. Output reaches 200 to 600 characters per second, or more than 2,000,000 per hour; this machine required magnetic tape.

The first book composed with a Lumizip, the *Index Medicus*, published in the United States in 1964, was as much of a landmark in the history of phototypesetting as the 42-line Bible had been in printing. Its more than 600 pages were completed in 12 hours. To produce the same work on a typecasting machine would have taken almost a year.

Third generation of phototypesetters: electronic. But magnetic tape was capable of production speeds even higher than those obtained on the fastest phototypesetters. To narrow this gap, a third generation of phototypesetters appeared in the 1960s, in which all mechanical moving parts were eliminated by omitting the use of light and therefore omitting the moving optical device responsible for operating in its field.

Cathode-ray-tube phototypesetters (RCA, Linotron, etc.) operate on a principle analogous to that of television: a narrow pencil of electrons analyzes an image matrix of each letter and commands the modulation of another pencil of electrons on a luminescent screen, which leaves an impression on photographic film. Performance exceeds 500 characters per second and even approaches 1,000, or over 3,000,000 per hour.

Digiset, a German development that appeared in 1965, pushed the use of the electron to its logical conclusion by suppressing even the image matrix of the character, simply keeping the binary analysis of its design available in its magnetic memory; this is, in fact, all that is needed to modulate the pencil of electrons on the final screen.

Speed of phototypesetters

Phototypesetters of this kind (called alphanumerical) have theoretical performance rates exceeding 3,000 characters per second, or more than 10,000,000 per hour, and should be able to approach 30,000,000.

Speeds such as these exceed even the production rate of magnetic tape. Consequently, to work at its most efficient output, such a typesetter must be directly connected to a computer with a similarly high rate of output.

Toward direct impression. The number of characters thus composed now approaches the number of characters printed on a press in the same time (for example, 5,000 sheets per hour, each with 100 lines of 60 characters per sheet, equals 30,000,000 characters). The narrower this gap becomes, the more a still further revolutionary possibility looms—that of eliminating the press itself, since the typesetter can be made to deliver each page as quickly as the press would have done.

To accomplish this it would be necessary only to replace the photographic film that the photographer imprints when it is conventionally used, with an inexpensive carrier capable of receiving an image in the same way without pressure.

Several pressureless printing processes have already been perfected. In 1923 an electrostatic onset system drew the ink of a cylindrical typeform to the paper by means of an electrical charge. In 1948 two Americans conceived another type of electrostatic printing in which the colouring agent is not ink carried on a typeform but a powder or a solution sensitive to the pull of an electric charge inscribed in a plate. This technique gave birth to xerocopy in office duplicating (see below *Office printing*) and, at the industrial level, to xerography for producing posters and maps.

Printing without pressure can also be accomplished on papers impregnated with photosensitive preparations and passed in front of a cathode-ray screen of a phototypesetter. The first experiment using this facsimile printing process was carried out in Japan in 1964 by the *Mainichi Shimbun*, a Tokyo daily newspaper. The image of the newspaper page formed on the cathode-ray screen was transmitted by radio waves, as in television. It was reproduced using the electrostatic system, which does not require chemical treatment of the paper after its exposure.

The Tokyo newspaper experiment

Serigraphy and collotype: a renaissance. Parallel to the evolution of the three major printing processes, letterpress, offset, and lithography, various other techniques have experienced a similar evolution, which has allowed them to survive or to establish themselves in the course of the 20th century and to preserve or win a place in printing.

The art of reproducing a design by forcing ink through the mesh of a silk screen partly blanked out with a stencil plate (serigraphy) had been practiced by the Chinese and Japanese long before the invention of letterpress. In the 19th century the textile manufacturers of Lyons, France, adopted it for printing textiles. In the 1930s in Great Britain and the United States the most varied materials (glass, wood, plastic) and even the most varied shapes (round objects, for example) were printed by serigraphy, which from a handcraft progressed to an industrial technique, with the screen prepared by photosensitization and printing carried out by semi-automatic or automatic machines.

Another process, patented in France in 1855 under the name Photocollography, was modified in 1865 under the name Phototypy (still used in France) and in Germany in 1868 under the name Albertypy (still used in Germany). This process used photosensitive substances not as agents in making plates for printing but to serve directly as the effective surface of such plates. Known elsewhere as the collotype process, the technique was in great favour between 1880 and 1914, was then neglected, and has recently been revived and mechanized for printing posters and transparencies in black and in colour.

Flexography is a letterpress process using rubber plates on the plate cylinder; it occupies a special place in printing on account of the fluidity of its inks. The first patents were taken out in England in 1890, and it was perfected in Strassburg a few years later. Flexographic printing is only of medium quality but is particularly suited to relatively coarse surfaces (pasteboard, wrapping paper, plastic or metal film). It can be carried out by sheet-fed machines but is chiefly used on powerful rotaries.

Three-dimensional printing (1960s). In the 1960s a three-dimensional print was developed, essentially an illustration bearing two views, superimposed, of the same image taken from slightly different angles, on a transparent mount striped with a multitude of imperceptible parallel strips (Xograph process.) On account of these strips, each eye, looking at the print from a different angle, sees only one of the two images selectively. The three-dimensional illusion is produced when this binocular vision is interpreted by the brain.

Office printing. The development of industry and commerce, in the 19th and 20th centuries, accompanied by an increase in administrative activity, created a demand for an abundance of printed information at various levels. In the field of office printing the first tool was the typewriter, perfected in 1867 (see TYPEWRITER). Thereafter, machines appeared that would reproduce large or small numbers of copies of typewritten texts and, later, texts or illustrations of every kind. Some of these machines rely on techniques very close to those of conventional printing; others turned to original techniques that were in turn extended into modern printing. In 1881 in England appeared the stencil duplicator, basically similar to the serigraphic technique. In 1900 a photocopying machine invented in France opened the way to facsimile printing. The offset printing process spread into the area of business printing with small offset duplicating machines whose simplified methods for preparing plates had repercussions that led to the simplification of the preparation of normal plates.

Stencil, facsimile, and xerography

The application of the electrostatic printing process to xerocopy, perfected in 1938, has since been taken over by industry.

All the various processes of duplication and reproduction of documents make up reprography, a name bestowed during the first congress devoted to these techniques, which was organized at Cologne in 1963. If its boundaries with conventional printing are poorly delimited, to the extent that reprography can compete with conventional printing when a medium number of copies are concerned, reprography nevertheless represents an original field. In response to the increased need for quality reprography, the typewriter has been improved since the 1950s and given the capability of providing justified composition suitable for conventional printing.

II. Modern printing techniques

COMPOSITION AND TYPESETTING

Mechanical composition and typesetting. In the first decades of the 20th century all type was set and composed into columns and pages by hand or by mechanical means. These methods are still widely used.

Letterpress composition by hand. The font, which constitutes a complete set of characters of a given typeface, with duplicate numbers of each letter in proportion to the frequency with which each is used, is stored in the compartments of a case; capital letters, proportionately less frequently called for, are in the upper compartments, whence their name, uppercase, and the small letters in the lower compartments, which are more easily accessible and whence their name, lowercase.

The typographer works standing in front of the case. His principal tools are the composing stick, a metal angle iron with one fixed end and a "knee" with a screw or lever for locking; the line gauge, a ruler graduated in units of typographic measurement; and tweezers.

He locks the knee of the composing stick at the justification; that is, at the length of the line to be composed. Against the inside edge of the stick he places a lead, a strip of nonprinting lead alloy that later enables him, using a second lead, to grip the finished line in order to remove it from the composing stick. Holding the composing stick in one hand, he uses the other to select the individual type characters from the case. He can tell by touching which way up they should go, thanks to a nick indicating the top or bottom of the body (the bottom in English-speaking countries and Germany; elsewhere, the top), and he places them side by side in the composing stick. Having completed the proper number of characters to fill the length of the line with a whole word or at the correct division in a word, he adds as necessary to the nonprinting pieces already in place to mark the spaces between the words until the exact justification is obtained.

Having composed and justified the line, the typographer takes it, gripped by its two leads between the thumb and forefinger of both hands, to place it in a galley, a wooden or metal tray with a raised edge on two or three of its sides.

Semimechanized composition. The Ludlow is considered a combination machine; though it automatically casts slugs, it is related to hand composition by the way the matrices are assembled. The matrices are bronze blocks bearing the letter or sign engraved in intaglio on their lower side and with two shoulders on their upper side.

Operations of a Ludlow caster

The composer gathers them individually from the case, which is one of the drawers of a desk, and arranges them side by side in a special composing stick. This steel composing stick is hollowed out in the middle to receive the matrices supported on their shoulders with an adjustable stopscrew for fixing the length of the line. Justification is ensured by blank unengraved matrices in various sizes equally distributed between the words.

The caster resembles a steel workbench with a hollowed-out slot on its surface in which the composing stick is inserted with the matrices face down. A lever starts the casting process by turning on an electric motor. A mold with an opening rises and positions itself under the aligned matrices; a plunger in the melting pot containing the molten alloy forces enough alloy into the mold to cast one line; casting is completed in less than ten seconds, the mold withdraws and releases the solidified line, and the lever, which releases the composing stick, rises automatically.

Since the body size of the font is a uniform size, the upper part of characters whose body size exceeds its measurement projects beyond each side and has to be supported, when it is being used, with leads.

Since the width of the slugs is also uniform, when shorter lines are being cast the composing stick is furnished with thick, blank matrices; once cast, the line is clipped off to the proper length. For longer lines, composing sticks are used with justifications in multiples of that of the mold. Fractions of the line are cast one after another and fit together exactly. The Ludlow is used especially for casting lines of large type for use as titles and subtitles, using typefaces varying from 12 to 144 points (one point equals 1/72, or 0.0138, inch).

The Ludlow caster is complemented by an Elrod caster. This automatically casts nonprinting leads and rules, narrow pieces of nonprinting alloy; both items come in various thicknesses.

Another type of mixed typecaster with manual assembly of the matrices is represented by the All-Purpose Linotype, a sort of Linotype from which only the casting part has been retained (see below *Linotype*). It is used primarily in United States printing establishments. An Italian equivalent, the Nebitype, is used, though less widely, in Europe.

Mechanical composition: slugcasting typesetters. The Linotype and Intertype slugcasting typesetters produce lines of letterpress composition in a single operation, starting with the assembling of the movable matrices. The letter matrices are thin, brass 19 × 32-millimetre (0.7 × 1.3-inch) plates, with two ears and a system of 14 notches arranged in a V on the upper surface and two heels in their lower part. The letter is engraved in intaglio on the face surface; usually two copies of the same letter are superimposed (duplex matrices)—one normal, or ro-

man, the other a variant, either italic (sloping design) or boldface (stronger design). Thus, their thickness varies according to the letter and the body of the character.

The set of matrices is stored in a magazine, a flat, trapezoid metal box consisting of 90 channels in which the matrices are aligned one behind the other, duplicates of 20 or 24 for each letter or sign, lying face down, resting on an ear and a heel.

Introducing blanks into the line is done in two ways: either by using unengraved blank matrices, included in the magazine in three standard sizes, or by using spacebands designed to ensure justification.

The operator sits in front of a keyboard with 90 keys, corresponding to the channels in the magazine, on the left the lowercase letters, on the right the uppercase letters, in the middle the small capitals, numbers, and various symbols. A special bar operates the release of spacebands.

The nine steps in slugcasting

Slugcasting typesetters function as follows: (1) Touching a key releases the matrices, which are brought in proper order on a conveyor belt to a composing stick made of slide-bars and held by their ears. The spacebands, which are stored directly above the composing stick, fall into place between the words. (2) When the matrices and spacebands in the composing stick visibly take up the amount of space planned for the length of the line, the operator completes the line, either with a whole word or by dividing the last word, and pushes a lever to move the line. Since the remaining operations are done automatically he can go on to set the next line. (3) The assembled matrices and spacebands are moved three times in succession: vertically upward on the composing stick; sideways to the left on a transfer slide rest; vertically downward on an elevator that puts them in front of the opening of a mold mounted on a cogwheel called a mold wheel, connected to an electric melting pot containing the molten lead alloy. (4) A justifying hammer forces the long pieces of the spacebands upward, forcing them to separate by equal spaces until all the matrices and spacebands are locked between two steel jaws fixed at the precise justification of the line (see Figure 3). A piston plunges

By courtesy of Harris-Intertype

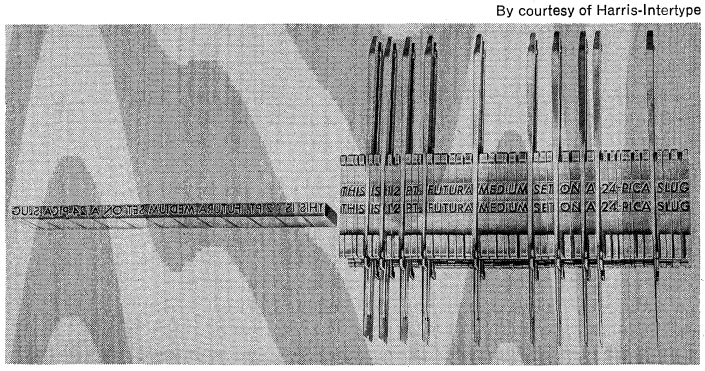


Figure 3: Lineup of matrices and justifying spacers (spacebands) and a cast line of type in a linecasting operation (see text).

into the melting pot and forces the alloy into the mold to cast the line. (5) While the mold wheel rotates three-quarters of a revolution and the solidified line is finished to its exact letterpress height before it is ejected into a galley, the matrices and spacebands are again moved upward by the elevator. (6) They are pushed to the right toward a triangular bar bearing 14 grooves corresponding to the 14 notches in the matrices. (7) Raised by a catcher arm, this bar removes the matrices, which are caught by their notches; the unnotched spacebands are released and immediately return to the place where they are stored. (8) When the catcher arm is at its highest position, the matrices are pushed to the right toward another triangular bar with 14 grooves along its length and flush with the top part of the magazine; this is the distributor bar. (9) The matrices move along the distributor bar until at a certain point the arrangement of grooves ceases to provide support for the notches, which of course are different for each letter or sign. Each letter's matrix is then released at the opening of its own channel in the magazine.

The automatic cycle of the typesetter is controlled by several large cams mounted on a single shaft driven by an electric motor.

Modern typesetting machines are equipped with several magazines of varying type sizes that can be used alternately. Some so-called double-distribution machines permit two magazines to be used at once by pressing a supplementary key.

The performance of recent models has been improved by accelerating the revolution of the matrices, intensifying the cooling system of the mold, and increasing the number of molds on the mold wheel to six.

The slugcasting typesetter, which furnishes solid, easy to handle, composed type, is particularly suited to printing newspapers. It has the disadvantage that to correct any error, however trivial, the whole line must be recomposed.

The All-Purpose Linotype is a combination manual and automatic machine that retains only the casting part of the Linotype. Special matrices, solidly rectangular or with notches, ears, and heels, are assembled by hand in a composing stick. Justification is done with blank matrices of various sizes. The line of matrices, held by the composing stick, is placed against two set squares fastened to the bedplate of the machine and manually pushed on a slide rest, which takes it to the elevator. The elevator places the matrices in front of the opening to the mold for the casting operation, which delivers the slug. The matrices are then distributed by hand.

Typesetter casting aligned characters. The operation of the Monotype typesetter, which casts individual aligned characters, is based on a system of measuring the width of characters, called the set. In each font, letters and symbols have sizes determined in units of set, from five units for the narrowest, such as the "i" or the "l," to 18 units for the largest, such as the "W" or the "M."

The Monotype keyboard

The Monotype keyboard, separate from the caster, consists, on the standard model, of 274 keys, 30 of which, in two rows numbered from 1 to 15, are called justifying keys. Typing out the text results in the perforation, by an automatic punch, of one or two holes, each letter or symbol having its own pattern, in the width of a paper tape that allows 31 possible arrangements. An automatic calculator adds the widths of the letters and symbols typed out. Moving his forefinger across a scale that he has before him, the operator knows when the end of a line is near. Once he has finished a line, he places his other forefinger on the justifying drum, which indicates which two justifying keys he must now press. This move results in the perforation of one or two holes whose position indicates the quotient of the number of units lacking for the line to be completed by the number of spaces between the words of this line, plus a third hole in a position specifically fixed for the justification process.

The typesetter is composed of an electric melting pot containing molten alloy situated under a mold in the shape of a vertical chimney, the internal dimensions of which can vary according to the measurement in units of set of the characters or spaces to be cast.

The matrices are in the form of small bronze cubic blocks measuring five millimetres (0.2 inch) square, arranged in a steel frame nine centimetres (3.5 inches) square containing 15 rows of 15, enough for five complete alphabets, typically uppercase and lowercase in roman, italic, and boldface, and small capitals, as well as double or triple letters, numbers, and punctuation marks. Each row includes only matrices of letters and symbols of the same unit of set, from the smallest (five units) in the first row to the largest (18 units) in the back row.

The frame can slide horizontally in either direction to place any matrix from any row above the opening to the mold.

The process of setting and casting type on the Monotype is as follows: the roll of perforated paper tape is placed in the pneumatic tower of the typesetter—a row of 31 pipes circulating compressed air. As it unrolls, the tape prevents the compressed air from entering all pipes except those corresponding to the perforations.

The tape unrolls in the direction opposite to that in

which it was rolled up—that is, in the opposite order to the way it was typed, the last line appearing first and the justifying perforations being inserted into the pneumatic tower before those for the letters and symbols. The compressed air that the perforations allow to pass into three (or two) of the pipes causes pieces of metal (justifying quoins) to fall into position in such a way as to control the internal measurement of the mold each time that the spaces between the words of the next line are to be cast.

The perforations for each letter or symbol allow the compressed air to pass into two (or only one) of the pipes connected to two blocks, each of which also contains a series of graded pins. Compressed air raises a pin in each block and halts the movement of the matrix frame in either of its sideways movements. In this way the matrix's row and the matrix's place in its row are selected. Selecting the row is the same as selecting a measurement given in units of set. Positioning the row is automatically linked to setting a piece of metal (the set quoin) whose position regulates the dimensions of the mold for casting the letter or symbol.

For the casting, a centring device places the selected matrix precisely against the opening of the mold, and a plunger in the melting pot forces the alloy up to cast the character or the space (that is, the unengraved matrix).

The composed line emerges from the machine completely assembled and justified and is placed in a galley.

The Monotype can cast type ranging in body size from five to 24 points (with a special mold for each). The addition of a speed-reducing device enables it to cast in 48 points. The maximum width of lines assembled is 60 picas.

In the early 1970s Monotype models could be equipped with a frame carrying 15 rows of 17 matrices (255) or 16 rows of 17 (272), with six or seven complete alphabets. The keyboard then has 310 keys.

A special model of the keyboard permits simultaneous perforation of two tapes for composing the same text in identical or different kinds of type and lengths of line.

Advantages of Monotype

The advantages of the Monotype system are the quality of its composition and the ease with which corrections can be made without having to reset the whole line. It is not well suited to newspaper printing because of the difficulty of handling lines of movable type and because, since typecasting begins with the end of the tape, composition must wait until all the type has been cast.

Automatic composition (perforated tape). The Teletypesetter (TTS) system extends to slugcasting machines the principle of separation of function originally characteristic of the Monotype: it enables Linotype or Intertype machines to be controlled by a perforated tape produced on a separate keyboard, even situated in a different city, since the combination of the perforations on the tape can be sent telegraphically.

The Teletypesetter tape is six-channelled; that is, it contains six possible positions for perforations across its width. This allows 64 different combinations of from one to six perforations.

Ths limited capacity, less than the number of keys on the keyboard of the typesetter, is corrected by an arrangement whereby each combination of perforations may have two different uses (for example, the uppercase and the lowercase of the same letter) according to whether it follows one or other of two special signals (themselves represented by combinations of perforations) that control passage of one or other of these uses.

The Teletypesetter keyboard

The keyboard for preparing the Teletypesetter tape looks like a typewriter with, in addition to the usual 44 keys and the space bar, 20 special keys. Striking each key establishes contact with the electric circuit or circuits that operate the perforators and at the same time acts on a calculating mechanism: a needle moving across a screen warns the operator of the end of each line.

Usually, as the tape is perforated the text is also typed out on a sheet of paper, which allows the work to be checked, reread, and corrected. For use with a Teletypesetter, the typesetting machine is equipped with a mechanism that translates the tape. In this mechanism the tape passes under six sensors that register electric contacts as the perforations pass. In accordance with the combination of electric contacts thus established, relays control the action of the keys or of the bar that causes the spacebands to drop and, at the end of each line, the starting of the casting cycle.

The most recent typesetters specially designed for use with the Teletypesetter offer such technical refinements as the elimination of the composing stick and immediate dispatch of the line to the elevator, simplifying the path taken by the matrices; and electromagnetic, rather than mechanical coupling, which speeds the starting up of the casting cycle.

Programmed composition (prepared by computer). The use of a computer eliminates manual intervention in preparing the perforated tape, in assessing the length of the lines, and even in deciding how to end them; *i.e.*, whether by completing or dividing a word.

Normally, the operator types out a continuous tape called idiot tape in the United States (kilometre tape in France) without concern for the length or division of lines. This band is inserted into the computer's input device, a tape scanner, which operates by means of either electric sensors or photoelectric cells and converts letters, signs, and orders into combinations of electric impulses. The computer semi-automatically or automatically processes this raw information in accordance with its programmed instructions and immediately communicates the result to an electromagnetic perforator, the output device, that produces a second tape, like the first but that also bears, in the proper places, perforations ordering the ends of lines.

A general program establishes the operation of the computer in its application to the work of composition. Individual programs adapt it to the machinery of the company concerned (models of typesetters, available magazines of matrices) and to the kind of work carried out (usual length of line, method of indenting paragraphs, etc). Finally, special instructions punched on the tape by the operator at the same time as the text can interrupt the execution of the programs registered in the computer with directions valid for this text alone, in its entirety or in certain parts: choice of typeface among those available, transition from one kind of typeface to another available for the same typesetter, length of line and changes in the length, alignment to right or left, squaring of lines, indentations for ornamental capitals, spaces for borders or illustrations, and other details.

Having identified the combinations of the perforations on the tape and separately retained the service signals addressed directly to it, the computer proceeds to estimate the amount of the space occupied in a line by its letters and symbols, referring to instructions registered in its memory regarding each. In the same way, it determines a justification zone in which a division in the line is necessary and possible, the minimum and maximum limits of this zone being fixed by the limits of expansion of the spacebands on the typesetter.

If the end of a whole word comes within the justification zone, the computer itself signals the end of a line after this word and suppresses the space that would normally follow it. Otherwise, the last word must be divided. The process is said to be semi-automatic if a special operator, seated at a keyboard linked to the computer, must intervene to decide where to place the division in the word submitted to him on, for example, a cathode-ray viewing screen. The process is automatic if the computer is designed and programmed to make the decision itself; the operation then is carried out by starting a subprogram in which all the divisions possible in the word considered are listed (after a prefix, between syllables). This list is tested against prohibited divisions (according to the rules of etymology, phonetics, typography) stored in the rapid-access memory of the computer. From among the positions that are not eliminated during this test, the computer chooses the one situated nearest the end of the word. It automatically inserts the signal for the hyphen and orders the end of the line.

Computerized division of words

The computer can also carry out the correction of mistakes before composition. Various methods are possible,

of which two will be described briefly. In one, the perforation of the justified tape delivered from the computer includes the introduction at the beginning of each line of a numbered signal and is accompanied by a proofing copy of the text with a corresponding reference number for each line. When the mistakes have been corrected on the proofing copy, an operator prepares another, much shorter correction tape, which consists of the corrections preceded by the reference to the line on which they occur. The justified tape and the correction tape are jointly introduced into a double reader, the mixer, which determines anew the length of the line and where the division should occur, as well as for such succeeding lines as need to be modified, before producing a final tape.

In the second method, the proofing copy can be typed out or shown on a cathode-ray viewing screen with the lines numbered but without the tape. At a keyboard connected to the computer, an operator types out the corrections, preceded by their line reference. If a viewing screen is used, the text reappears immediately in its corrected form, and the output perforator immediately delivers a justified and corrected tape.

The computer is usually programmed to sort out and correct even mistakes or anomalies in typing, such as the presence of two consecutive spaces, in which case it cancels one. It can, if its capacity allows, receive a makeup program independent of the tape of the text; following the specifications of the layout (positioning and size of headings, text, and illustrations) coded in binary language, the computer itself introduces onto the perforated tape the special instructions concerning kinds of typeface, length of lines, changes in lines, etc.

Because of the quantity of information needed for composition, the six-channel Teletypesetter tape is being increasingly replaced by seven- and eight-channel tape.

Computer processing using a continuously typed tape can be applied equally well to the Monotype system. The programmed operation for dividing lines is in this case carried out by the automatic calculation of the width of the spaces between words and by the perforation, before the end-of-line signal, of a signal signifying the appropriate position of the justifying quoins. To enable the text to be read by the pneumatic tower of the typesetter, a converter transcribes the perforations from a narrow six-, seven-, or eight-channel conventionally perforated tape to the wide tape of the Monotype system.

The use of computers is now widespread in preparing photocomposition jobs, with programs adapted to the specifications. The computer's output device can produce magnetic tape instead of perforated paper tape.

Optical scanning devices

One intake device no longer reads perforated tape but is an optical mechanism for scanning a typewritten text. The Retina reader, for example, is a sort of artificial retina made up of a group of photosensitive units able to identify each letter typed by a special typewriter, using only three data: height, width, and gray value; that is, the surface area occupied by the outline of its design.

Cold type. Cold type is the expression used, particularly in the United States, to describe a simple and economic method of preparing text by machines resembling ordinary typewriters but capable of producing justified lines in type the width of which varies according to the letter involved. Justification is achieved in several ways by different versions of the machine. In the IBM Multipoint, a first typing calculates the total measurement of the type pieces up to the beginning of the justification zone and causes a coded sign to appear. A button is set in position over the coded sign thus assigned to each line before a second, final typing is done. The position of this button determines the automatic adjustment of the spaces between the words to the amount needed to obtain justification.

In the Justowriter, the keyboard on which the uncoded, unjustified proofing copy is typed simultaneously perforates a paper tape with the code for the letters, as well as, for each line, the code for the amount of space between the words as indicated by a calculator. The tape then controls, on a second unit of the machine, the electric typing of the final justified copy.

In the IBM Multipoint with magnetic tapes, a magnetic tape produced at the keyboard is processed by a computer for justification and, if necessary, for corrections. The final tape delivered by the computer controls the action of an output unit, which carries out the final typing.

If the copy thus produced on paper is to be photographed to prepare printing plates by photogravure, cold type cannot be directly incorporated into photocomposition because of the intermediate operation.

Optype is a hybrid process that simultaneously carries out the operation of justifying a text typed directly in cold type and transmits it to photographic film. By means of optical distortion, each line is stretched to the exact length of line projected on the film. The same mechanism also enables the line to be magnified or reduced or set in italics.

Phototypesetting. Using phototypesetting, a direct image of the text is obtained, positive or negative, according to need, on a photosensitive, usually transparent surface by exposing the surface to light through transparent matrices, negative or positive, of the letters and symbols.

Manual phototypesetters. Several small machines permit phototypesetting of short texts and titles in conditions to a greater or lesser degree short of automation. Among them are the following:

Dantype uses separate transparent plastic matrices, which are assembled in a composing stick and placed in direct contact with the photosensitive film inside the machine.

Typro makes use of letters and symbols on a negative film that moves to and fro to place the desired type piece in contact with the photosensitive film.

Headliner incorporates letters and symbols that appear in negative on an interchangeable plastic disk whose position is controlled from outside. The film is exposed by contact.

Hadego uses plastic matrices assembled in a composing stick, exposure taking place through an adjustable photographic lens that permits enlarging or reducing. With just two series of 350 matrices, one with a 20-point body, the other with a 48-point body, all sizes of type from eight to 110 points can be obtained.

The Starlettograph, comparable to an ordinary photographic enlarger, can be used only in a darkroom. The type, inscribed on a semirigid plastic tape, is set in position one piece at a time, using red light that does not affect the photosensitive film.

Letterphot works on the same principle as the photographic enlarger but on a luminous table. A first projection is made of all the characters of a line without the sensitive surface. Then the sensitive surface is placed on the luminous image of the line, which appears transparently and cannot therefore make an impression. Letters are successively printed in a two-part operation. First the letter is projected in normal light to cause it to coincide with its luminous image; the normal light does not make an impression on the sensitive surface, because the latter has a special composition. After this adjustment has been made, the letter is projected in actinic (photographically active) light, which exposes the sensitive surface.

Diatyp and the Monotype photoheadliner (as well as the Varityper, which is similar in composition) are more elaborate phototypesetters, easier to operate and permitting production speeds of nearly one character per second. The image of each character on the matrix disk is controlled by a symbol that is read by photoelectric cells and which automatically moves the film forward the same amount as the space taken up in the line by the character. A totalizing calculator informs the operator of the rate at which the line is being completed, and justification can be achieved by a first typing without having the source of light in operation; in a second typing, the spaces between the words are adjusted the necessary amount. Adjusting the lens of the Diatyp produces characters ranging in size from four to 36 points and, using the Monotype, from five to 84 points.

Automatic phototypesetting. The first Linofilm was a direct adaptation of the Linotype. Its photographic matrices were the normal Linotype matrices, the only difference being that, instead of bearing an intaglio engraving

of the character on their face, they bore its outline in black on a white background. Lines were composed in exactly the same way as on the typesetter, justification being carried out by expanding the spacebands. The justified line is then passed a single time in front of a lens to be photographed.

Fotosetter operations

The Fotosetter is an adaptation of the Intertype machine but with functional differences. The matrices resemble matrices used for casting; they have the same notching and different thicknesses, depending on the character. But the outline of the character, instead of being inscribed on the face, is a transparency (*i.e.*, a photographic negative), in a capsule set into the level surface of the matrix. These special matrices are called fotomats. In place of spacebands there are space fotomats of different thicknesses.

The Fotosetter is equipped with magazines of 117 channels, 27 more than the typesetters, with an enlarged keyboard of 114 keys.

Once the line has been assembled and justified, using space fotomats of the necessary sizes, the fotomats move inside an optical apparatus that sends a brief flash of light toward the sensitive film. After each exposure, the support of this film is moved slightly sideways by a rack-and-pinion system commanded by the withdrawal of the next fotomat from its alignment; the matrix moves in proportion to the thickness of this fotomat. When all the type pieces in a line have been photographed, the film unwinds the correct amount to present a clean surface ready for the phototypesetting of a new line, while the fotomats are carried off to the distribution bar.

Equipped with a turret of 14 different lenses, the optical apparatus produces 14 sizes of type from three to 72 points, from the same set of fotomats of uniform 12-point size.

The Monophoto is a direct adaptation of the Monotype system with, on the one hand, an independent keyboard that produces a wide perforated tape in the Monotype code and, on the other, a phototypesetter operated by inserting this tape. The type pieces are chosen by positioning a frame, which carries 17 rows of 20 cubelike matrices in which the letter or symbol appears as a transparency, in negative, in the path of a beam of light (see Figure 4). This beam, after proper processing, is directed toward the sensitive film, on which it makes an impression. It first travels through a combination of magnifying glasses and prisms whose position in relation to each other is adjusted to obtain the desired ratio of enlargement or reduction. The sensitive film remains stationary on the drum carrying it as a composed line, while the element that enables the beam of light to move from letter to letter is a set of two mirrors placed face to face at a 90° angle and mounted on a mobile carriage. Before each exposure, this set of mirrors shifts, parallel to the direction of the sensitive film, the same amount of space as the width of the character about to be composed. This amount of space depends also on the number of units of set of the letter or symbol and on the ratio of photographic enlargement or reduction. The movement of the mirrors is thus subjected to the command mechanisms of two factors: the position of the frame, since the matrices are arranged in rows of the same units of set, and the adjustment of the combination of magnifying glasses and prisms.

Justification is accomplished, as on the typesetter, by predetermining the width of the spaces between the words. Since the justification perforations appear before those for the type pieces, they establish for the line to come the amount of space the set of mirrors has to shift at each space command punched in the perforated tape.

After all of the type in a line has been photographed, the set of mirrors returns to its original position, and the drum bearing the sensitive film turns the amount necessary to continue on to the composition of the next line according to the degree of line spacing (leading) chosen.

Using matrices of a single eight-point size, the Monophoto makes available the whole range or type size from six to 24 points. For perfect photographic reproduction it is usually found preferable to use two or three sizes of matrices to cover this range. Given the quality of its

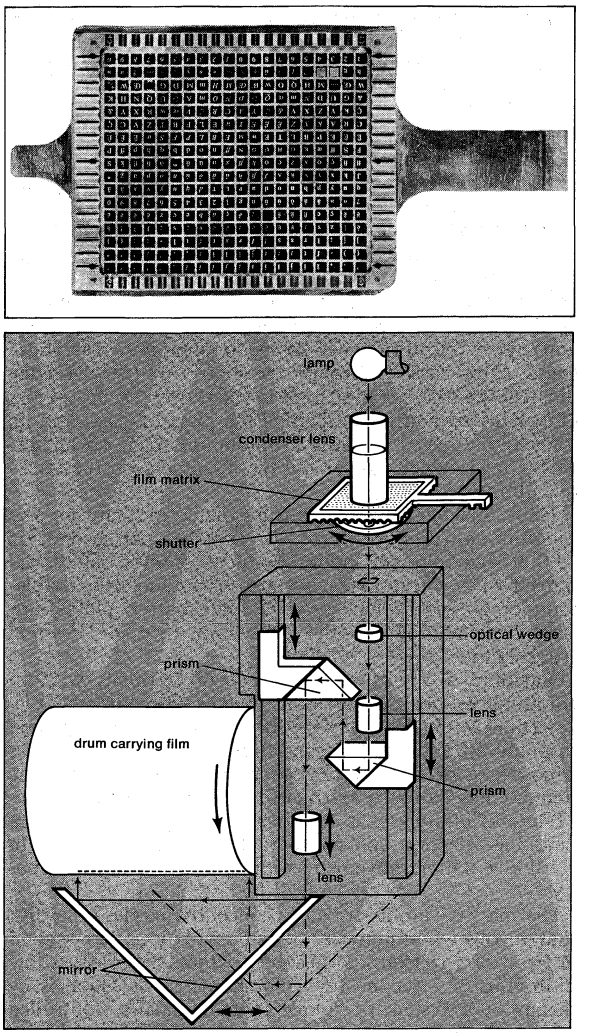


Figure 4: Basic operation of the optical system used in Monotype photocomposition (see text). A film matrix is shown at top.
By courtesy of Monotype

production, the Monophoto, sometimes linked to a unit programmed to prepare the tape, is popular for work that demands careful composition.

Functional phototypesetters. A second generation of phototypesetters consists of functional machines that are analogous neither in structure nor in operation to typesetters using lead. Outwardly they resemble metal chests comparable to office furniture. Their design aims at reducing mechanical parts, inertia, and friction to the minimum. Their technical characteristics vary according to model, use, and cost.

The keyboard, which is hardly more complex than that of an ordinary typewriter, can be attached or separated; in the latter case, information regarding the text to be composed is inserted by perforated tape. Computer units can be integrated, either merely to direct the machine's operations or to ensure completion of the justification process, division of words, and correction, whether from the adjoining keyboard or from a continuous perforated tape. Selection of the matrix image of each character can be done either by using a mobile support for the matrices (plastic tape, disk, drum) in front of a fixed source for the beam of light or by using a mobile beam of light in front of a fixed support for the matrices (glass or plastic plate); alignment of type pieces is carried out in either case by mobile prisms or mirrors.

Versatile capabilities of the phototypesetter

In addition to enlargement or reduction, the optical apparatus can be designed to carry out special effects, such as converting roman to italic type or stretching a line to make it longer or higher. The photographically composed text can be delivered either on paper or on film, in

positive or negative, or in straight reading or reverse reading. The source of light is usually an electronic flash the intensity of which can, if necessary, be made proportionate relative to enlargement or reduction.

Some of the characteristics of phototypesetters are outlined below.

Linofilm (new method): The matrices of the 88 characters in a set are inscribed on a plate of glass that remains stationary during composition. The character is chosen by the shutter of the photographic lens. This shutter consists (as in a commercial camera) of very thin, overlapping metal blades, eight in number. Instead of always opening at the same point at the moment of exposure, it opens facing the desired character, each being set in position by an electromagnet so as to obtain this arrangement. After passing through the matrix of the character thus chosen, the beam of light is taken over by one of the 88 small lenses arranged behind the plate of glass and its trajectory directed towards a mirror mounted on an undercarriage, which carries out the alignment on the sensitive film.

Using this very light electromagnetic mechanism, the Linofilm can produce up to 12 exposures per second, or 43,000 symbols per hour. Eighteen matrix plates arranged in a turret magazine are instantaneously usable, producing 1,584 characters. Three matrix plates are enough to photograph the same type face in 16 sizes, from six to 36 points.

Diatronic, a phototypesetter made in Germany with an adjoining keyboard, uses matrix plates with 126 symbols. Selection is made after the beam of light has passed through all the symbols on the plate, through prisms which take up the position necessary to retain only the light coming from the matrix of the chosen character.

Continuous high-speed photo-typesetting

Photon-Lumitype was the first phototypesetter to introduce the selection and photographing of the character in a rapid circular movement without interrupting continuity.

The matrices are inscribed in concentric circles on a disk that revolves continuously at ten revolutions per second in front of an electronic flashtube whose light lasts a few millionths of a second for each character. Selection is by means of a system of rotary contact makers, controlled by the telegraph system. A nylon drum is integrated with the matrix disk and turns with it in the same movement; the drum is encircled with as many tracks as there are channels in the binary code used to define the characters. These tracks are the transmitting and isolating elements that pass under a row of electric sensors. A special combination of transmitting and isolating elements corresponds to each character matrix positioned ready to be photographed.

Whether by striking a keyboard or by perforating a tape, selecting a given character consists of the precise formulation of the combination that establishes the electrical contact and initiates the flash of light. This selection can be acted on only at the precise moment when, as the disk revolves, the matrix of the desired character moves into position to be photographed.

Whether textual information is fed into the machine on an adjoining keyboard (as on the early Linotypes) or on a keyboard directly connected to the photographic unit or whether it is done on perforated tape, this information is in every instance preserved, line after line, in a memory, formerly mechanical but magnetic on the later models, which at the same time permits calculation of the size of the spaces between the words and ensures that the character's binary signal is presented during the 1/10-second period of time available.

It is possible to attain a production speed of ten symbols per second, or, theoretically, about 36,000 per hour; in practice the figure averages less.

Each of the eight concentric circles on the matrix disk contains two complete sets of 90 characters, which can be filmed in 12 sizes, from five to 72 points. In other words, a total of 17,280 characters are immediately available.

Another Photon-Lumitype model operates on the same principle, but the disk is replaced by a drum revolving 30 times per second around an axis that coincides with the source of light. The type matrices are inscribed in negative on two films carried on the surface of the drum, and the source of light consists of two electronic flashtubes, one for the upper, the other for the lower half of the drum (see Figure 5, right). The lens and mirror system for producing images of the matrices on a moving light-sensitive film is shown at the left in Figure 5.

By courtesy of International Photon Corp.

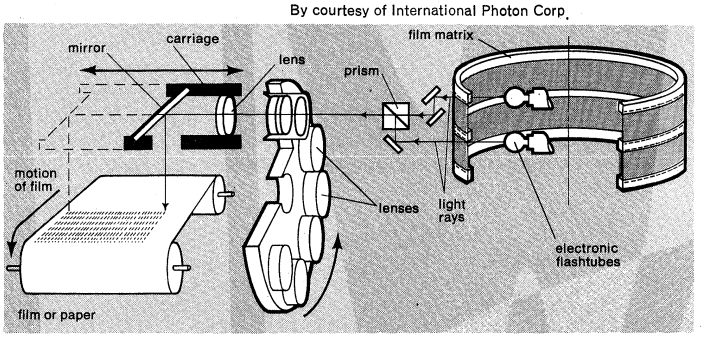


Figure 5: Basic operating principle of the Photon 713-10.

This model's total capacity in characters (four complete sets and eight ratios of enlargement or reduction) is three times smaller, but its speed of composition is three times faster: it can attain 80,000 symbols per hour.

By putting the same films of type matrices in the upper and the lower part of the drum—that is, by cutting in half the number of characters stored—speed of production can be raised to 120,000 symbols per hour.

Europa-Linofilm is similar in design to the Photon-Lumitype just described, with a permanently revolving drum but with an electric type-matrix selection system. These matrices are small individual plates bearing not only the negative image of the letter or symbol but also a series of transparent marks whose arrangement constitutes its binary identification code. In the revolving drum this coded part of the plates passes in front of a scanner made up of a series of photoelectric cells. As soon as the scanner picks up a coincidence between the code of the type matrix and the code of the character selected for composition, it activates the shutter release of the electronic flash.

The Europa-Linofilm drum is composed of four superimposed levels, each containing 120 duplex type matrices —for example, with the same letter in both roman and italic—easily interchangeable in order, since their identification is not linked to their place.

The Photon-Lumizip

Photon-Lumizip is based on a different principle. The performance speed of drum phototypesetters can hardly be increased because of the technical problems posed by the rapid rotation of the drum. To increase speed, the Lumizip abandoned rotary movement. The type matrices are stationary and are aligned in negative on a large-sized plate. There is an individual electronic flash behind each type matrix. The sensitive film is stationary while a line is being composed. The only moving element is the component of the lens placed between the plate and the film, which carries out a rectilinear to-and-fro movement, parallel to them. The flashtube situated behind a given matrix emits its beam of light each time that the component of the lens finds itself, in the course of its to-and-fro movement, in the axis that joins this matrix to the position where the letter to which it corresponds is to appear in the line to be composed on the sensitive film. Thus, the order in which the characters are photographed is neither that in which they appear in the text nor that of the matrices on their plate but that determined by an angular relationship between them both.

A computer is built into the Lumizip. As soon as the coded signals for a line to be composed reach the computer, it determines the order and the exact moment when each flashtube is to operate, synchronized with the movements of the component of the lens.

In practice, the type matrices are aligned on the plate not in a single row but in 11 horizontal rows. The component of the lens always moves in the same horizontal

plane, which is the same as that of the line to be composed and also that of the median row, the sixth, of the matrix plate. Alignment of characters from the other rows is achieved by means of two level, horizontal mirrors placed on either side of this horizontal plane, parallel to it and face to face at a small distance from one another. The beams of light coming from the flashtubes of the matrices of the median row pass between these two mirrors without touching them. But the beams of light coming from the flash tubes of the matrices of the other rows strike these mirrors at an angle that is sharper, depending on how far the row is from the median row, and between these mirrors they are repeatedly reflected, the number of reflections depending on the angle: one for rows five and seven, up to five for rows one and 11, the last reflection falling into alignment on the sensitive film.

Mechanical movement in the Lumizip is reduced to the extreme minimum, because the component of the lens is the only moving part. Since it depends on an alternating, rectilinear movement and is therefore handicapped by factors of inertia, its speed, which cannot be as great as that of a continuous rotary movement, is only ten to-and-fro movements per second. But during a single one of these to-and-fro movements all the several dozen characters for one line are photographed. Thus, the Lumizip attains a performance rate perhaps 20 times superior to that of the Lumitype. Theoretically, it can perform at a rate exceeding 2,000,000 symbols per hour and in practice has produced over 1,000,000.

Electronic phototypesetters. In phototypesetters of the third generation, the beam of light is replaced by a flow of electrons, which offers the advantage that the electrons can be deflected by means of magnetic fields without the intervention of mechanical parts such as mirrors and lenses. Television systems are based on this characteristic, and an early type of electronic phototypesetter is structurally comparable to a closed-circuit television system. A reading device analyzes, by fine scanning, the outline of the matrix of the letter to be composed and converts the luminous information it obtains into electronic signals. The cathode-ray screen of an output device reconstitutes, in accordance with these signals and by a fine scanning device synchronized with the reading device, a luminous image of the letter, which makes an impression, through an optical reducing device, on a photosensitive surface. A computer, depending on the text to be composed, directs the position of the reading device's scanner towards that part of the matrix plate that bears the matrix of the letter selected, and simultaneously directs the position of the output device's scanner toward that part of the screen that corresponds to the position of this letter in the line being composed.

On some models, the scanning device consists of the equivalent of a television camera whose electron beam is selectively deflected towards the chosen matrix and directly analyzes the luminous information coming from it. On others, a cathode-ray tube takes the place of the emitter of a regular beam of light by scanning behind a plate on which the matrices appear in transparency. On the other side of the plate, photoelectric cells collect this beam at the moment it passes through the matrices and react by emitting electronic signals directed to the output device.

Electronic matrix selection

For matrix selection, the face of the emission tube is divided into 16 square sections (four by four), of which only one is illuminated at a time by selective scanning directed towards it. There are 16 photoelectric cells arranged in a square (four by four), only one of which is in operation at a time. Thus, there are 256 (16 by 16) possible arrangements of the chosen section and of the chosen cell. Each combination corresponds to an optical trajectory belonging to one or the other; that is to say, to the precise positioning of one of the matrices over the plate.

On the screen of the output device, the letters have a definition of 650 lines per inch for ordinary work and 1,300 lines per inch for quality work. The line structure is invisible after the letters have been reduced for photographic reproduction.

A more complex model of the Linotron scans, on the screen of the output device, the surface of a whole page, composing as it goes all examples of the same letter in all the places where it occurs on the page. Composition of the page is completed after all the matrices have been exposed once. The average speed of production is on the order of 1,100 symbols per second, or almost 4,000,000 per hour.

Carrying the system of electronic composition to its logical conclusion, designers have replaced the matrices, whose outline had to be repeatedly electronically analyzed, by the results of analyses previously carried out and preserved in binary form in a magnetic rapid-access memory, setting up for each letter the output program for its luminous image on the cathode-ray screen when it is selected for composition. Electronic phototypesetters of this kind are called alphanumeric.

Hell-Digiset carries out a preliminary analysis by inscribing the outline of each letter on a very dense grid of 3,000 to 6,000 small squares, according to the body size of letter envisaged. Those squares covered by the outline are assigned the symbol 1 of the binary code; the others are assigned the symbol 0. The result of the analysis is first inscribed in perforations on an eight-channel tape. Tape containing perforations for an entire set of type in a given style is inserted into a special Digiset reader to instruct the magnetic memory, in a few dozen seconds, concerning type production. All that is necessary to change the style of type is insertion of the tape belonging to another set.

The Digiset 50 T 2 can reach a production speed of 3,000 characters per second, or more than 10,000,000 per hour. One Digiset is designed to permit a whole newspaper page to be composed photographically in a single scanning operation; not only the words but also the illustrations are analyzed in binary code.

Fototronic-CRT and APS (Alphanumeric photocomposition system) reduce the amount of coded information by interpreting each letter as a series of closely packed adjacent vertical lines whose distinguishing parameters are their height and their position. Vertical scanning on the screen of the photographic output device reproduces these lines one after another according to these parameters (see Figure 6).

The number of lines varies from about 50 to 90, depending on the width of the letters, and the number of units calculating the measurement of parameters of height can go up to 80, which amounts to a definition perceptibly as fine as the Digiset grid, or 800 lines per inch in two dimensions on the screen of the output mechanism.

The APS electronic phototypesetter has a production speed of 3,000 to 10,000 characters per second, the latter figure amounting to 36,000,000 per hour.

Makeup of letterpress copy. Preparing a form suitable for use in printing from letterpress copy, whether in individual type pieces or in lines of lead alloy, is an operation called makeup. This is preceded, if the same form is to include several smaller pages to be printed together, such as a book, by an operation called imposition, which consists in laying out the pages in the form so that they are in their numerical order after the printed sheet has been folded into a signature of eight, 16, or 32 pages.

Newspaper makeup

In the case of a daily newspaper, one form is generally used for each page. When the manuscript or typescript copy containing information regarding the type and the justification required arrives at the printing plant, it is divided among several Linotype machines. The titles are handed over to be composed, depending on their body size, in movable type on a Ludlow or Linotype machine. When ready and corrected from galley proofs, the titles and columns of text arranged in galleys are brought to the compositor, together with, where relevant, plates of illustrations or advertising matter mounted on lead blocks the same height as the type. Standing at a level casting table, the stone, and following the instructions on a layout, the compositor arranges the elements inside a rectangular steel frame, the chase, usually equipped with a locking system using quoins that slide along the inside of two of its adjacent sides.

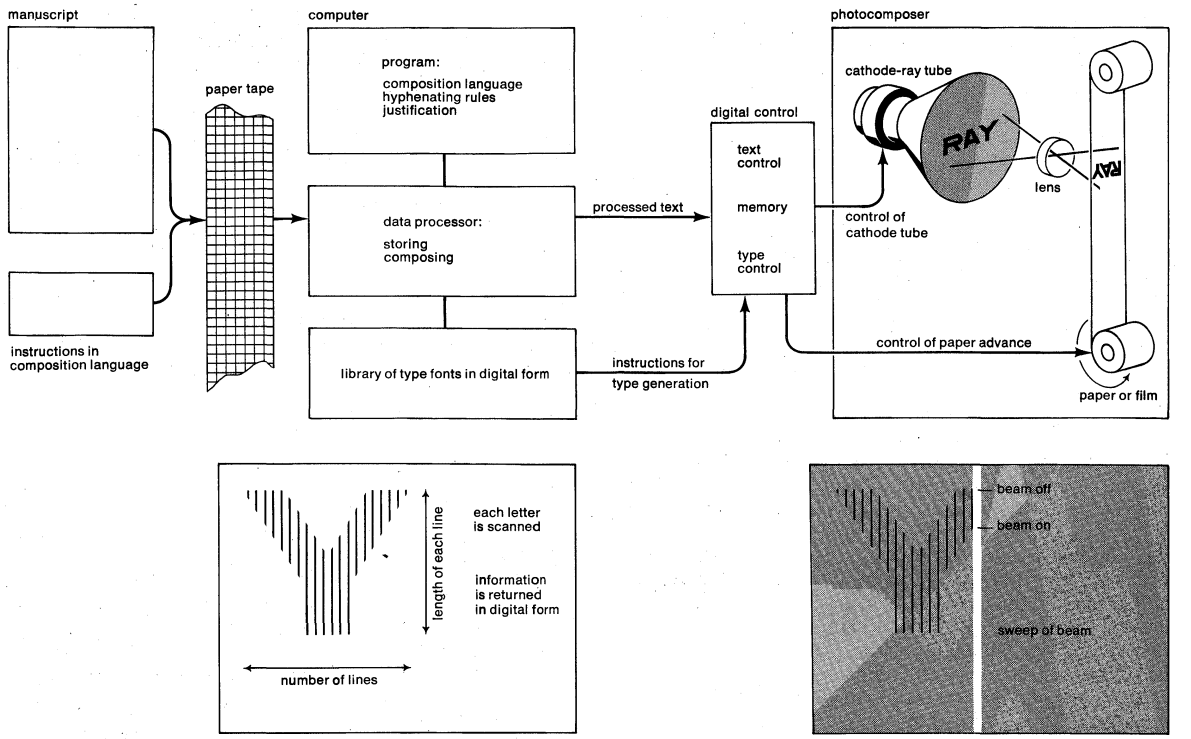


Figure 6: Basic components of the Alphanumeric photocomposition system (see text).
By courtesy of Alphanumeric Inc.

The compositor inserts leading between paragraphs to bring each column on the page to its proper height. He separates columns and articles using leading or rules cut to the desired dimensions. Once the chase has been locked, proofs are taken to check the page for final corrections before it is pressed on to a metal frame to be brought to the press.

Mounting composition on film. Composition on film consists of a mounting operation carried out on a luminous table. Films of the text and titles and positives or negatives of the screened or unscreened photographs, depending on the printing method to be used, are arranged on a sheet of transparent plastic of the dimensions indicated on the layout, placed under the plastic sheet and lighted from beneath. The film is glued or fixed using pieces of transparent adhesive.

Conversion systems. One composition process can be converted to the other. A page composed on film, in negative, can be used for photogravure plates or for metal or engraved plates intended for letterpress printing (see below *Printing*). Contrariwise, a page composed in type can be converted to a positive or negative, direct or inverted transparency by any of a number of techniques.

The whole page, including the screened plates or photographic illustrations, can be subjected to conversion, or the text can be taken alone, leaving the positioning of illustrations until the positives or negatives of the illustrations, screened or not, according to the printing process to be used, are included in the makeup at a later stage.

PRINTING (PRESS OPERATION)

Printing in the second sense of the word—that is, press operation—is the technique by which ink or other colouring agent, transferred to paper or any other material, is localized on printing surfaces delimited by the composition of texts or the making of illustration material.

Colour printing. Juxtaposition of colours is achieved by submitting each sheet to successive impressions by typeforms each of which prints only on areas designed to carry a single colour and inked only in that colour.

There are sufficient wavelengths in the three primary colours, blue, red, and green, to reconstitute all the colours of the spectrum. The colour perceived from an inked impression results from the fact that the ink reflects some waves of the range of colours of the spectrum and absorbs the others; that is, prevents them from being seen. Thus, three colours of ink can reconstitute the visual effect of all the range of colours by combining them appropriately. Yellow totally absorbs all waves of colour that are basically blue and reflects waves of colour that are basically red and green; magenta (deep crimson) totally absorbs green and reflects red and blue; cyan (close to turquoise) absorbs red and reflects blue and green.

Effects of combining coloured inks

When two of these inks combine, each annuls in the other the capacity to reflect that one of the two primary colours that it cannot itself reflect; the eye then perceives only the one primary colour that both inks reflect; for example, red, by combining yellow and magenta inks. All three inks combined no longer reflect any of the three primary colours but appear as black.

In trichromatic printing the screened plates necessary for each of the three colours of ink are prepared by selecting the colours through filters. A fourth plate is usually used for making a print in black ink to accentuate the contours and modelling in the picture, making the process quadrichromatic.

Printing colours by superposition requires the exact positioning on top of one another of the successively printed constituent parts of the picture (usually printed in this order: magenta, yellow, cyan, black). The finer the definition of the screen, the more precise this positioning must be.

Letterpress printing. Letterpress printing consists of transferring a thin film of ink from the printing surface of the typeform to that of the paper by pressing the two together.

Letterpress presses are thus made up of two principal elements, one bearing the type form, the other exerting the pressure. These elements may be either flat or cylindrical. There are three principal types of presses, according to the way these elements are combined: (1) plane to plane; (2) cylinder to plane; and (3) cylinder to cylinder (see Figure 7). In all types of letterpress presses the printing surface must be coated with a uniform layer of ink before receiving the paper. This inking is carried out by a mechanism composed of up to 20 rollers: take-up rollers carry the paste ink from the ink supply to the distributing and sliding rollers, which operate with a longitudinal to-and-fro movement to crush and spread the ink in an even

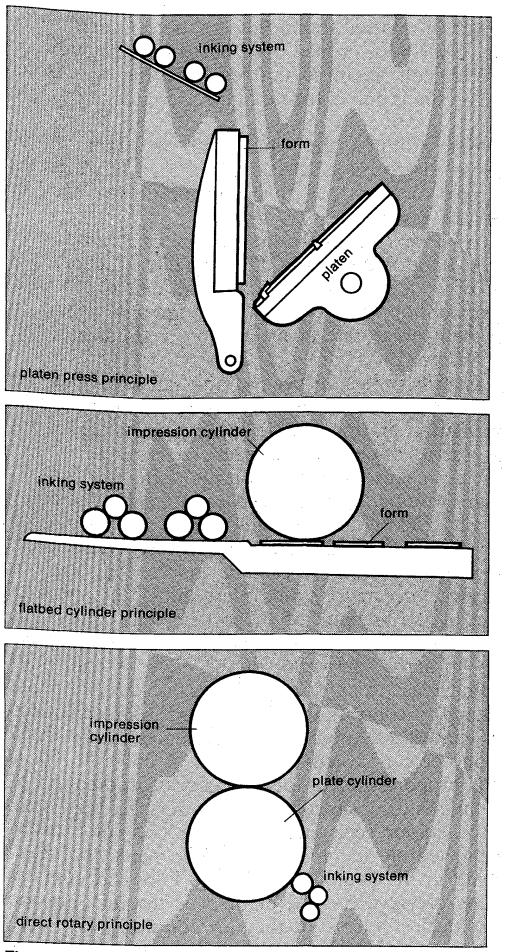


Figure 7: Operating principles of the three principal types of letterpress presses (see text).
From J.S. Mertle and G.L. Monsen, *Photomechanics and Printing*

layer on a level or cylindrical metal surface and finally contact rollers that transfer the ink to the printing surface of the typeform.

The first two kinds of presses are sheet fed. The third can be either sheet fed or roll fed (web fed), depending on the model and type of work. Sheets of paper must be fed into the press in a sequence synchronized with the movement of the press. Most modern presses are equipped with an automatic feeder that functions in conjunction with the movement of the press. The feeders operate by either friction or suction. In a friction-operated feeder, the sheets of a pile of stock on a slightly slanting surface are fanned out so that each sheet projects over the one beneath in such a way that the friction of a cylinder can dispatch them individually one after another toward the feedboard of the press, where three pegs guide each into position. In a suction operation the sheets remain piled up vertically; a wheel brushes a corner of the top sheet and separates it from the others, enabling a compressed-air blower to inject a cushion of air underneath the sheet. A system of vents connected to a suction pipe lifts the paper and carries it towards the press's feedboard. An automatic device lifts the pile of sheets continuously into place. On all letterpress presses, slight irregularities of the surface of the typeform are compensated for by packing the platen with a material soft enough to absorb the irregularities.

As the sheets leave the press, a powder is sprayed onto their surface to form a separative coating that prevents ink from one sheet being transferred to another. An alternative in modern high-speed presses is the incorporation in the ink of a special quick-drying agent.

Platen presses. Presses that operate plane to plane are called platen presses. A vertical clamping contrivance clamps the bed, which carries the form into which the composed type is locked, and the platen, which carries the sheet of paper while it is being printed. When this clamping contrivance is open, the typeform is inked by a series of rollers that descend and then reascend, and the printed sheet is removed and a new sheet placed in position on the platen.

The pressure exerted during impression is about 40 kilograms per square centimetre (570 pounds per square inch). A platen press can reach a production speed of 5,000 sheets per hour.

Cylinder presses. In presses that operate cylinder to plane, called flatbed presses, a cylinder provides the pressure while the typeform retains its flat surface, generally in a horizontal position. Generally, too, the bed is mobile to allow the typeform, as it moves back and forth, both to pass under the rollers of the inking system and to pass under the impression cylinder around whose outer surface the sheet of paper is wrapped, attached by a set of clamps. Flatbed presses fall into various categories, depending on the cylinder's operation.

The stop-cylinder press

In the stop-cylinder press, a toothed rack incorporated in the bed engages a cogwheel incorporated in the cylinder while the bed is moving forward. As it moves back again, the cogs disengage. A shallow cavity in the cylinder makes it possible for the typeform to be slid underneath. Printing speeds can reach 5,000 sheets per hour.

In the two-revolution press, the cylinder never stops revolving but is raised on its bearings as the bed moves back again in order not to touch the form. In its lowered position the cylinder's cogwheel engages a low-toothed rack incorporated in the bed; in its raised position the cogwheel engages a parallel high-toothed rack, which enables the cylinder to continue revolving in the same direction. Printing takes place during the first revolution; during the second revolution the cylinder runs free. Printing speeds are about the same as for the stop-cylinder press. But, by avoiding the mechanical jerkiness due to the stopping of the cylinder, the two-revolution press has a much smoother, more regular, and quieter action.

On a single revolution press, the cylinder does not stop revolving but must be raised while the bed moves back. Its diameter is twice that of the two-revolution cylinder press, but one half of its surface is hollowed out in order not to touch the form. Printing, then, takes place in the first half of the cylinder's revolution.

A perfecting press is a combination of two two-revolution presses bracketed together. It has two cylinders and a single bed bearing two different type forms. Thus, as it moves back and forth the bed prints two impressions, one for each form. The same sheet of paper is passed from one cylinder to the other and is thus printed on both sides. The padding on the second cylinder is constantly cleaned by a kerosene-coated roller to prevent a transfer from the first side of the paper to be printed. In fact, the second impression is better than the first, and the part of the job requiring a better quality is reserved for the second cylinder.

The two-colour press also combines two bracketed two-revolution presses, but a small auxiliary drum is interposed between the two cylinders to ensure that a single sheet of paper presents the same side to be printed twice. The typeforms, which are designed to complement one another, are each inked with a different colour.

Contrary to the general design of cylinder presses, the vertical cylinder press is composed of a vertical bed, and both bed and cylinder move vertically with a reciprocating motion, each in an opposite direction. The cylinder revolves only while it is moving up and down, which makes this kind of press similar to the stop-cylinder press.

Printing speeds exceed 5,000 sheets an hour for paper up to about 2,000 square centimetres (300 square inches).

Rotary presses. Presses that operate cylinder to cylinder, usually described as rotary presses, consist of two cylinders turning in opposite directions. The typeform is attached to the surface of one cylinder, and the impression cylinder provides the pressure.

Sheet-fed rotary presses produce the same kind of work as flatbed cylinders, but the size of paper can be slightly larger, and production speed is three times higher using the same size of paper. The inking system is much the

same as in flatbed cylinder presses. In both models the sheet of paper is held to the impression cylinder by a set of clamps. On the largest models passage of the paper between the cylinders is controlled only by precise positioning and presentation of the paper.

The two-colour rotary press combines two plate cylinders, each bearing a different typeform and each provided with its own inking system against a single impression system, in a so-called satellite arrangement. Thus, the same side of the same sheet of paper receives two successive impressions of two different colours in a single revolution of the impression cylinder.

The rotary perfecting press uses the same arrangement as the previous machine, but a second impression cylinder of smaller dimensions is interposed between the first impression cylinder and one of the plate cylinders. Thus, the sheet changes sides between one impression and the other and so is printed on both sides.

The two-colour cylinder and flatbed press combines a rotary and a single-revolution press by using the same impression cylinder. Clamped around this cylinder, the sheet of paper first moves against the curved form on a plate cylinder inked by the first inking system; then it moves against a flat form on a mobile bed inked in another colour by a second inking system.

Polychrome rotaries permit three, four, or five colours to be printed on the same sheet without handling the pile of paper. Some of these are designed on the planetary principle, using as many plate cylinders as colours, each being fed by its own inking system and grouped around a single impression cylinder. This type of machine is much more popular in North America than in Europe. Others are designed as a row of identical units each printing one colour, the paper moving from one unit to another by means of a transmission drum or a conveyor contrivance.

The planetary colour press

Roll-fed rotaries are exceptionally large, with high rates of production, and are used almost exclusively for printing daily newspapers. Their principle is extremely simple; the continuous roll of paper drawn from a reel moves between a plate cylinder and an impresssion cylinder. In practice, the cylinders of some have a circumference twice the height of a page of newsprint. Thus, each revolution of the cylinder prints two copies of the same page. On others the circumference of the cylinders corresponds to the width of four pages of newsprint side by side; each revolution prints eight copies of the same page.

In another type, the basic unit of the rotary, called the group, is composed of a symmetrical arrangement of two plate cylinders with their impression cylinders. Within the same group the roll of paper moves from one of the plate cylinders, where it is printed on one side, to the other cylinder, where it is printed on the other side. Each revolution of the machine thus produces a group of two times eight pages of newspaper printed on both sides. Inking is effected by distributing, sliding, and contact rollers that receive ink from dozens of openings distributed across the width of the cylinder, each of which can be precisely adjusted. In each group paper is fed in by a barrel-like device with three axes for supporting the reels, which makes it possible to move over from a finished roll to a new one by a simple gluing operation and a 120° revolution. The reels weigh up to 600 kilograms (about 1,300 pounds). The rotary press is made up of a certain number of identical groups in alignment.

Folding and cutting operations

After printing, the roll of paper is automatically folded down the middle so that on both sides only two pages remain face to face. Moving along a triangle with rounded sides and then passing between rollers, each half roll is folded in the middle to form a newspaper. A cutting mechanism, synchronized with the action of the rotary, separates each newspaper from its neighbour.

Depending on the number of pages of the newspaper, various arrangements are possible: rolls coming from different groups can be added to by accumulation to produce two issues in a multiple of four pages. A roll half as wide as the others coming from one of the groups can be added to the middle of one or more of the usual rolls, producing two issues in multiples of four pages, plus two. Two parallel rollers, called turning bars, arranged at an angle of 45° in relation to the flow of the roll, enable the roll, after it has been printed and folded in the middle, to be folded in half again so that each of the two issues produced by the group is formed into a signature of eight pages.

For colour printing, the same roll of paper moves through several groups in succession, in each of which the plate cylinders carry typeforms appropriate for each of the colours necessary.

Modern rotaries can revolve at the rate of 35,000 revolutions per hour (500 metres [1,600 feet] of paper per minute), or at a theoretical production rate of 140,000 newspapers an hour in two issues from the final group. In practice, the average rate of production is about half this figure.

At speeds such as this, inspection and safety precautions are more reliably carried out by electromagnetic devices based primarily on the use of photoelectric cells; for example, a series of cells on the track, over which a roll of paper moves, react to tearing by stopping the machine.

In colour printing, photoelectric cells ensure matching by selectively picking up the guide marks printed in each colour as they go by and by reacting to any irregularities in the distance between these guide marks. Any error is automatically corrected by modifying either the speed of one group or the pressure of the rollers that control the tension of the paper between one group and the next. Matching up across the width of the paper is controlled by cells that react to any lateral digressions as the paper moves along by controlling a lateral-shifting mechanism.

Quality control of colour reproduction is also carried out by photoelectric cells that emit a current whose strength is proportional to the intensity of impression of the guide marks for each colour. By comparing this intensity with a colour scale, an electronic computer determines from minute to minute the adjustments that need to be made in the composition of the inks and controls the opening of valves to add pigment if the inks are too light or colourless varnish if they are too dark.

Preparing stereotypes and plates. The curved shapes with which the printing cylinders of the rotaries are fitted are the stereotypes, or plates. These may be produced from ordinary flat typeforms by reproducing the relief surface of the type copy and of the plates for illustrations made from halftone photoengravings and line engravings. Alternatively, screened photographic illustrations are used to make plates from mounted positives of transparencies by means of photoengraving.

Casting stereotype plates

In making stereotype plates, a flong, or mat, a thin sheet of pasteboard, pliant enough to register an impression and sufficiently heat-resistant to tolerate the molten type metal, is placed on the type form with paper and cotton packing. It is subjected to heavy pressures in a press at a moderately high temperature to ensure that it dries; it retains an intaglio impression of the relief surface of the type form and is placed against the inside wall of a curved casting box into which a lead alloy is injected. The cooled stereotype plate resembles a rigid shell, solid or ribbed according to its thickness.

The plate is mechanically finished to ensure uniform thickness and to provide properly bevelled edges; metal from nonprinting areas is routed out to avoid ink smudges. Finally, the plate is electroplated with a thin layer of nickel to increase resistance to wear.

Though it is rarely done, stereotype plates can equally well be cast flat, then curved while heated after being finished.

The stereotype process is the fastest and most economical process for obtaining curved plates, but such plates are not suited to the precise matching up necessary in colour printing because of the irregular behaviour of the mat; its performance depends heavily on humidity and temperature conditions.

In making electrotype plates, an impression is made of the typeform using a substance that is a conductor of electricity or can be made conductive by treating with black lead or dusting with a powdered metal such as silver. Any of the following may be used: a sheet of wax submitted to heavy pressures in the press; a sheet of lead

submitted to extra heavy pressure; a sheet of Tenaplate, a kind of vulcanized plastic covered with a film of black-lead wax, submitted to pressure only slightly less than that for lead; a sheet of celluloid or a sheet of plastic (Vinylite, tenalite composed of a layer of aluminum between two layers of Vinylite, etc.).

After metallizing to ensure conductivity, these molds are electroplated with a thin copper shell, delicately reproducing the relief surface of the type form. This shell is stripped from the mold and reinforced with a backing of lead alloy poured over its underside. Nickel plating can again be employed to increase its wear resistance.

Electrotype plates can be curved while cold after they have been backed or while hot, during leading, when the lead has not yet solidified. There are also impressions that can be curved before electroplating to obtain curved copper shells. The plate is then reinforced by spraying molten metal against the shell while it is fixed against the inside wall of a rapidly spinning drum.

Making electrotype plates, especially in a curved shape, is a costly process, but they produce the best quality print. An impression made with a sheet of lead is best suited to printing in colour because it is least sensitive to variations in humidity and temperature. Metal-shell plates are attached to the plate cylinder mechanically.

Stereoplastic plates are made in two successive moldings. A hot mold is made in the press from the typeform in a thermosetting material; that is, one that can be melted only once and can thereafter tolerate high temperatures without damage, such as Bakelite. This first molding itself serves as the mold into which is pressed the hot material for the plate—usually cellulose acetate, vinyl resin in sheets or in powder to which a plasticizer has been added to make it more durable, or a sheet of rubber gum that will be vulcanized when pressed. New liquid plastics can also be employed using a method similar to casting.

The plates are trued up by milling or filing to the desired thickness and glued either to the plate cylinder of the rotary or to a metal plate wrapped around it.

Stereoplastic plates are light and easy to use, particularly on small rotaries. Their printing quality is good enough for texts and line illustrations, although they are ill suited for fine-screened halftone illustrations.

Whether wraparound plates (wraproud plates in Great Britain) are made of metal or of plastic, they utilize photosensitive materials in their preparation.

Metal wraparound plates may be of copper, magnesium, or zinc. Only one type of zinc, microzinc, whose molecular structure permits finer prints than ordinary zinc, is used. The plates are covered with a layer of photosensitive material and processed like photoengravings, using negatives of the pages on which the photographic illustrations have already been screened. Because the engraving is only half as deep as letterpress engraving, ink rollers of a larger diameter are used.

The plate can be curved after engraving, but curving is usually done beforehand, both to avoid breaks in the metal and to obtain an absolutely uniform degree of bend in the various plates for colour printing.

Plastic wraparound plates

Plastic wraparound plates utilize the property of photosensitive polymers to lose their solubility in certain solvents when exposed to light. Exposure to light through the negative of a page fixes the insolubility of the polymer and limits it to the areas that are to constitute the printing surfaces. A suitable solvent then eliminates all the nonprinting areas and sets off the type in relief.

New polymers with this property and with new qualities are constantly being perfected. Among the better known examples are nylon, Dycril, and KRP. Nylon is sensitized in bulk by immersion in a solution of acetone containing the sensitizing agent. The plate is exposed to ultraviolet light, and the nonprinting areas are dissolved by a bath of methyl and ethyl alcohol. It takes 24 hours for the plate to attain its maximum hardness. Dycril is sensitized by immersing it for 24 hours in a carbon dioxide atmosphere. The nonprinting areas are removed by sprinkling with a solution of sodium hydroxide. Because it is preferable for the plate to be curved before being engraved, exposure takes place on a rotary drum turning in front of an arc lamp and the bath on a rotary drum turning in a trough. The total amount of time required to make a plate is about 45 minutes.

KRP (Kodak Relief Plate) is a sheet of cellulose acetate that is superficially sensitized by the deposit of a thin coat of photographic emulsion. After exposure to light, this emulsion remains only on the printing areas, which it protects from the action of the solvent. Engraving the KRP plate can also take place on a rotary drum.

The polymer of the plastic wraparound plates is usually mounted on a base consisting of a metal sheet. The depth of the engraving can equal the actual thickness of the polymer. The type therefore stands out in sharp relief. Whether metal or plastic, the wraparound plates are easily and rapidly attached by means of register hooks, which ensure perfect tension of the plate on the surface of the rotary's plate cylinder.

Scope of letterpress. Letterpress printing is characterized by the sharpness and strength of bite of the type. It produces good reproduction of illustrations on sheet-fed machines, with these two reservations, that screening prevents the reproduction of pure white, and it is not possible to use more than four colours without risking a speckled moiré pattern. Roll-fed printing still preserves a good sharpness in the text but produces only an acceptable or average photographic illustration, depending on the quality of paper, and then only in black. Colour printing of illustrations on roll-fed rotaries can hardly rise above the mediocre level, even with the use of special rotaries in which the printing groups are drawn as close together as possible.

Rotogravure. Rotogravure is a system of intaglio printing. It consists of transferring to paper fluid ink contained in the cells of the printing cylinder, while the projecting nonprinting areas on the surface of this cylinder are kept free of ink by constant wiping.

The density of the print at each point depends on the depth of the cell at that point and the quantity of ink it contains, rather than on the printing surface, as in the letterpress process. The screen no longer plays an optical role. It is used to establish the partitions that separate all the cells of the honeycomb from each other and that form a surface of uniform height, while the cells are all of different depths, so that the ink is taken up on the engraved surface in an exactly defined quantity. The screen also prevents the wiping mechanism from penetrating the cells of the cylinder and withdrawing the ink. For this reason, line drawings and even the text type must be screened, as well as the photographic illustrations.

Rotogravure rotaries are thus composed of two principal elements: the cylinder bearing the printing form and the impression cylinder; the paper moves between these two cylinders either in sheets or from a roll.

The nature of the ink used in rotogravure and the operation of the wiping mechanism make the use of plates technically more delicate because of the indentation that would be necessary in the places where they were affixed to the cylinder. Plates are used only on certain sheet-fed machines, because they are easier to prepare and stock. Generally speaking, the image is etched directly onto the printing cylinder, which must therefore be easily handled.

Ink of sufficient fluidity is used to allow direct inking of the engraved cylinder without the use of a roller to spread and distribute it. While the paper tangential to its upper surface is being printed, the lower part of the cylinder is bathing in the pool of ink. In fast-working machines the ink can be poured on through a spout or sprayed onto the surface of the cylinder, to avoid flying ink. On the plate machines, inking must be done by roller to avoid filling the hollows of the clamping system with ink.

The operation of the doctor blade

Between inking and printing, the wiping mechanism comes into action. It consists of a thin blade of soft steel, the scraper, or doctor blade, which moves slowly to and fro lengthwise. By rubbing against the cylinder with a precisely regulated degree of pressure, it causes the excess ink to drop off before the cylinder moves over the paper.

In sheet-fed machines the impression cylinder is like that of a letterpress press, with a diameter large enough

to carry the sheet wrapped around its surface, gripped and released by a system of articulated clamps. Production speed can attain 6,000 sheets per hour.

Roll-fed rotaries generally consist of several printing units in line (up to as many as 18), each composed of a printing cylinder, an inking system, and a hard-rubber impression cylinder of small diameter. Each unit can turn in either direction to allow the paper to move through in any combination.

When printing both sides, the same roll of paper must pass through two successive printing units. In colour printing the same roll must move through as many printing units as there are colours used on each side of the paper. Several rolls coming from different printing units can be added together by accumulation.

Rotogravure rotaries normally include the same folding and cutting systems as letterpress rotaries, along with the electronic devices necessary for controlling the operation. Because of the fluidity of the ink, all rotogravure machines, whether sheet-fed or roll-fed, require a drying process. The paper passes over a heated drum or over sources of infrared rays or through a compartment ventilated with hot or cold air.

Preparing rotogravure cylinders. The printing cylinders or plates are prepared from positives of pages on which neither text nor illustrations are screened. The photosensitive substance necessary for carrying out the photoengraving operation, carbon tissue, is separately packaged to permit it to be treated before being affixed to the metal. Carbon tissue is paper, in sheets or on rolls, coated with a layer of gelatin. This layer is sensitized before use by being plunged into a solution of potassium bichromate. The carbon tissue is first exposed to intense light through a glass plate bearing a transparent screen on an opaque background and then exposed a second time through a positive of the pages. During these two exposures the gelatin is hardened by the light, completely in the white and screened areas, more or less deeply in the halftones of the photographic illustrations, and not at all in the areas covered by text and lines.

The carbon tissue is then applied to the copper surface of the cylinder (or plate). The tissue is then peeled off, leaving the gelatin film fused with the metal. The cylinder (or plate) is plunged into a warm-water bath, which dissolves the gelatin in inverse proportion to the degree to which it was hardened: totally in areas where it received no light like the text and line drawings, more or less deeply in the halftone photographic illustrations, and not at all in the screened areas.

The etching of the copper is begun by sprinkling it with an acid solution (ferric chloride), which attacks the metal in the areas where the gelatin is totally dissolved and bites into it more or less deeply depending on the thickness of the surviving gelatin that it must penetrate. After being etched, the printing surface can be reinforced by chromium plating.

Improvements on this classic method of preparing the rotogravure plate are continually being made. The carbon tissue may be replaced with silver emulsions on a plastic base; a direct copy without carbon tissue may be made by dusting the cylinder with a photosensitive substance and projecting the image on the surface of the cylinder by optical means and finally by electronic engraving.

Plates for rotogravure are made of solid copper, and after use they are discarded unless preserved for reprinting. Rotogravure cylinders consist of a steel mandrel on which a layer of copper has been deposited by electroplating. After printing, the etching is removed by grinding and a thin film of copper deposited to restore the cylinder to its original diameter. Restoration is simplified if the film can be prevented from adhering to the cylinder so that it can be ripped off after printing. This can be done by coating the surface of the cylinder with a copper–mercury amalgam before electroplating. After plating, the new copper film is then polished. Some electroplating baths produce a shiny copper finish, dispensing with the polishing operation.

Scope of rotogravure. Even in long runs, rotogravure produces quality illustrations with rich colours. It is less suited to printing small typefaces because they are cut up by the screen.

Offset. The first characteristic of offset printing is that, in principle, there is no appreciable physical difference between the printing and the nonprinting surfaces. The difference is simply that the printing surfaces repel water when moistened but absorb the ink with which they are coated, whereas the nonprinting surfaces absorb the water and repel the ink. The form used in offset printing is a metal plate on which the surfaces have been prepared to give them these opposite properties.

The second characteristic of offset printing is that the ink is not transferred directly from the printing form to the paper but is first transferred to a rubber blanket. It is the ink retained by the blanket that is transferred to the paper.

The three-cylinder offset press

Offset presses utilize three principal elements: the plate cylinder; the blanket cylinder; and the impression cylinder, which applies the paper, in sheets or a roll, against the blanket cylinder.

The plate cylinder is like the cylinder for wraparound plates on a letterpress rotary, with a groove in which the attaching and tension mechanisms are housed. A moistening system and an inking system are connected to this cylinder. The inking system is similar to that on letterpress presses, with a series of alternating hard and soft rollers that grind, spread, and evenly distribute the ink in a uniform thickness.

The wetting system, which dampens the plate before it is inked, consists of a series of rollers that transmit regularly controlled quantities of water taken up by puddling in a tank or sprayed from rotary brushes grazing its surface.

The blanket cylinder also has a groove containing the mechanism for attaching and controlling the tension of the blanket, which consists of several layers of fabric and rubber. During printing the groove of the blanket cylinder and that of the plate cylinder coincide at each revolution.

On sheet-fed machines the impression cylinder has a cavity in which a system of articulated grippers is recessed. The movement of the machine is synchronized so that at each revolution the slightly projecting grippers enter into the groove in the blanket cylinder in such a way as to cause no damage to the latter. This problem does not arise on the impression cylinders of roll-fed rotaries, and their movement does not therefore have to be synchronized with that of the other cylinders.

In sheet-fed presses the three cylinders, usually placed side by side, are of the same diameter and turn at the same speed. The movement of the cylinders is synchronized by meshing sprocket wheels. Paper is fed in by a feeder similar to that of letterpress presses.

Offset machines can be used to print very large sheets. The fastest reach a production speed of 10,000 sheets per hour.

In printing two or more colours, during the offset process the dampness of the plate is partly transferred, via the blanket, to the paper. This slightly alters the paper's dimensions, causing difficulties in register. The separate colours are therefore printed as simultaneously as possible.

On some two-colour machines a single impression cylinder applies each sheet successively against two blanket cylinders, which receive a transfer from their own plate cylinders.

More generally, sheet-fed offset machines printing in two or more colours are designed as a series of printing units each consisting of three cylinders with their wetting and inking systems, the sheet being transferred from one unit to the next on drums of large diameter with a system of articulated grippers. Size of paper and speed are the same as for the one-colour machines.

Blanket-to-blanket machines

In the offset process an unusual arrangement, without impression cylinders, is possible. Simultaneous printing is done of both sides of the sheet as it passes between two juxtaposed blanket cylinders. Each receives a transfer from its own plate cylinder and prints one side of the sheet while at the same time acting as an impression cylinder for printing the other side of the sheet. Thus, the

blanket-to-blanket machine consists of a total of only four cylinders.

The various kinds of machines can be used in combination: for example, three one-colour printing units and one blanket-to-blanket machine can print one side of the sheet in four colours, the other side in black.

Roll-fed rotaries work on the same principle as sheet-fed presses, with two differences—the groove for the plate fasteners on the plate cylinder is very narrow, and the impression cylinder, lacking grippers, has an entirely level surface.

In-line rotaries are designed as a succession of one-colour printing units or, more frequently, blanket-to-blanket units. On these, as on the others, printing in several colours is done by sending the paper directly and horizontally through several successive units.

On blanket-to-blanket rotaries, the back is printed simultaneously with the front. On rotaries made up of one-colour units, a turning bar reverses the direction in which the roll of paper presents itself. Feeding the paper into and taking delivery of it from the machine can be combined so that the rolls are superimposed to form a signature.

Satellite rotaries (drum presses) work on the principle of a single impression cylinder of very large diameter that has arranged around its outer surface four, five, or sometimes six blanket cylinders. Each receives a transfer from its own individually moistened and inked plate cylinder. Thus, colour printing one side of a roll of paper is accomplished in a single revolution of the impression cylinder.

Drum cylinders are most often made up of two symmetrical groups printing one side after the other. For quality colour printing with a heavy thickness of ink, rapid drying is necessary to prevent smudging. The roll therefore moves horizontally into a drier heated by a ramp of gas burners or hot-air blowers and is cooled as it passes over metal drums refrigerated by circulating water.

Offset rotaries work at a speed of 15,000 to 20,000 revolutions per hour and have the same cutting and folding systems as letterpress rotaries.

Preparing the plates used in offset involves defining and fixing the joint presence of two mutually repellant materials, one water-receptive for the nonprinting areas, and the other ink-receptive for the printing areas. As in letterpress the screen is necessary to translate the halftones of photographic illustrations into surface densities.

Even though there is no question of creating reliefs, the preparation process is very similar to that of photoengraving. But, because the blanket intervenes between plate and paper, text and illustrations appear on the same offset plate in straight rather than in reverse reading as on letterpress plates.

Offset plate making. Depending on the work they are intended for, various kinds of plates and processes for preparing them exist.

Monometal plates are made of a metal with hydrophilic (water-receptive) properties, such as zinc or aluminum, the surface of which is treated to make it more porous. The plate is coated with a thin layer of a photosensitive substance, covered with a negative of the texts and previously screened photographic illustrations, and exposed to intense light. This hardens the photosensitive substance in the areas of the negative where the light passes through; that is, in those corresponding to the printing areas. The photosensitive substance is then washed away from the nonprinting areas, where the metal, now stripped, is wetted; the hardened photosensitive substance is inked.

Presensitized plates are monometal plates coated with a photosensitive layer that has a life-span, away from light, of up to six months. Certain presensitized plates for use in short runs can be made of paper or plastic.

Deep-etch plates are monometal plates prepared by inversion. The plate is coated with a photosensitive substance and then exposed to light through a positive of the text and screened photographic illustrations. This exposure hardens the photosensitive substance on the nonprinting areas. Washing eliminates the photosensitive substance on the printing areas, exposing the metal. The plate is then given a mild acid bath, which etches the metal of the printing areas to a shallow depth. An ink-receptive lacquer is spread over the surface of the plate; that is, over the hardened photosensitive substance wherever it still survives, as well as over the etched-metal areas. The hardened photosensitive substance and the lacquer are both removed at the same time by dissolving and brushing, exposing the metal of the nonprinting surfaces to be wetted. Lacquer remains in the cavities of the etched areas, ready to be inked. Plates treated by this process are suitable for longer runs (up to 250,000 copies).

Bimetal or trimetal plates are composed of two superimposed metals, one of which is hydrophilic (aluminum, stainless steel, chromium, nickel) and the other ink-receptive (copper, bronze). Whichever of the two covers the other in a microscopic film is partly eliminated by the photoengraving operation. In the case of a hydrophilic metal, photoengraving is carried out using positives of the text and illustrations. In the opposite case, negatives are used.

Among the combinations used in various commercial plates are chromium on copper or nickel on bronze for use with positives and copper on stainless steel or copper on aluminum for use with negatives.

The combination of two metals can, in certain plates, take the form of a double film deposited on a third metal (steel or zinc), which simply serves as a base. Bimetal or trimetal plates are very durable and can be used to produce an average of 500,000 copies.

Alongside these various types of conventional offset plates, special processes are being developed that tend to further simplify the work of preparation. Electrostatic (or xerographic) transfer plates were developed from the principle that certain materials (such as selenium) that are insulators while in the dark become conductors of electricity when exposed to light. A selenium plate is given a positive electric charge in the dark and then exposed to light through the positive of the text and illustrations. In the areas touched by the light, the plate becomes a conductor, and the positive charge disperses. A fine, negatively charged powder is sprinkled over the plate and is attracted by the positive charge in those areas of the plate where it still remains. The image thus revealed is transferred to an aluminum plate, which is positioned on the selenium plate and then given a positive electrical charge. The powder transferred to the aluminum plate is fixed by heating to constitute the ink-receptive printing surface.

Innovations in offset plates

This whole operation, which is carried out automatically in a compact apparatus, takes only three minutes. Xerographic plates are used only on small machines.

"Immediate" offset plates consist of a polymer layer that is sensitive not to light but to heat, which makes it become hydrophilic. Those areas not touched by heat retain the opposite property. The plate is then ready for printing without further treatment.

Scope of offset. In offset printing the letters print a little less sharply and strongly than in letterpress, except on special qualities of paper. The reproduction of photographic illustrations, in black and in colour, also depends on the quality of the paper and may rival rotogravure reproduction when printed on offset rotaries with driers.

Other printing processes. *Letterset.* This process, also called indirect letterpress, or dry offset, is a combination of letterpress and offset. Like offset, it has a blanket cylinder as the transfer element between typeform and paper, but, like letterpress, it uses a relief typeform. Thus, letterset presses use the same three cylinders as offset, but there is no dampening system.

Because of the use of a blanket, the text and illustrations appear on the typeform straight and not reversed. Consequently, the use of plates is excluded, since these would be made by making a mold of a conventional typeform. On the other hand, all types of metal or plastic wraparound plates, prepared using positive or negative transparencies, can be used. These plates are thinner than those used for letterpress, and the height of their relief is less. Rollers of a quite large diameter must be used for inking, and con-

tact between plate and blanket cylinder must be extremely light.

All the combinations that the blanket allows on offset machines are equally possible on letterset machines: blanket-to-blanket and drum press. Machines are manufactured that can be used interchangeably, both for offset and for letterset; in the latter case, the action of the dampening system is suspended. The size of paper and performance rates are the same in letterset as in offset.

Serigraphy (screen printing). Serigraphic printing consists of forcing an ink, by pressing with a squeegee, through the mesh of a netting screen stretched on a frame, onto the object to be printed. The nonprinting areas of the screen are protected by a cutout stencil or by blocking up the mesh.

Silk and other screens

The screen is usually made of fine but strong silk gauze (available in different mesh sizes) but can also be made of synthetic gauzes (nylon, tergal) or of wire gauze (phosphor bronze, stainless steel, nickel) or of combinations (nylon–copper, nylon–bronze).

Preparation can be carried out by hand. The design to be reproduced is transferred to the screen by drawing with a benzene-soluble ink, spreading glue over the whole surface of the screen, then dissolving the ink so that the glue remains only on the nonprinting areas. The nonprinting areas can also be covered with glued paper or film, cut out in the shape of the desired image and attached either by heat or by a solvent.

Direct or indirect photomechanical processes are being more and more widely used, however. In direct photomechanical processing the screen is covered with a photosensitive layer and then exposed under a positive, the photographic illustrations having been screened previously. In indirect photomechanical processing, the printing and the nonprinting areas are prepared by exposure under a positive (screened, if necessary) of a photosensitive film (carbon tissue, the same as that used in rotogravure, or presensitized film), which is then bonded to the screen.

Screen printing is still done largely by hand, the frame being lifted up after each operation, but semi-automatic and automatic machines driven mechanically or by compressed air can carry out the series of operations: positioning the object to be printed, moving the frame, inking, spreading the ink with the squeegee, delivering the printed sheet or object, and transferring it to a drying apparatus.

Serigraphic printing can be applied to a wide variety of surfaces—paper, cardboard, glass, wood, plastic, posters, bottles, electronic circuits, etc.—and to varied shapes: for printing cylindrical objects, the squeegee is stationary, while the screen revolves under it at the same time as the object to be printed.

Modern machines reach speeds of from 1,000 to 6,000 copies per hour.

Collotype. A photosensitive layer, spread on a plate of glass and exposed under a negative, hardens and loses its hydrophilic properties in the areas where it receives light and in proportion to the intensity of the light received through the halftones of the photographic documents. It is then able to retain both ink and water in all areas, the ink repelling the water in inverse proportion to the intensity of the exposure.

Collotype printing, which is based on this property of mutual repulsion, is thus related to lithography. But it is also related to rotogravure in the fact that the thickness of the film of ink is not uniform but in proportion to the shades of tone in the original image. Collotype, the only printing process that can reproduce photographic documents without a screen, is characterized by its fidelity of reproduction.

Collotype presses consist essentially of a bed bearing the glass plate prepared for printing, an impression cylinder carrying the sheet to be printed on its outer surface, and a system of ink rollers. The photosensitive preparation retains its own moisture, so no dampening system is necessary.

Speed of collotype printing

Printing speed is slow, rarely more than 200 copies per hour, and the useful life of the printing surface is very limited. from 2,000 to a maximum of 5,000 copies. Increasingly, the glass plate is being replaced by a cellophane film as the surface to which to apply the photosensitive layer. The prints can be cut out and, when glued onto wooden or metal blocks of type height, can be placed alongside a typeform for a limited number of copies to be printed from a flat surface.

Collotype is used for printing limited editions of works that require an excellent quality of photographic reproduction in one or more colours, such as reproductions of documents or pictures, posters, and transparent illustrations for advertising or various artistic uses.

Flexography. Based on letterpress-printing principles, flexographic presses are composed of the same basic elements as letterpress cylinder-to-cylinder presses; that is, the impression cylinder covered with a rubber packing and an inking system, which is simplified, owing to the fluidity of the ink used. Although several sheet-fed models exist, flexographic presses are usually roll-fed rotaries that can reach large dimensions and high printing speeds. They consist of a group of several identical units that print an equal number of colours (up to six).

Flexography provides poor- or average-quality economical printing, either of solid lines or with quite a coarse screen, on unfinished surfaces, wrapping paper, cardboard, or plastic film.

Electrostatic printing. Electrostatic printing is a process of printing without contact, without a typeform, and without ink. The paper is coated with a very thin layer of zinc oxide, which makes it an insulator while in the dark and a conductor of electricity when exposed to light.

This paper is given a negative electrical charge in the dark. It is then exposed by light projected through a positive film of the document to be reproduced. The zinc oxide layer becomes conductive wherever it is illuminated, and the negative charge is dissipated in those areas that correspond to the blank surfaces of the document. Finally, the paper goes through a bath containing pigmented particles that are attracted by what remains of the negative charge and are fixed by drying.

Electrostatic machines have been designed for printing geographical maps. These are composed of five successive units, each carrying out the same complete cycle of processing the paper to produce an edition in five colours at speeds of about 2,000 copies per hour.

Improvements in the bath of pigmented particles are making possible the application of the electrostatic process to printing small books.

Printing inks. Printing inks contain three components, the vehicle, the colouring ingredients, and the additives. The vehicle, responsible for transferring the colouring ingredients from the ink fountain to the typeform, can be either a vegetable base (linseed, rosin, or wood oils), which dries by penetration and oxidation and at the same time ensures fixation, or a solvent base derived from kerosene, in which case drying takes place by evaporation. The colouring ingredients come in several forms: pigments, which are fine, solid particles manufactured from chemicals, generally insoluble in water and only slightly soluble in solvents; agents made from chemicals but soluble both in water and in solvents; and lacquers, obtained by fixing a colouring agent on powdered aluminum. The additives stabilize the mixture and give the ink additional desirable characteristics.

The nature and proportions of the ingredients vary according to the printing process to be used and to the material to be printed. The proportions must be checked and sometimes modified during printing.

Letterpress and offset use greasy inks. For printing on sheet-fed presses, thick greasy inks are used in which the vehicle is generally made of vegetable oils with the addition of hard natural or synthetic resins dispersed in mineral oils. Roll-fed rotaries use fluid greasy inks in which the vehicle is made up of heavy mineral oils.

Sources of ink pigments

The colour black is generally obtained from an organic pigment, carbon black, derived from the incomplete combustion of oils or of natural gas. Colour pigments are made from chrome (yellow, green, and orange), molybdenum (orange), cadmium (red and yellow), and iron (blue).

Inks for offset are more highly coloured than those used in letterpress, because they must be transferred to the blanket before they reach the paper. Furthermore, the pigments must resist being picked up by the water from the dampening system.

Inks with various special qualities exist for both letterpress and offset. In high-gloss inks, the vehicle is not homogeneous, as with ordinary inks, but heterogeneous, based on synthetic resins dissolved in a solvent, with lead and cobalt additives. This ink glazes as it dries. When printing several colours, the whole series of operations must be finished before the ink has time to dry so that the inks can attach themselves to the surface.

Quick-setting inks utilize a vehicle that also has a base of resins dissolved in a quick-drying solvent.

Heat-set inks require the application of heat to facilitate both the oxidation process and the evaporation of the solvent, as well as the penetration of certain elements that had rendered the ink more fluid.

Cold-set inks are hardened by chilling after printing, having been kept fluid by heat until they were applied to the typeform.

Moisture-set inks dry when moistened, either by the paper or by being sprayed with water. In these inks, which are used more in the United States than in Europe, the vehicle is a solvent, soluble in water, that, on contact, penetrates the paper, leaving the pigment on its surface. Odourless, they are used for printing food packaging.

Among other special-characteristic inks are metallic inks containing powdered copper, bronze, aluminum, or gold mixed with the pigment; magnetic inks, containing a powdered magnetized iron mixed with the pigment for "recognizing" the shape of printed characters as they pass before electronic reading equipment; and fluorescent inks.

Rotogravure uses fluid inks in which the colouring agent, fixed on a natural or synthetic resin, is integrated in a fluid solvent to which, just before printing, a second, extremely volatile solvent is added.

Flexography also uses fluid inks whose pigments or colouring agents are dissolved in pure alcohol, in alcohol solutions, or in water.

Serigraphy uses inks of extremely varied consistency, depending on the surface to which they are to be applied; some are little different from ordinary paint, except that their composition must not be such that rapid drying would clog the mesh of the screen.

THE FUTURE OF PRINTING

Economic considerations determine the direction of the present evolution of printing and indicate its probable short- and long-term aims. The elimination of lead, now technically possible, has begun and can only accelerate. Up to the present, this has been demonstrated by the advance of offset at the expense of letterpress. There is a movement toward offset daily newspapers when for a long time the flexibility of letterpress composition and makeup and the simplicity of stereotype plates guaranteed the least delay between event and publication. The high performance rate of photocomposition, the success of the computer, and the simplified processes for preparing plates have all reduced and even removed this advantage of letterpress.

Letterpress has reacted by borrowing some of offset's advantages, by turning to the use of metal or plastic wraparound plates, or by developing indirect letterpress. The superior sharpness of reproduction of texts still attributed to letterpress indicates that competition between the two processes is far from ended.

The convenient use of rapidly made, easily replaceable plates to obtain the best quality of image reproduction has also enabled offset to gain ground at the expense of rotogravure, preparing the cylinders for which is a long and difficult job. It does not appear that even electronic methods for engraving cylinders can halt the recession in rotogravure.

Phototypesetting has seen considerable growth in its programmed applications. The only obstacles it has encountered were the large investments in lead-composing equipment and the uncertainty of printers when faced with equipment that might not pay back its purchase price, given the explosive rate of technical advance. In the 1970s phototypesetting promises to become widespread and expand at a regular rate. The most important problems that it poses in the future are on the human level: the preparation of idiot tape means the mechanization of a human activity; that is, the downgrading of a trade that was considered an art by those who practiced it. On the other hand, to make a profit from phototypesetting machinery capable of very high output, a great volume of work is necessary. It is thus likely to stimulate competition and create difficulties at the level of medium-sized establishments. The enlarged range of offset equipment and its facility in use help these medium-sized establishments to continue to compete. Furthermore, at their level, cold type augurs well as a supplement to letterpress that is in harmony with offset printing.

In the longer run a completely new development may be foreseen: the rise of home printing. Owning or renting an electrostatic printing machine could allow any person, without the slightest training in printing, to produce his own magazine or book simply by using diapositives that might be obtained by subscription and distributed by mail.

Also to be considered is the possible presence in each house of an electronic image maker, at the moment used only to receive television broadcasts. Such a device might be justified in the future by extending its range of uses. Among these would be electrostatic printing by facsimile, following the principle of the homemade newspaper with which the *Mainichi Shimbun* of Japan experimented in 1964. Transmission of print would obviously not be confined to reproducing a newspaper. The subscriber might be offered, on different channels and following a programmed timetable, a choice of magazines, novels, and books that would not take longer than a few minutes to print and that would be made up automatically before being delivered from the machine.

Home printing presents no major technical difficulty. The only obstacle, cost, will disappear under the conditions of mass production and distribution of such equipment for printing written and illustrated information. Such a development would be a response to the conditions of the times, similar to and with the same objectives as Gutenberg's printshop in his.

BIBLIOGRAPHY. The history of printing has been the subject of many volumes. See D.C. MCMURTRIE (ed.), *The Invention of Printing: A Bibliography* (1962). Books retracing the origins of typography include: L.P.V. FEBVRE and H.J. MARTIN (eds.), *L'Apparition du livre* (1958); P. BUTLER, *The Origin of Printing in Europe* (1966); E.G. DUFF, *Early Printed Books* (1893, reprinted 1968); and T.L. DE VINNE, *The Invention of Printing* (1876, reprinted 1969). Biographical works on contributors to the invention of printing are: S. JENNETT, *Pioneers in Printing* (1958); J. GUIGNARD, *Gutenberg et son oeuvre* (1963); G.P. WINSHIP, *Gutenberg to Plantin* (1968); and A. VAN DER LINDE, *De Haarlemsche Costerlegende* (Eng. trans., *The Haarlem Legend of the Invention of Printing, by Lourens Janszoon Coster*, 1871, reprinted 1968). The subject of an entire family devoted to the typographic art is treated by A.J. GEORGE in *The Didot Family and the Progress of Printing* (1961). The various stages of evolution of the techniques of printing are developed in: S.H. STEINBERG, *Five Hundred Years of Printing*, 2nd ed. rev. (1962); M. AUDIN in M. DAUMAS (ed.), *Histoire générale des techniques*, 3 vol. (1962–68; Eng. trans. of vol. 1–2, *A History of Technology and Invention*, 1965–69); I.B. SIMON, *The Story of Printing from Wood Blocks to Electronics* (1965); J. WATSON, *The History of the Art of Printing* (1713, reprinted 1965); C.A. CLAIR, *Chronology of Printing* (1969); and I. THOMAS, *The History of Printing in America, with a Biography of Printers* (1874; 2nd ed., 2 vol., reprinted 1967). One particular aspect of the history of printing is treated in C.H. BLOY, *A History of Printing Ink: Balls, and Rollers, 1440–1850* (1967). Both history and technology are covered in P. LUCKOMBE, *The History and Art of Printing* (1771, reprinted 1965); G. BAUDRY and R. MARANGE, *Comment on imprime*, 3rd ed. (1966), and also in G.A. STEVENSON, *Graphic Arts Encyclopedia* (1968). The specialized language of the profession and the technical terms that are appropriate to it are collected in the following dictionaries: *Pocket Encyclopedia of Paper and Graphic Arts Terms* (1960); W.M. PEPPER, *Dictionary of Newspaper and*

Printing Terms: English-Spanish, Spanish-English (1959); E.M. ALLEN, *Harper's Dictionary of the Graphic Arts* (1963); W.A. SAVAGE, *Dictionary of the Art of Printing* (1841, reprinted 1966); W.W. PASKO (ed.), *American Dictionary of Printing and Bookmaking* (1894; reprinted with new introduction, 1967); C.T. JACOBI, *The Printer's Vocabulary* (1888, reprinted 1969); and R. HOSTETTLER, *Technical Terms of the Printing Industry*, 5th rev. ed. (1969), which gives the equivalents of the words in five languages. In addition to general technical works on printing, such as F. PATEMAN and L.C. YOUNG, *Printing Science* (1963); and R.R. COUPE, *Science of Printing Technology* (1966); there are numerous treatises for professionals, such as: C.A. HURST and F.R. LAWRENCE, *Letterpress, Composition and Machine Work* (1963); E.A.D. HUTCHINGS, *Printing by Letterpress* (1964); V.S. GANDERTON and H. COPELAND, *Cylinder Presses*, 2nd ed. (1965); L. HEITNER, *Introduction to Offset* (1964); R.R. KARCH and E.J. BUBER, *Graphic Arts Procedures: The Offset Processes* (1967); J.E. COGOLI, *Photo-Offset Fundamentals*, 2nd ed. (1967); A. KINSEY, *Introducing Screen Printing* (1967); A. KOSLOFF, *Photographic Screen Process Printing*, 3rd ed. (1968); A.H. PHILLIPS, *Computer Peripherals and Typesetting* (1968); E.A. APPS, *Printing Ink Technology* (1958) and *Ink Technology for Printers and Students* (1963). On the economics of the printing industry, see *Printing Industry in Britain, U.S.A. and Japan* (1964), ed. by the NATIONAL PRODUCTIVITY COUNCIL, NEW DELHI. Some philosophical and social elements are discussed in MARSHALL MCLUHAN, *The Gutenberg Galaxy: The Making of Typographic Man* (1962); J. CARTER, *Printing and the Mind of Man* (1967); and W.M. IVINS, *Prints and Visual Communication* (1969).

(Ro.L.)

Printmaking

To the modern reader, the word print might suggest mechanically mass-produced commercial products, such as books, newspapers, and textiles. In this article the print refers to the original creation of an artist who, instead of the paintbrush or the chisel, has chosen printmaking tools to express himself.

The fine print is a multiple original. Originality is generally associated with uniqueness, but a print is considered original because the artist from the outset intended to create an etching, woodcut, or other graphic work and thus conceived his image within the possibilities and limitations of that technique. Without doubt, early printmaking was strongly influenced by a desire for multiple prints. Artists quickly discovered, however, that when a drawing is translated into a woodcut or engraving it takes on totally new characteristics. Each technique has its own distinctive style, imposed by the tools, materials, and printing methods. The metamorphosis that takes place between drawing and print became the strongest attraction for the creative artist. It is important to understand that the artist does not select his printing method arbitrarily but chooses the one in which he can best express himself. Thus, any of the proofs printed from an original plate is considered an original work of art, and, although most fine prints are pulled in limited quantities, the number has no bearing on originality, only on commercial value.

Difference between a reproduction and an original print

What is the difference between a reproduction and an original print? In the very early days of printmaking this was not a serious problem because the print was not looked upon as a precious art object, and prices were low. The question of originality became an issue only in the 18th century, and, in the 19th century, artists started to hand sign their prints. Since then, the signed print has been accepted by most people as the proof of its originality.

With regard to the name with which he signed his works, the Japanese artist followed a bewildering custom: he adopted and discarded names at will. If he admired another artist, he simply adopted his name. Thus, in the art history of Japan, it is common to find several unrelated artists bearing the same name and one artist bearing many names; during his long life, Hokusai, for example, used some 50 different names.

In fact a signature by itself means little or nothing. For instance, Pablo Picasso has issued many signed reproductions of his paintings; on the other hand, many of his original etchings have been published in split editions, some signed, some not. These unsigned etchings are original, while the signed reproductions are not. The crucial difference is that Picasso made the plate for the original print, while the signed reproduction was photomechanically produced.

In 1960 the International Congress of Plastic Arts drafted a resolution intended to regulate contemporary prints. The crucial paragraph reads:

> The above principles apply to graphic works which can be considered originals, that is to say to prints for which the artist made the original plate, cut the woodblock, worked on the stone or any other material. Works that do not fulfill these conditions must be considered "reproductions."

Although this is a straightforward statement, recent developments have proved it to be highly controversial. Since the rise of the Pop and Op movements, a great number of photographically produced prints have been published and sold as signed originals. Because museum curators, art critics, and artists did not take a firm stand on the question, any print that the artist declares to be original is now accepted as such, regardless of how it was made. Although the art world is divided on the solution, nearly everybody agrees that something should be done to clarify the situation.

Many artists believe that the answer lies in the giving of honest information. In the 17th and 18th centuries in the West, most prints carried all the relevant information on their margins. The name of the individual was followed by a Latin abbreviation indicating his role in the work. Common examples are *del.* (*delineavit*): "he drew it"; *imp.* (*impressit*): "he printed it"; and *sculp.* (*sculpsit*): "he engraved it." This type of information, together with the total edition number, should be furnished by the artist or the dealer to the buyer. Clearly, it is impossible to make completely rigid rules to define originality. Probably the most realistic solution is to establish degrees of originality, based on the degree of the artist's participation in the various steps in the creation of the finished print.

Numbering editions and signing prints

There may also be confusion about edition numbering. In contemporary printmaking, an original print in limited edition should carry information about the size of the total edition and the number of the print. A problem can arise because, in addition to the regular edition, there are "artist's proofs" or the French "H.C." (*hors de commerce*) proofs. These were intended for the artist's personal use and should be no more than 10 percent of the edition; but, unfortunately, this practice is often abused, and additional quantities are printed. All the prints pulled between working stages are called "trial proofs." These can be of great interest because they reveal the artist's working process and of great value because the number of proofs is small.

With prints of old masters in the West, originality is a very complex and difficult issue. These artists did not publish their prints in limited editions but printed as many as they could sell and without signing or numbering their works. There are arguments even between experts about the authenticity of many old prints. Important works of the masters are documented in catalogs and, although these must be revised from time to time, they furnish the only firm information available.

After the edition is printed, the modern artist usually either destroys the plate or marks ("strikes") it in a distinctive manner to guarantee that any reprint from the plate is identifiable.

The 19th-century American painter and etcher James McNeill Whistler was one of the first Western artists to hand sign his prints. Signing is now regulated by a convention. Upon completing the edition, the artist signs and numbers each print. Usually the signature is in the lower right corner; the edition number is on the left. Some artists put the title in the centre.

This article contains the following major sections:

I. Major techniques of printmaking
Relief processes
Intaglio processes
Surface-printing processes
Special techniques

Process prints
Contemporary experimentation
Mounting and care of prints
II. History of printmaking
Printmaking in the 15th century
Printmaking in the 16th century
Printmaking in the 17th century
Printmaking in the 18th century
Printmaking in the 19th century
Printmaking in the 20th century

I. Major techniques of printmaking

The techniques of printmaking are divided into three major categories: relief processes, intaglio processes, and surface processes. The surface processes are subdivided into two categories: planographic (lithography) and stencil methods. The methods can be and often are combined.

RELIEF PROCESSES

In relief processes, the negative, or nonprinting part of the block or plate, is either cut or etched away, leaving the design standing in relief. Or, instead of cutting away the background, the relief print can be created by building up the printing surface. The relief is the positive image and represents the printing surface. The most familiar relief-printing materials are wood and linoleum, but many other materials can be used, such as aluminum, magnesium, and plastics. Any metal or plastic plate incised or worked in relief can be first inked in the depressions (intaglio inked) and then surface rolled, thus combining relief and intaglio processes.

Relief printing lends itself particularly to a bold conception of design, expressed more in areas than lines. This varies, however, depending on the material used: metal allows more intricate detail than wood, for example.

Woodcut. Woodcut, which appeared in the 8th century in the East and in the early 15th in the West, is the earliest relief-printing method. In this method, the design is first either painted directly onto the wood block or pasted on it. Then the surface of the wood is cut away around the design. For fine details and outlines the knife is used; larger areas are removed with gouges. The depth of the relief depends on the design: open areas must be cut deeper than the fine details so that the roller will not deposit ink in these areas. Although woodcuts are generally conceived in bold lines, or large areas, tonal variations can be achieved with textures, a variety of marks made with gouges, chisels, or knives. In contemporary woodcuts many other methods such as scraping, scratching, and hammering are also used to create textures.

Originally, woodcut was a facsimile process; *i.e.*, the cutting was a reproduction of a finished design. With most contemporary woodcuts, however, the artist creates his design in the process of cutting.

Wood for blocks

As wood is a natural material, its structure varies enormously and this exercises a strong influence on the cutting. Wood blocks are cut plankwise. The woods most often used are pear, rose, pine, apple, and beech. The old masters preferred fine-grained hardwoods because they allow finer detail work than softwoods; but modern printmakers value the coarse grain of softwoods and often incorporate it into the design.

The printing of woodcuts is a relatively simple process because it does not require great pressure. Although presses are used, even hand rubbing with a wooden spoon can produce a good print. The ink used to print woodcuts must be fairly solid and sticky, so that it lies on the surface without flowing into the hollows. The printing ink can be deposited on the relief either with dabbers or with rollers. Japanese rice or mulberry papers are particularly suitable for woodcuts because they make rich prints without heavy pressure.

Colour woodcut. The standard procedure for making a woodcut with two or more colours is to cut a separate block for each colour. If the colour areas are distinctly separated and the block is large, one block can be used for more than one colour. All blocks must be the same size to assure that in the finished print the colours will appear in their proper relation to one another, that is, properly registered.



"Head of Ludwig Schames," woodcut by Ernst Ludwig Kirchner, 1918. 58.4 cm × 26 cm.
By courtesy of the Museum of Modern Art, New York, gift of Curt Valentin

The first, the key block, is generally the one that contains most of the structural or descriptive elements of the design, thus serving as a guide for the disposition of the other colours. After the key block is finished and printed, the print is transferred to the second block. This procedure is repeated until all the blocks are finished.

The registering system depends on the method of printing used. On a press the registering presents no problem: the wood block is locked into position and the uniformly cut paper is automatically fed into the proper position by the press. For hand rubbing, several registering methods can be used. One method uses a mitred corner nailed to a table or special board. A sheet of paper is attached to one side of this corner, after which the wood block is placed securely in position and the print is made. Once the first colour has been printed, the paper is folded back and the first block is replaced with the second, and so on.

In woodcut colour printing, the artist must consider whether he can print wet on wet or whether the print should dry before it is overprinted. Usually a second colour can be printed immediately but, if the ink deposit is heavy, the print will have to dry before additional colours can be printed. This problem arises mainly with oil colours, which dry more slowly than water-base colours.

Use of small movable blocks

Movable small blocks are also used today by a number of printmakers. These involve some planning in order to print them in register with the large blocks. The easiest way is to put a light cardboard, exactly the size of the main block (the key block), in position. Once the small blocks are registered, their location can be marked on the cardboard. Then the small blocks can be glued down to the cardboard in order to avoid the danger of shifting.

The conception and technique of the Japanese colour woodcut was totally different from that of the European woodcut. Except for chiaroscuro prints, no real colour

woodcut existed in Europe before the 19th century. In the West, the woodcut was primarily a reproductive facsimile process: usually, the artist made a completed drawing that was copied by the cutter. The Japanese print, on the other hand, was the result of intricate, perfectly coordinated effort by the designing artist, the cutter, and the printer. Instead of painting a complete picture to be copied, the artist furnished a separate drawing for each colour. The engraver or cutter pasted each drawing on a wood block and cut away the white (negative) part. In this process the drawing was destroyed. Printing started only after all of the blocks had been cut. As the Japanese used water-base colours, often blending tones, printing itself was a very delicate and crucial operation, requiring perfect coordination and speed. Only after the completion of this process could the artist see the total image.

Wood engraving. Wood engraving is a variation of woodcut. The main difference is that, for wood engraving, the block—usually pear, apple, cherry, sycamore, or beech—is cut cross-grain rather than plankwise; on the end-grain block the artist can thus cut freely in any direction, allowing him to do much more intricate work with much finer tools. The image is created by fine white lines

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



Wood engraving by William Blake, 1820–21, for R.J. Thornton's *Pastorals of Virgil.* 3.5 cm × 7.2 cm.

and textures. On most wood engravings, the whites appear as the positive image against a dominant black. The blocks are usually cut at the same height as printing type so that they can be printed on a press. Invented in the 18th century, wood engraving was primarily used by illustrators.

Linoleum cut. Since linoleum is easy to cut and does not have a grain, the linoleum cut is used often to introduce children to printmaking. The process was held in low esteem until, in the 1950s, Pablo Picasso made a series of brilliant colour linoleum cuts.

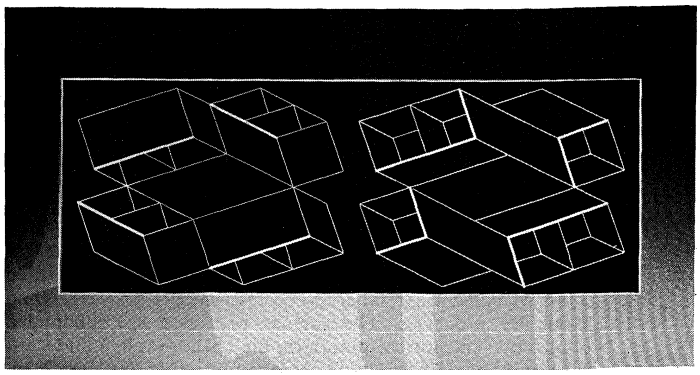
The printing of linoleum cuts is similar to the printing of woodcuts or wood engravings. They can be printed by hand rubbing or, properly mounted, can be printed on a press. The colour printing process follows the woodcut principles.

Metal cut. At times artists have used soft metals, such as lead or zinc, to make prints similar to woodcuts or wood engravings. In the 19th century, lead cuts were often used for newspaper illustrations. The distinguished Mexican artist José Guadalupe Posada, for example, used lead frequently for his prints. Lead was used primarily because it was inexpensive and easy to work. Because metal cuts were printed like woodcuts or wood engravings, it is often difficult to tell from the print which material was used.

Cardboard (paper) cut. Elementary school children are often introduced to printmaking by making cardboard cuts, and sophisticated artists use the same material to print complex abstract images. Cardboard and paper

Advantages of cardboard and paper

are not only inexpensive, readily available, and workable with simple tools but, when properly prepared, have also proved to be remarkably durable. Cardboard cuts can be made either by building up or cutting out. In the first process, cut-out pieces are glued to a support. When the plate is finished, it is coated with a plastic varnish to make



"Duo H," inkless intaglio print (relief of design etched on plastic and mounted on Plexiglas) by Josef Albers, 1966. Image 34.3 cm × 12.7 cm.

By courtesy of Multiples, Inc., New York and Los Angeles; photograph, Ralph Tornberg and Associates

sure the surface is tough and nonabsorbent. In the cutting-out method a heavy laminated cardboard is used and the cut out sections are simply peeled off to the desired depth. When finished, the cut is varnished. The printing of cardboard plates follows the same principle as woodcuts or linoleum cuts.

Relief etching. When large areas of a metal plate are etched out (see below *Etching*), leaving the design in relief to be surface printed, the process is generally called relief etching. Usually the method is used for areas, but it can be also used for lines. The English artist and poet William Blake was the first printmaker to experiment extensively with relief etching. He devised a method of transferring his handwritten poems onto the metal plate to be etched together with the illustrations. In contemporary printmaking, relief etching is used extensively for colour printing. The different levels of the plate can be inked with different colours. Relief etching is also a popular method of making inkless intaglio prints (shallow bas-reliefs on paper).

Rubbings. Simply by inking the surface, placing a fine paper over it, and rubbing, practically any surface can be used for printing—including, as in Japan, the body of a fish. Rubbings were probably the earliest prints made by man. In India and China rubbings were made of tombstones and temple bas-reliefs. In addition to fish rubbings, the Japanese made rubbings of metal ornaments. Today

By courtesy of the Philadelphia Museum of Art, gift of Horace H.F. Jayne



Scenes from the tomb of the Wu family in Shantung, China, stone rubbing, AD 147 (Han dynasty). 7.5 m × 6.3 m.

many museums sell rubbings of bas-reliefs in their collections. Rubbings can be made with the simplest tools and materials: rice, mulberry, or any other woodcut paper; water-base inks (because they can be washed off, avoiding permanent stains on porous surfaces); a hard rubber roller to deposit the ink, and a wooden spoon or small polished wood block for rubbing.

Dotted print (criblé). A traditional technique of the goldsmith long before engraving for printing purposes was developed, criblé was also used to make the earliest metal prints on paper. Criblé was a method of dotting the plate with a hand punch; with punch and hammer; with a serrated, flatheaded tool called a matting punch; with various gouges; or sometimes, with a hollow, circular-headed ring-punch. Criblé plates were relief printed like woodcuts. On most dotted prints, a black background dominates a fine lacelike design.

INTAGLIO PROCESSES

Intaglio printing is the opposite of relief printing, in that the printing is done from ink that is below the surface of the plate. The design is cut, scratched, or etched into the printing surface or plate, which can be copper, zinc, aluminum, magnesium, plastics, or even coated paper. The printing ink is rubbed into the incisions or grooves and the surface is wiped clean. Unlike surface printing, intaglio printing—which is actually a process of embossing the paper into the incised lines—requires enormous pressure. The major working methods for intaglio printing are engraving, etching, drypoint, and mezzotint. Intaglio processes are probably the most versatile of the printmaking methods, as various techniques can produce a wide range of effects, from the most delicate to the boldest. The intaglio print also produces the richest printed surface, as it is three-dimensional.

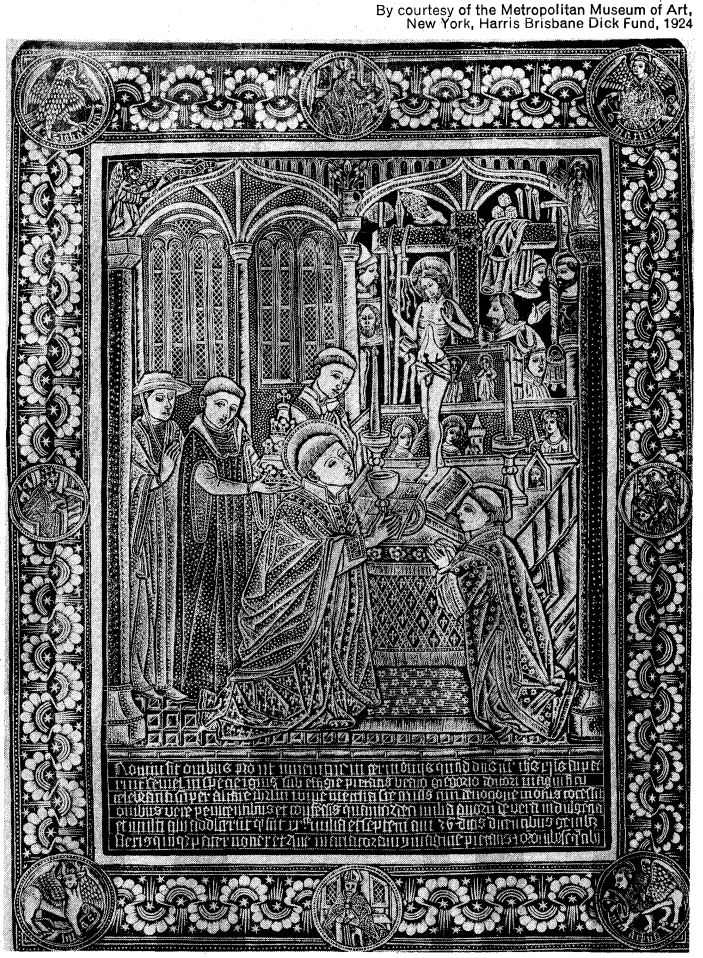
Pressures required for intaglio printing

Engraving. In engraving, the design is cut into metal with a graver or burin. The burin is a steel rod with a square or lozenge-shaped section and slightly bent shank. The cutting is done by pushing the burin into the metal plate. The deeper it penetrates into the metal, the wider the line; variations in depth create the swelling–tapering character of the engraved line. After the engraving is finished, the slight burr raised by the graver is cleaned off with a scraper. The engraved line is so sharp and clean that it asserts itself even if cut over a densely etched area. In the print, the engraved line is notable for its precision and intensity. In engraving, the hand does not move freely in any direction but pushes the graver forward in a line; change of direction is achieved by the manipulation of the plate with the other hand. Although copper, zinc, aluminum, and magnesium plates are used—and in the past, soft iron and even steel were used—the best all-around metal is copper. It is preferred because it has the most consistent structure and is neither too soft nor too hard.



"St. John on the Isle of Patmos," engraving by Master E.S. (active c. 1440–c. 1467). 20.6 cm × 14.1 cm.
By courtesy of the Art Institute of Chicago, Mr. and Mrs. Potter Palmer Collection

By courtesy of the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1924



"The Mass of St. Gregory," criblé, or dotted, print by an unknown German artist, 15th century. 35.2 cm × 25.1 cm.

Drypoint. Next to engraving, the drypoint is the most direct of the intaglio techniques. In printing, however, it represents the opposite end of the spectrum. Engraving is precise, drypoint is rugged, warm, and irregular.

Drypoint is made by scratching lines into metal plates with steel- or diamond-point needles. In this method the penetration into the plate is negligible; it is the metal burr raised by the point that holds the ink. As the burr is irregular, it prints as a soft velvety line. The angle of the needle has much more effect on the width of the line than the pressure does: if the needle is perpendicular to the plate it throws burr on both sides, which then produces a thin double line; for wide lines the optimum angle is 60 degrees.

Importance of needle angle

Copper plate is the best for drypoint. The plates are fragile because the burrs are easily flattened down by the printing pressure. Even a too vigorous wiping can damage a plate. Thus, unless the artist is satisfied with a very limited number of proofs (three or four), the plate must be faced with steel, a process in which steel is deposited by electrolytic means on the copper plate. This coating is



"Self-Portrait with Bowler Hat," drypoint by Max Beckmann, 1921. 31.3 cm × 24.5 cm.
By courtesy of the Museum of Modern Art, New York

very thin and if it is properly done, the burrs are hardened without affecting printing quality. Zinc and aluminum, however, cannot be steel-faced.

Mezzotint. In mezzotint the metal plate is roughened with fine burrs until it prints a rich, velvety black. The plate is then worked back toward the lighter values with scrapers and burnishers. For this reason, mezzotint is also called *manière noire*, or the "black manner."

Mezzotint flourished throughout the 18th and 19th centuries and was primarily used for portraits or to reproduce paintings. None of the important printmakers of the past used the technique. After the invention of photoengraving, mezzotint was nearly forgotten, but in recent years a few printmakers have started to work again with this exotic medium.

Use of the rocker

The first step in preparing a mezzotint plate is to rough up the whole plate surface as evenly as possible. The tool used is the rocker, a blade with a curved serrated edge. The rougher the rocker, the heavier is the burr. The rocker is held with its cutting edge at a right angle to the plate, and the curved edge is rocked systematically over the entire surface. If this is properly done, the entire plate is covered with uniform burrs. Then the work with scrapers and burnishers begins. Where lighter tones are desired, the burr is gradually removed, and in the white areas the plate is burnished back to its original finish.

As with drypoint, mezzotint plates must be steel faced if a large edition is desired. The printing of mezzotints differs slightly from the printing of etchings or engravings. Since the layer of burr on the mezzotint acts as a blotting paper, the ink must be selected with this fact in mind. The inking and wiping must be done gently with soft rags. Printing pressure should be considerably less than that used for engravings or deeply etched plates.

Crayon manner and stipple engraving. Invented in the 18th century, crayon manner was purely a reproduction technique; its aim was the imitation of chalk drawings. The process started with a plate covered with hard ground (see below *Etching*). The design was created using a great variety of etching needles (some of them multiple). After the design was etched in, the ground was removed and the design further developed with various tools. Fine corrections and tonal modifications were made with scrapers and burnishers. Finally, engraving was used for additional strengthening of the design. Pastel manner is essentially the same as the crayon manner except that it is usually used to imitate pastel drawings.

Stipple engraving, also a reproduction method, is closely related to the crayon manner. The exact date of its invention is not known, but it is reasonably certain that it came after the crayon manner. The first step in stipple engraving was to etch in the outlines of the design with fine dots made either with needles or with a roulette, a small wheel with points. The tonal areas were then gradually developed with tiny flick dots made with the curved stipple graver. For very fine tonal gradations, roulettes were also used. The only artist of any importance to use pure stipple engraving was Giulio Campagnola in the 16th century.

Etching. Etching is a process in which lines or textures are bitten (etched) into a metal plate with a variety of mordants (acids). The metal plate is first covered with an acid-resistant coating (ground). The design is then scratched or pressed into the ground, exposing the metal in these areas. Finally, the plate is submerged in an acid solution until the desired depth and width in the exposed areas is reached.

Although the basic principle of etching is very simple, there are many possible variations that have a strong influence on the final result. The materials themselves offer a wide range of possible variations; for example, copper, zinc, aluminum, or magnesium plates can be used; and nitric acid, hydrochloric acid, or ferric chloride can be used for the etching process. Other variations include the strength of the mordants, the biting time, the kinds of grounds and the ways in which they can be worked, and, finally, all the possible methods of printing.

Although all these matters seem purely technical, every tool or material used, every step followed, is an integral part of the creative process. The biting action of the acid is just as much part of the drawing as the incising into the ground. The selection of the paper or the method of wiping the plate can completely change the nature of a print.

Hard-ground etching. Any acid-resistant coating used to make an etching is called a ground. In the past a great variety of different grounds were used, and each master had his own formula. Most of them had wax as a basis, combined with various oils and varnishes. Today, the most commonly used ground consists of two parts Egyptian asphaltum, two parts beeswax, and one part resin. These ingredients are either dissolved and mixed or fused by heat. Ground comes either in lump or liquid form.

The plate is cleaned before the ground is applied because grease or dirt can affect the ground's adhesion, making it peel or crack. If ground in solid form is used, it is melted on low heat and rolled out evenly. Liquid ground is brushed on the plate, then heated to make it more even and to evaporate the solvents. In both cases, after the plate cools, the ground should be solid (not sticky).

By courtesy of the Museum of Modern Art, New York



"Minotauromachy," etching by Pablo Picasso, 1935. 49.5 cm × 69.7 cm.

Normally, a good ground is dark enough to offer sufficient contrast with the plate to see the work. If, however, a black ground is desired, it can be achieved by darkening the ground with the smoke of a candle.

Variety of modern etching tools

In etching the ground, any number of tools and instruments may be employed. The old masters were restricted, but the contemporary printmaker uses a whole arsenal, including electrical drills and gravers. The line produced by the etching needle is threadlike and uniform in thickness. Other instruments are used to introduce a great variety of marks. The character of the etching is further influenced by the choice of the metal and the type of acid used. For controlled, regular bite, it is common to use Dutch mordant (nine parts of water saturated with potassium chlorate to one part of hydrochloric acid) on copper. For a rugged, irregular bite, nitric acid (one part to nine parts of water) is used on zinc. A plate can be etched in stages by covering some of the already etched areas with stop-out varnish (rosin dissolved in alcohol), which resists the acid, and then etching the rest for a longer period. This procedure can be repeated many times. Most artists develop their plates by repeated bites. After the etching is finished, the ground is removed with solvent (such as kerosene or benzene), and the plate is printed.

The first print is a state, or trial, proof. If further work is desired, the plate is cleaned and covered again with ground, the previous work remaining visible through the new ground. The whole process is repeated as many times as is necessary.

Soft-ground etching. Soft-ground etching is basically the same as hard-ground etching except that the ground contains about one-third grease, which keeps it in a semihard, or tacky, condition.

Initially, in the 19th century, soft ground was used primarily for offset drawings. The artist placed a paper on the grounded plate and made his drawing on the paper with a sharp pencil or other drawing instrument. Under the pressure, the paper picked up the ground and produced a soft granular line. Then the plate was etched normally with a fairly weak acid.

At the present, soft ground is used more often to etch various textures into the plate. Textured materials are placed on the soft ground and the plate run through the press. A thin, even ground picks up the finest textures. The design is controlled by applying a stop-out varnish to areas that should not be etched. The remaining textures are etched into the metal in the same way as in conventional hard-ground etching. This technique lends itself well to collage-type effects on the plate.

Relief etching. To make a relief etching, the areas not to be removed by acid are protected with liquid ground or varnish. The varnish used has to be tough (asphaltum, or ground) because the relief bite takes a long time, and when large areas are bitten, the plate has a tendency to heat up. If various levels are desired, relief etching can be done in stages, as in regular etching.

Aquatint. Aquatint is a process used to etch tonal areas on the plate. The first step is to give the plate a porous ground by dusting it with rosin powder and fusing the powder to the plate by means of heat. When the plate is etched, the acid goes through the pores in the ground and bites tiny cavities in the metal. These cavities hold the ink. A variety of tones and textures can be created, depending on the density, width, and depth of the cavities.

The aquatint method was invented in the 18th century, and, although a great number of pure aquatint plates were done, the technique was mainly used with line etching. Theoretically, there is no limit to the range of tones that can be etched with aquatint.

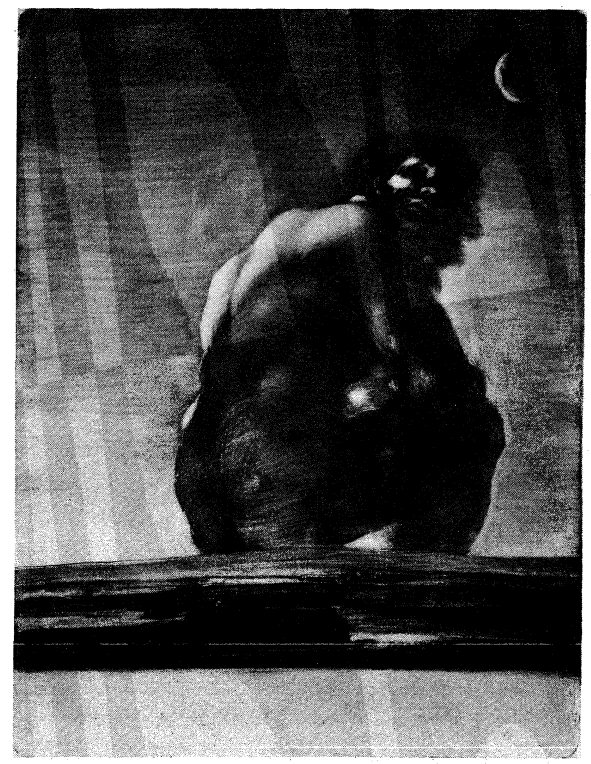
For aquatint, the plate is cleaned, as in hard-ground etching, and then dusted with rosin. Care in this step is crucial, as an incorrectly distributed rosin ground will produce uneven, spotty tones. To achieve even tones, a fine-grain rosin is used. The quantity should cover about 50 percent of the surface, neither too thin nor too thick. The dusting can be done either with a dust box or with dust bags.

Use of the dust box

The dust box is a completely enclosed container with a sliding tray (usually made of steel mesh) that holds the plate in position above the dust tray, which is filled with fine rosin dust. After the plate is placed in the box, the rosin dust is agitated either by a bellows, by an electric fan, or by shaking.

Dusting bags are made of various materials; the finer the material, the finer the dust coming through. The dusting bags have the advantage of allowing the artist to visually control the amount of dust deposited and also to use different textures in different areas.

By courtesy of the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1935



"The Colossus," aquatint by Francisco de Goya (1746–1828). 29.3 cm × 21.6 cm.

After dusting, the plate is placed on the heating plate, and the rosin is fused to the metal. When the plate has cooled, the design is applied with a stop-out varnish. To achieve various tones the plate is bitten in stages, much as in hard-ground etching but with one important difference: aquatint is much more delicate, and the time element is more critical. A biting time of a few seconds can produce a fine gray, but a proportionately longer time is needed as the artist proceeds toward the darker tones.

Plastic sprays are also used to make aquatints. These lacquers and enamels are sold in pressurized spray cans and are sufficiently acid resistant to use for moderately long bites. They are easy to control and simpler to use, but they must be used in spray booths or other well-ventilated places.

Lift-ground etching (*sugar-lift aquatint*). In lift-ground etching, a positive image is etched on an aquatint plate by drawing with a water-soluble ground. In the conventional aquatint technique, the artist controls the image by stopping out negative areas with varnish, thus working around the positive image. But for lift-ground etching, he uses a viscous liquid (such as India ink, gamboge, or ordinary poster paint mixed with sugar syrup) to paint directly on the plate. After the painting is finished and dried, the whole surface is covered with thin, liquid hard ground. When dry, the plate is placed in lukewarm water that dissolves the painted design, lifting the ground and dislodging it from the places that had been painted, thus exposing the metal surface to be etched. Aquatinting can be handled two ways: either the whole plate can be aquatinted before painting with lift ground or it can be aquatinted after the design is lifted. Lift-ground etching is particularly well-suited to free, spontaneous, calligraphic designs.

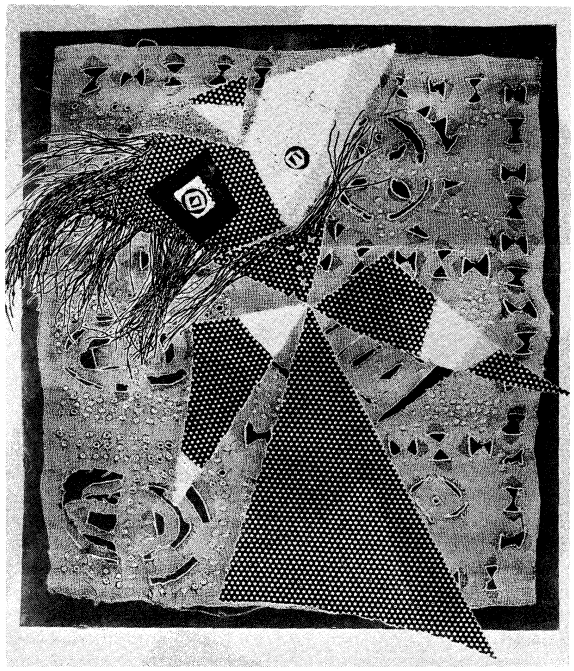
Action of nitric acid and Dutch mordant

Acids and the etching process. The acid bite of the plate is a critical stage in making an etching. The printmaker must be familiar with the characteristics of the materials he is using. On a zinc plate, nitric acid is used. In the process of biting, this acid develops air bubbles over the bitten area. Under the bubbles, the acid action is slower, and, therefore, if the bubbles are not constantly moved around by brushing, the etched line will be uneven. Nitric acid also has a tendency to underbite, that is, to bite not only straight down but sideways. For this reason, areas of dense texture must be watched closely.

Nitric acid also can be used on copper, but, except to bite out large areas, Dutch mordant is much better suited for this metal. The action of hydrochloric acid on copper is much more even and controlled than that of nitric. Thus, for a bold, rough bite, nitric on zinc is fine; but for delicate, controlled etching, Dutch mordant on copper is preferred.

Metal graphic. This method was originated by Rolf Nesch (1893–), the German-Norwegian printmaker. In all the intaglio methods previously discussed, the artist's design was created by making incisions in the plate. Nesch's method is the reverse of this process: the design is built up like a montage, by cutting out metal shapes and soldering them on the plate surface. Instead of the etching needle and the graver, the tools are shears, wire cutters, and a soldering iron. These plates are in deep relief

By courtesy of the Philadelphia Museum of Art, gift of Lessing J. Rosenwald



"Sort Dame" ("Dark Lady"), "metal graphic" print by Rolf Nesch, 1953. 57.1 cm × 50.2 cm.

and thus produce a heavily embossed print. Often such plates are combined with conventionally etched or engraved sections. In addition to metal shapes, wood and plastics may be used. Because of the extremely high relief, the printing of the plates requires specially prepared presses. A few contemporary artists work in such a high relief that the ordinary etching press cannot print their work and standard printing papers cannot be used. In some cases, the high relief is created by compressing paper pulp into molds with hydraulic presses.

The use of embossing is not new. Some Japanese woodcuts have sections decorated with "goufrage" (blind pressing). In contemporary printmaking, embossing has become a major interest, and many artists are exploring the possibilities of the intaglio print by using shallow paper bas-reliefs to exploit the interplay of shadow and light.

Printing by intaglio processes. The most important piece of equipment in intaglio printing is the etching press, a simple machine whose basic principle has not changed for centuries. Motorization and the use of pressure gauges are the only major improvements. The press consists of a solid steel plate (bed) driven between two rollers; a screw mechanism on both sides of the top roller adjusts the pressure. Large modern presses are motor driven.

The etching press

The print is made by placing the inked plate face up on the bed. Dampened paper is placed carefully on the plate and covered with several layers of pure wool printing felts. The bed is then driven through the rollers. The felts, squeezed between the metal rollers and the plate, are pushed into the crevices of the plate, forcing the paper into contact with the ink and thus transferring the image.

A fairly heavy pure rag paper is normally used. It is soaked until its fibres are softened and then, before printing, it is blotted until no surface water is visible.

Inks for intaglio printing are especially made for this purpose. The consistency of the ink must be such that it comes off the plate cleanly during the wiping operation, but at the same time it must have enough body to retain its relief on the paper. The printing ink must also have sufficient viscosity to stick to the damp printing paper to produce a clear and rich image. For inking, the plate is placed on a heater and kept warm throughout the inking and wiping steps. Heat makes the ink looser and thus facilitates both of these processes.

Wiping is the operation in which the ink is removed from the surface of the plate, while leaving it in the recesses. Usually a carefully folded starched cheesecloth (tarlatan) is used. When a clean, crisp print is desired, the plate is given a final wiping with the palm of the hand.

After the print is pulled, it is dried, either between blotters or taped to a large, stiff board. This choice depends on the size of the print and the type of paper used.

Intaglio colour printing. The intaglio colour print is made with two or more intaglio plates successively overprinted on the same paper. Each plate represents one colour and its possible graduations. In principle, it is possible to take four plates—the three basic colours, yellow, red, and blue, plus black—and make a print that will have the full range of colours. If the colour areas are distinctly separated, more than one colour can be printed from one plate. This method involves an extremely meticulous inking and wiping process.

Problems of colour registration

One of the greatest problems with intaglio colour printing is registering the successive colours in their precise location. If the colours can be printed immediately, wet on wet, then it is relatively simple, but often this is not possible. If the first plate has high relief and is overprinted while wet, the second plate will crush it completely. In this case the first print must be thoroughly dried and then rewetted for the second printing. Because the paper shrinks in the drying process, it is difficult to get it back to the original size when rewetted.

Several methods of registering can be used, depending on the particular problem. For wet-on-wet printing the process is simple. After both plates are inked, the first plate is placed on the press bed and its position is marked. Paper is placed over the plate and secured at one end with masking tape, or, if there is enough margin, the paper is run through so that one end remains caught under the printing roller. The print is then folded back and the first plate is replaced with the second.

Another method uses mats. The paper to be used in the edition is cut to the same size. A cardboard or metal mat is cut, corresponding to the size of the wet paper. The plate position is either cut out or marked on the mat. Registration consists of lining the paper up with the mat.

The most precise registering is with pinholes. Two pinholes are punched in the opposite corners of the mat. Corresponding pinholes are punched through all the printing papers. In printing, the paper is picked up with two heavy needles through the already punched holes. The needles are then inserted in the corresponding holes on the mat and the paper is released. The holes should be placed close to an edge that will be trimmed after the print is dry.

Stencilled colours with an intaglio plate. Stencilling is one of the simplest ways to use a number of colours

combined with an intaglio plate. This method has advantages and also limitations. The main advantage is that it eliminates the registering problems of intaglio colour printing. On the other hand, it is limited to flat, sharply defined colour areas. One method does not replace the other, but each may be used to solve a particular problem.

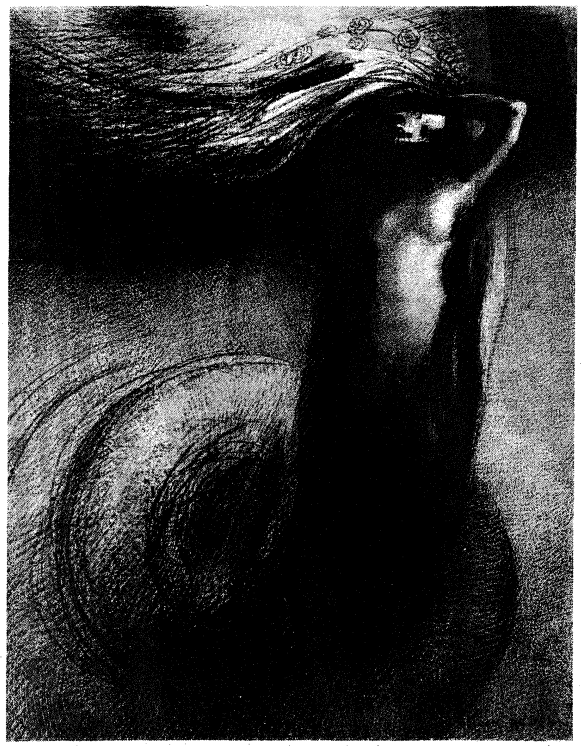
The procedure itself is very simple. The intaglio plate is inked and wiped normally. The desired colour shape is cut out on a stencil paper. The stencil is placed on the already inked plate and the colour is rolled onto the surface of the plate using a gelatin or soft rubber roller. For surface rolling, regular artist oil colours can be used. The use of stencils allows a great number of colours to be printed with a single run on the press. This is done by surface rolling colours through stencils onto the intaglio inked and wiped plate surface.

For more complex colour combinations, it is possible to combine colours stencilled directly on the paper with colours offset from the intaglio plate. For more sophisticated stencilling, silk screen can be used also in combination with the intaglio plate. When intaglio and stencilling are combined, the process is often designated as mixed or combined technique. This is essentially the same procedure as conventional stencilling except that with silk screen more complex designs and textures can also be stencilled on the plate (see below *Stencil processes*).

Intaglio and surface colour with relief etching. In this technique the main colour structure is defined by the plate surface, which is etched to different levels. The linear or textural elements moving from one level to another bind the whole together.

Inking to different levels

The sequence of printing begins with the intaglio inking and wiping of the plate. Next, the first surface colour is rolled on with a soft gelatin roller that penetrates the lower levels of the relief. The high areas are inked with a hard rubber or composition roller. The sequence of rolling can change, according to the demands of the particular colour problem.



"La Mort: Mon ironie dépasse toutes les autres!" *(Tentation de St. Antoine)*, lithograph by Odilon Redon (1840–1916). 26.2 cm × 19.7 cm.
By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection

SURFACE-PRINTING PROCESSES

Surface printing comprises those techniques in which the image is printed from the flat surface of the metal, stone, or other material. The major surface method is lithography, a planographic process. Although many experts place silk screen and stencilling in a separate category, they can be considered surface-printing processes. In lithography, the control of the design is achieved by the chemical treatments of the drawing surface. In stencilling, the design is created by holes in the stencil and the printing ink is either rolled or squeezed through the stencil onto the paper. Silk screen is a special form of stencilling.

Lithography. Lithography is based on the fact that water and grease do not mix. The image is drawn or painted on the stone or metal plate with greasy litho crayon or a greasy black ink (tusche). Once the drawing is finished, it is fixed with an etch to prevent the spreading of the grease. A heavy, syrupy mixture of gum arabic and a small quantity of nitric acid, the etch is used to protect the drawing from water and to further desensitize the undrawn areas to printing ink. The nitric acid opens the pores of the stone, enabling the gum and the grease to enter easily. The gum arabic surrounds the greasy sections, forming an insoluble surface film which sticks to the negative areas and crevices of the grain. This coating around the image repels the water applied during printing and establishes a grease reservoir. It does not smear, and it prevents seepage that would blur the image.

The etch in lithography

Because of the antipathy of grease and water, the image attracts oily ink, but repels water. Thus, when the stone is dampened with a sponge and an ink-charged roller is passed over it, the ink is deposited on the greasy drawing but not on the wet stone.

In lithography, the assumption is that the drawing made on the stone or plate will be closely duplicated on the print. While intaglio processes produce prints unlike any drawing technique, lithography is quite reproductive. Although a complex method, if lithography is well done, the effect of the print is deceptively simple and direct, making the technique attractive to artists who wish to avoid the more idiosyncratic printmaking methods.

A highly skilled technician is needed to produce a good lithograph, and most lithography is done in workshops where well-trained men are available. The artist usually works on the stone or plate under the guidance of master printers. When the artist finishes his drawing, the master printers etch the stone and do the printing.

In the basic technique, the first step is the preparation of the stone or plate. If a stone has been used before, its surface must be reground. The stone is placed in a sink, thoroughly wetted, and carborundum powder sprinkled over it. Then, either with a levigator (a heavy steel disc with a handle) or by rubbing two stones together, the surface is thoroughly reground. From time to time the surface should be tested with a steel straightedge to make sure it is level, otherwise it will print unevenly. After the stone has dried, it is ready for work. It is very important to keep the stone clean because any dirt, particularly grease, would show up on the print. Smudges and dirt can be cleaned off with erasers and abrasives.

Metal plates (zinc or aluminum) can also be used and these, too, may be reground. Although metal plates are satisfactory, stone is far superior, particularly for producing subtle tones and details.

Use of crayons and tusches on stone

With litho crayons and tusches the artist can work on the stone as he would on paper. A whole arsenal of effects is available, including pen, pencil, splashing, sprinkling, spraying, texture transfers, and scraping. After the drawing is finished and before etching, the image must be protected from the etching solution by rubbing rosin and then talcum powder on the stone. The acid-resistant rosin protects the drawing; the talcum absorbs the excess grease, allowing the adhesion of the gum etch to the edges of the drawing.

Next, the whole surface of the stone is coated with undiluted gum arabic, applied with a wide, soft brush. The subsequent etching process is done in stages. The weakest acid solution is usually brushed first on the lightest areas of the drawing. After an appropriate interval, the next strength solution is brushed on, and this continues until the strongest etch has coated the darkest areas.

After the allotted time has elapsed, the excess etch solu-

tion is blotted with newsprint paper. The surface is then wiped down and buffed with cheesecloth to a smooth, even layer. When properly handled, the stone should appear dry. It should be allowed to stand for two hours before washing out, the next step.

The wash out is done by pouring a small amount of turpentine or Lithotine over the drawn areas. Gently rubbing the drawn areas with a clean dry rag removes the drawing through the gum-etch coating. The image is preserved by the absorbed grease in the porous limestone.

Next, the stone is rubbed with liquid asphaltum or printing ink dissolved in turpentine. This procedure saturates the image and protects it at the same time.

After the stone is dry, it is ready to be inked (rolled up). First, it is dampened with a wet sponge. (In between the rollings, the stone should be redampened.) Ink rolling should be carried out according to a set pattern, gradually building up the image. To facilitate the even distribution of ink it is important to use a roller wider than the image.

The lithographic press prints with scraping pressure. The press itself consists of a metal frame that accommodates a travelling steel plate (the bed), which passes with the stone under a scraping bar (or yoke). The bed can be lowered (to position the stone) and raised (to print). The pressure on the scraping bar can be adjusted.

Lithographs can be printed on either dry or damp paper. The advantage of dampening is that it is possible to use less ink and less pressure, thus minimizing the risk of clogging the image.

To print, the printing paper is first placed on the stone, then a newsprint paper, then a blotter. Last comes the tympan, a sheet of smooth, tough material that can withstand great pressure without stretching. After the bed is raised to printing position, grease is spread evenly in front of the scraping bar on the tympan to allow it to slide easily. Then the print is made.

The prints of the French artist Henri de Toulouse-Lautrec demonstrate that lithography offers endless possibilities in colour printing. Because the effect of lithography is much more painterly than either woodcut or intaglio printing, it is natural that the strong preoccupation with pure colour in contemporary art has created a revival of interest in this medium. The planning and the principle of colour separation are similar to those for the colour woodcut or intaglio colour.

Stencil processes. In stencilling, one of the simplest methods of duplication, the design is cut out of paper (or any other suitable thin, strong material) and is then printed by rubbing, rolling, or spraying paint through the cutout areas.

The silk-screen process

Silk screen is a sophisticated stencil process, developed around 1900 and first used mainly for advertising and display work. About 1950, fine artists started to use the process extensively, giving it the name serigraphy.

The silk-screen process got its name from the fine mesh silk which, when tacked to a wooden frame, serves as a support for a cut paper stencil. The stencil is glued to the silk. In the basic process, the open mesh of the silk lets the paint through, while the paper stencil blocks it out. A design can also be blocked out on the screen with glue or other suitable substance.

By courtesy of the Kennedy Galleries, Inc., New York



"Pleiades," silk-screen print by Ben Shahn, 1960. 52 cm × 67.3 cm.

A common method of stencil preparation is to cut the stencil with a knife. In this method the artist can use commercially produced screen process printing plates or conventional stencil papers. For fine, accurate work, process plates, which consist of a film on a backing, are preferred. Areas to be printed are cut out of the film and peeled off, leaving the rest of the film on the backing paper. After the plate is attached to the screen, the backing paper is removed; thus, the screen is covered with film except in the printing areas. Process plates are available in different colours to make registering easier, and they are attached to the screen either by heat or by the use of a special solvent.

Another common method is the so-called tusche-and-glue method, which is similar to lift-ground aquatint etching. The design is painted on the screen with tusche and when dry the whole screen is covered with glue. When the glue dries, the design is washed out with either kerosene or turpentine. The tusche comes in liquid form for brushing or in solid crayon form. The use of the crayon results in screen prints deceptively resembling lithographic prints.

Stencil plates can also be made photographically. These plates are made by placing a photographic positive on a photosensitized gelatin stencil plate in a vacuum printing frame. Exposure to light hardens the gelatin under the transparent areas but leaves the gelatin soft under the dark areas. In warm water the soft areas wash out. The stencil is attached to the screen in the same manner as other stencils.

Silk-screen prints

To make a silk-screen print, the wooden frame holding the screen is hinged to a slightly larger wood board. The printing paper is placed on the board, under the screen. The consistency of the ink is important: it must be liquid enough to pass through the screen but not so liquid that it runs. The ink is pressed through the screen with the squeegee (a rubber blade set in a wooden handle, usually the same width as the screen). Any number of colours can be used, a separate screen for each colour.

SPECIAL TECHNIQUES

Monoprint (*monotype*). A monoprint is a unique print. The artist paints on various surfaces such as metal, plastic, and glass and then transfers the wet design to paper, either by rubbing or with an etching press. The primary reason for making a monoprint is that, in offsetting the image from the plate to the paper, the print achieves a separate quality and luminosity totally unlike a painting made directly on paper.

Cliché-verre. The method of printing known as *cliché-verre* was used by a few artists in the 19th century during the period when photography was a new and exciting invention. The *cliché-verre* follows the principle of photography but does not have its tonal variations. The print was made by covering a piece of clear glass with an opaque pigment or emulsion; the design was then scratched through with a sharp etching needle or stylus. When the drawing was finished, the glass plate (negative) was placed on a photosensitized paper, exposed to light, and developed. The result was a (positive) print with strong black and white contrasts.

The cellocut. The cellocut method was named by its originator, U.S. printmaker Boris Margo (1902–), one of the first to experiment extensively with plastics.

In this method, liquid plastic dissolved in acetone is poured onto a rigid support backing, such as fibre board or cardboard. The solidified plastic can be textured, raised into relief, and worked with various tools. It can be engraved, scratched, sanded, and filed. The resulting plastic plate can be printed either as a relief or as an intaglio plate, or even both. It can be printed alone or in combination with other techniques. Thin layers of plastics can easily be placed on top of intaglio plates and printed together.



"Clown with Monkey," monotype by Georges Rouault, 1910.
In the Museum of Modern Art, New York. 57.4 cm × 38.7 cm.
By courtesy of the Museum of Modern Art, New York, gift of Mrs. Sam A. Lewisohn

Collagraphy. Like the metal graphic process, collagraphy is an additive method; the printing surface is built up. It is essentially an intaglio method, but it can be combined with relief printing. The printing surface is created by gluing various materials and textures to a support. Today, with the variety of new material available, the possibilities are limitless.

The support (plate) for collagraphy must be thin and strong. A porous material, such as cardboard, must be treated with a sealer. To build up a tough, durable printing surface, a strong adhesive such as polyvinyl acetate must be used.

Among the materials that can be used for tonal areas are sawdust, sand, carborundum, sandpaper, and ground walnut shells. For specific textures, materials such as tarlatan, laces, and crushed paper can be glued into the adhesive.

After the plate has been constructed, the surface is sealed. The sealer can be either brushed or sprayed on. Plastics are preferred because they are tough and are not dissolved by the solvents generally used to clean the plate.

The printing of collagraphs is essentially the same as for intaglio printing.

Making a plaster print

Plaster print. Good proofs of an intaglio plate can be made by plaster casting. Fine plaster of Paris will pick up the most delicate details and can be particularly attractive if the plate has deeply etched or engraved sections.

To make a plaster print, the plate is inked in the same manner as it would be for normal printing. The inked and wiped plate is placed face up on a glass plate, and a precut wood frame is placed around the plate to contain the plaster. After the plaster is poured, it is allowed to cool and set, after which the plate is gently removed.

PROCESS PRINTS

Process-printing methods are primarily used for commercial reproduction. Today, however, many artists use commercial methods to produce fine art. Silk-screen printing itself began as a commercial process, and today it is one of the most popular techniques in printmaking because its character is well suited for hardedge geometric images. Photomechanical processes are incorporated in the work of many contemporary printmakers.

Linecut. The linecut technique is the simplest and least expensive of all the photoreproductive processes. As it cannot register tone, it is used mostly to reproduce black-and-white line drawings. If tones are needed in a linecut, they are achieved with the use of screens consisting of dots (Ben Day screens). The linecut is similar to the woodcut in that both are used in relief printing.

Linecuts are usually made on zinc plates coated with an emulsion of albumin or gelatin mixed with potassium bichromate. This emulsion hardens on exposure to light. The light passing through the transparent part of the negative hardens the emulsion. The areas of the emulsion that are protected by the black on the negative remain in their soluble state.

The plate is then rolled with greasy ink and soaked in water. The unexposed soft emulsion is washed out by the water. The plate is then dried and dusted with powdered rosin, which adheres to the remaining inked emulsion areas. Heating causes the rosin to melt, forming an acid-resistant coating. The plate can then be etched so that the design stands up in high relief.

Halftone cut or plate. Halftone is more sophisticated than linecut, since it is capable of reproducing fine tonal variations. The subject is photographed first through a glass plate that has fine lines printed on it at right angles. The result is an image broken up into tiny dots corresponding to the openings in the screen. When printed, these dots create the optical illusion of continuous tones. There are great variations in screens from coarse (50 lines per inch) to very fine (175 lines per inch). The selection of the screen is dictated by the paper to be used for printing.

After the photonegative of the image is finished, it is printed on a sensitized copper plate. For halftone work, copper is used because of its ability to record fine details. The procedure of wash out and etch is similar to that used with linecuts.

Rotogravure. To make a gravure plate, a screen is used that is the reverse of the halftone screen, in that the lines are

By courtesy of the Brooklyn Museum, New York



"The Alchemist, No. 2," cellocut by Boris Margo, *c.* 1947
85.3 cm × 58.3 cm.

Function of the gravure screen

opaque squares. When the sensitized plate is exposed to light through the screen, the emulsion on the plate hardens under the lines, but leaves the squares soft. Then the plate is exposed again through a diapositive (a positive transparency) of the subject. This time the soft emulsion squares harden in proportion to the range of grays. In etching, the softest squares are affected by the acid first and the hardest ones last. The result, after the etch, is a plate covered with squares of equal size but varying depth. As the deep squares hold more ink than the shallow ones, the tones in the reproduction are controlled in the same manner as in all intaglio printing methods. The rotogravure plate is inked by an ink-carrying cylinder and wiped by a steel blade that removes all the excess ink from its surface.

Although rotogravure is an intaglio printing process it is printed on dry paper with light pressure and thin ink. Hence, there is hardly any embossing.

Offset lithography. Offset lithography is the application of lithography to commercial mass production. The plate, instead of being a stone, is of specially treated zinc or aluminum, suitable for mounting on a cylinder. The image is photographed on the sensitized lithoplate through a screen. The offset method involves double printing. The image from the plate is printed on another roller covered with a rubber blanket, and from this roller it is transferred to the paper. Because the image is reversed twice, the final print corresponds to the original plate. Since the litho offset ink is thin, to speed up inking and facilitate transfer, the tonal areas lose some of their richness and tend to print gray. Litho offset is often used for colour printing. The colour separation is made photographically.

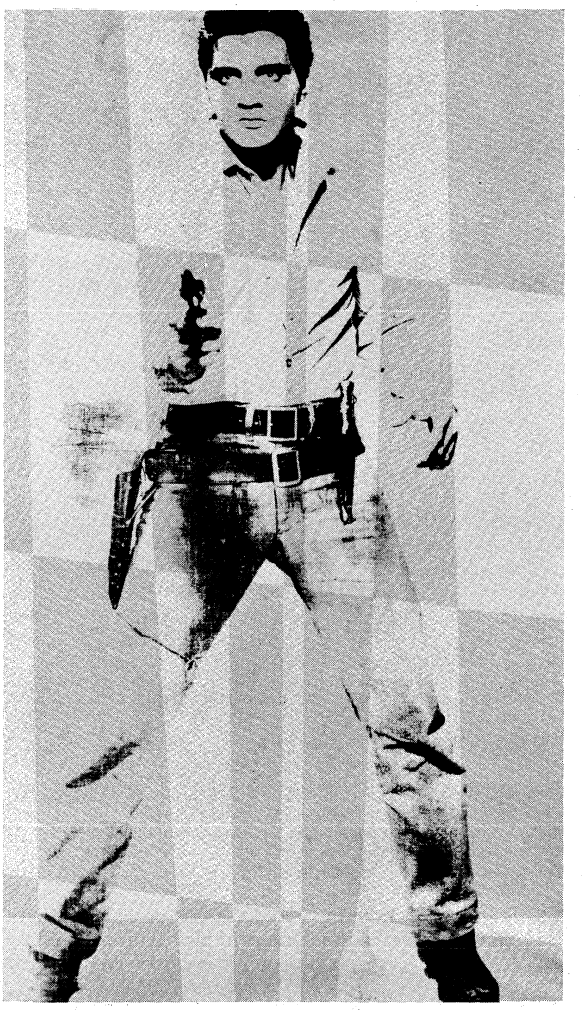
CONTEMPORARY EXPERIMENTATION

One of the most crucial changes in the 20th century involved the size of the print. All through its history, with few exceptions, the print was considered an intimate art form, enjoyed by the few. The change started with the Lautrec posters: the print started to grow until it became mural size. As the dimensions of the print changed, so did its character. It became increasingly bolder and more colourful. Today, the print often competes with painting, a situation deplored by many people who feel that in the process the print is losing its particular character and beauty. A result is that major print shows tend to exhibit a limited number of small, delicate prints.

Next to the size of the print, the greatest change has been in the technology of colour printing. In this area, techniques have become so varied that practically any effect is possible. This development has contributed to the vitality of printmaking, because it has encouraged the participation of colour-oriented artists. The combining of

Combining media

various media is closely related to the experimentation in colour printing. Each medium has its own capabilities and limitations; combined, the media often complement each other. It is now common to see three or four different techniques combined.



Use of photography shown in "Elvis Presley," silk-screen print on canvas by Andy Warhol, 1964. 2.08 m × 1.2 m.
By courtesy of the Leo Castelli Gallery, New York

Another area of experimentation is in three-dimensional surfaces. The trend started with embossing, and today artists are creating completely three-dimensional printed objects.

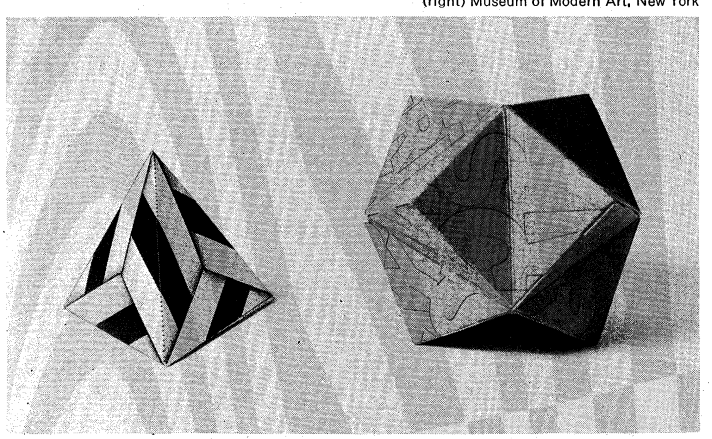
The shaped print is a printed paper sculpture, made by cutting and folding the printed paper, or by assembling precut printed surfaces. Many of these surfaces are metal or plastic objects with printing, usually done by the silkscreen process.

Instead of using rectangular painting surfaces, many painters now work on shaped canvasses. In the same way, printmakers, instead of using rectangular plates, are using many different shapes. Printing with movable plates, which became particularly popular in intaglio colour printing and with colour woodcuts, was the logical extension of this freedom. In this method cut out small plates are placed on top of larger plates and printed together, or they can be assembled on a cardboard support and printed. This procedure facilitates the use of many colours and also offers great freedom in composing.

Photography is profoundly affecting printmaking. Photographic methods can be combined with intaglio, lithography, or silk screen to enrich their vocabulary. The possibilities are nearly limitless. Yet photography can be corrupting when it reintroduces reproductive ideas, and, unfortunately, it is often used for this effect.

Kinetic (moving) art, such as the mobile, is a major contemporary preoccupation in painting and sculpture. At the present there are few attempts in this direction in printmaking, but there will probably be more. The prob-

By courtesy of (left) theARTgallery Magazine, (right) Museum of Modern Art, New York



Three-dimensional prints.
(Left) "Cheops' Prize," offset-printed paper sculpture by Omar Rayo, 1970. Side 8.9 cm. (Right) "Polyhedron," offset-printed paper Christmas card by Paul Rogers Harris, 1968.

lem in such works is how to combine the print with motion without destroying its very nature.

MOUNTING AND CARE OF PRINTS

Very few people know how to display prints and how to take care of them properly. It is heartbreaking to see a great master's print glued to a cheap cardboard or the border of a fine print ruined with tape.

Because paper, particularly old paper, is fragile, it should be handled as little as possible and never picked up with one hand since this might put too much stress on the paper and tear it. To protect it, a print should be mounted as soon as possible.

The surface of the print, especially an intaglio print, is delicate, and rubbing might permanently injure it. Prints should not be stacked without protective layers of tissue paper between them. Wood-pulp papers should not be used, as the acid content in these might burn the print. A print should not be exposed to intense sunlight; this is true particularly of colour prints, for very few colours are stable enough to withstand long exposure to direct sun. Light can also affect the paper. Because wood-pulp board contains chemicals that in time could burn or discolour the paper, a permanent mat should be constructed out of pure rag board. A properly constructed mat consists of two parts: the backing board to support the print and the covering frame to display it. The width of the mat frame should be related to the print's dimensions so that the mat does not overpower the image. The window size of the mat should never obscure the printed image itself, or the signature and edition number.

Temperature changes in a damp climate might cause condensation and the print could develop fungi, so prints should be kept from direct contact with the glass. The simplest protection is a deep enough pure rag mat. If this is not sufficient, a filler should be inserted into the frame to increase the space between the mat and the glass.

The backboard of the frame should be also rag board, or at least faced with rag paper, although the latter is not the perfect solution. The back of the frame should be sealed with tape to prevent the penetrating of dust. In damp climates it is advisable to keep the frame away from the wall by placing corks on its four corners. This facilitates the free circulation of air. Air conditioning and humidity controls are the best protection.

II. History of printmaking

Engraving is one of the oldest art forms of man. Engraved designs have been found on prehistoric bones, stones, and cave walls. The technique of duplicating images goes back several thousand years to the Sumerians (*c.* 3000 BC), who engraved designs and cuneiform inscriptions on cylinder seals (usually made of stone), which, when rolled over soft clay tablets, left relief impressions. They not only conceived the idea of multiplication but also the mechanical principle, the roller, which in more sophisticated form became the printing press.

On the basis of stone designs and seals found in China, there is speculation that the Chinese may have produced a primitive form of print—the rubbing—about the 2nd century AD. The first authenticated prints rubbed from wood blocks were Buddhist charms printed in Japan and distributed between AD 764–770. It is believed that the first wood-block prints on textiles were made by the Egyptians in the 6th or 7th century; but the earliest printed image with an authenticated date is a scroll of the *Diamond Sūtra* (one of the discourses of the Buddha) printed by Wang Chieh in AD 868, which was found in a cave in eastern Turkestan.

In Europe, stamping (to imprint royal seals and signatures) preceded printing by rubbing or with a press. The earliest documented impressed royal signature is that of Henry VI of England, dated 1436.

Textile printing, however, was known in Europe in the 6th century, the designs consisting largely of repeated decorative patterns. Printing on paper developed from textile printing, following the introduction of paper from the Orient. The first European paper was made in 1151, at Xativa (modern Játiva), Spain. Soon afterward paper manufacturing began in France and then in Germany and Italy, notably by Fabriano, whose enterprise was established in 1276.

The first woodcuts on paper, printed in quantity, were playing cards. The term *Kartenmahler* or *Kartenmacher* ("painter or maker of playing cards," respectively) appears on a German document dated 1402; and documents from both Italy and France from the middle of the 15th century mention wood blocks for the printing of playing cards. The earliest dated woodcut is a "Madonna with Four Virgin Saints in a Garden" from the year 1418. Woodcuts of playing cards

Many documents from the 15th century indicate that a clear distinction was made between the designer and the cutter of the wood blocks. From the outset, woodcut was primarily a facsimile process: the cutter copied a drawing provided by the designer.

Printing from a metal engraving, introduced a few decades after the woodcut, had an independent development. The art of engraving and etching originated with goldsmiths and armour makers—men who were thoroughly professional craftsmen, practicing an art that had a long, respected tradition. Since the armour makers and goldsmiths were designers themselves, the whole process was controlled by the creative artist.

PRINTMAKING IN THE 15TH CENTURY

Germany. Single prints (in contrast to those printed in a series or as part of an illustrated book) of the early 15th century were not signed or dated, and, because they were religious images carried by pilgrims from one place to another, it is nearly impossible to establish with certainty their place of origin. Their style alone must be relied upon for some indication of origin.

The first phase of woodcut, from about 1402 until about 1425, was dominated by boldly designed single figures against a blank background. Most of the cuts were made to be hand coloured. In the second half of the 15th century the cuts became more complex: architectural and landscape elements came into use, and often the image was framed in an elaborate border.

The first metal prints (criblé, or dotted, print) were made in the second half of the 15th century. The design was created by tiny dots punched into the metal and intermingled with short cuts. Surface printed, the whites are the positive part of the design, which is dominated by the dark background. Tiny holes in the borders indicate that most of these plates were intended as decorations to be mounted rather than as printing plates.

The earliest dated intaglio-printed engraving is from 1446: "The Flagellation," of a Passion series. Around this time, the first distinct personality to have great influence on German engraving appeared. He is known as the Master of the Playing Cards. His style was simple, nearly monumental; unlike the printwork of goldsmiths, his engravings lack ornamentation. For shading he used slightly diagonal parallel cuts. The Master of the Playing Cards heralds the beginning of a century of great printmakers in Germany. Another significant engraver, the Master of the Banderoles, was named after the ribbon scrolls characteristic of his prints, which are more decorative than those of the Master of the Playing Cards.

In the second half of the 15th century, the outstanding printmaker was Master E.S., who flourished about 1440–67 and was one of the first to use initials as a signature on his plates. Little is known about him, but the personality that emerges from approximately 317 plates is forceful and distinct. Although it is evident from his prints that, like most early engravers, he was first trained as a goldsmith, his work has strong pictorial quality. Engravings of Master E.S.

Martin Schongauer was the first great engraver who is known to have been a painter rather than a goldsmith. Although Schongauer's style was still Gothic in character, he composed with much greater freedom than his contemporaries, thus representing a transition into the Renaissance. He made about 115 plates, mostly of religious subjects, and was a powerful influence on the young Albrecht Dürer (see below *Printmaking in the 16th century*). During the second half of the 15th century, a group of brilliant engravers known only by their initials



"Blumen-Dame" ("Cyclamen Queen"), engraving by the Master of the Playing Cards, *c.* 1440. 13.9 cm × 9.2 cm.
Deutsche Fotothek Dresden

emerged in Germany. They are the Masters B.G., B.M., L.G.S., A.G., B.R., and W.H. The controversial figure of Israhel van Meckenem appeared at the end of the 15th century. A superb and extremely prolific engraver, he was a rather eclectic artist, borrowing from other masters and often copying them.

Italy. In the 15th century, Italian printmaking was dominated by the northern cities: Florence, Venice, and Milan. Throughout the century, printmaking was mainly concerned with playing cards and book illustrations, with a few single prints appearing in the second half of the century. While in Germany and the Netherlands the art was completely dominated by devotional, religious subject matter, Italian printmaking covered a relatively broad range. The awakening Renaissance attitude made the artists much more receptive to purely aesthetic, decorative, sensuous experience. In addition to religious subject matter, Italian prints included mythology, pure ornamentation, and some of the finest early portrait engravings.

Giorgio Vasari, the chronicler of the Renaissance, credited the Florentine goldsmith Maso Finiguerra with the invention of printed engraving, but present knowledge indicates that, at the same period in Germany and the Netherlands, printmaking was in a more advanced stage. In spite of the fact that book printing was originally introduced from the northern countries into Italy, engraving remained a national, regional development, free of strong foreign influence until the beginning of the 16th century.

Broad manner and fine manner engraving

Two methods of engraving were practiced in Italy, the broad manner and the fine manner. The fine manner, associated with the Finiguerra school, is characterized by closely cut and extremely fine lines combined with crosshatching intermingled at times with dots. The broad manner is less dense, and forms are modelled using diagonally cut parallel lines, interlaid at times with short cuts or dots. In shading, the spacing between the lines is wider than in the fine manner and there is no crosshatching.

Finiguerra himself was not an important artist. His significance lies in his influence on Antonio Pollaiuolo, a Florentine painter, sculptor, and architect whose reputation as one of the most distinguished engravers of the 15th century is based on his one authenticated print, "The Battle of the Nudes" (*c.* 1465)—a powerful image, beautifully engraved in the broad manner.

While Pollaiuolo worked in Florence, Andrea Mantegna, a great painter and certainly the most eminent Italian printmaker, lived and worked in Mantua. Mantegna produced approximately 20 plates (only seven of which are completely authenticated), all line engravings in the broad manner. A superb draftsman and a virtuoso engraver, Mantegna could achieve, in spite of the limitations of his method, an incredible range of colour in his prints, a quality lacking in the work of most of his followers.

In addition to the masters, talented engravers included Cristofano Robetta, a Florentine who made some rich, intricate engravings in the fine manner; and the Venetian Jacopo de' Barbari, who travelled in Germany and whose refined engravings show the influence of Albrecht Dürer.

Other countries. *The Netherlands and Burgundy.* The first half of the 15th century in the Netherlands and Burgundy was dominated by woodcut book illustrations. Although no single prints of great importance were produced, beautiful books were published. Antwerp and Delft were the main printing centres.

Parallel with, if not even a little earlier than, the emergence of distinguished printmakers in mid-15th-century Germany, a group of great engravers emerged in the Netherlands and neighbouring Burgundy. Superb artists, they are identified only by the subject of their most characteristic work: the Master of the Death of Mary, Master of the Gardens of Love, and Master of the Mount of Calvary.

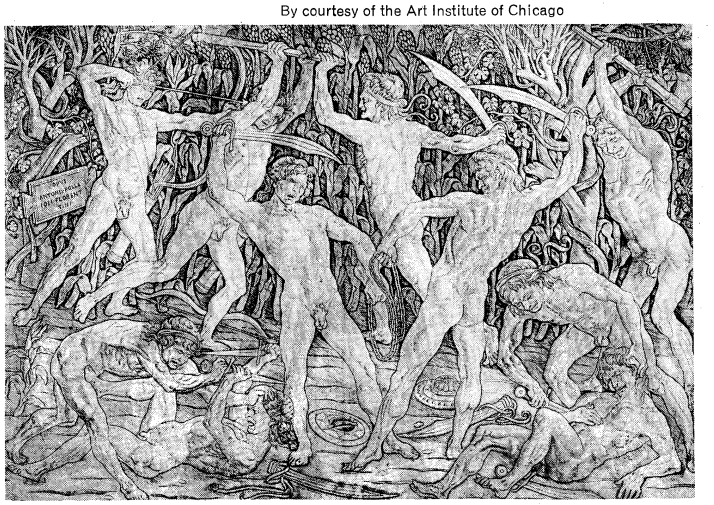
Toward the end of the century, the Netherlands produced a brilliant artist—rivalling Master E.S.—known as the Master of the Amsterdam Cabinet or Master of the Hausbuch, who worked between 1450 and 1470. He is known by this name because the finest collection of his extremely rare prints is in the Print Cabinet of the Rijksmuseum. His prints are painterly and almost expressionist in power. His role in the technical development of printmaking is also significant, as he was the first major artist to make drypoints.

Dominance of book illustrations in France

France. In France, book illustrations dominated printmaking throughout the century. Paris, the cultural centre, led in book publishing, although other prosperous cities, such as Lyon, produced many illustrated books. The publications printed by Dupré and Pierre Le Rouge are the glories of French medieval graphic art.

Strangely enough, there was little engraving of importance. Most of the French engravings of this period were either rather crude, provincial illustrations or playing cards.

By courtesy of the Art Institute of Chicago



"The Battle of the Nudes," engraving by Antonio Pollaiuolo (*c.* 1431/32–98). About 40 cm × 61 cm.

PRINTMAKING IN THE 16TH CENTURY

Germany. Albrecht Dürer was the master of 16th-century German graphic arts. One of the towering figures in the history of printmaking, he was a complex, truly Renaissance man, interested in philosophy and science as well as art. He was one of the first to break the provincial isolation of Germany by travelling to Italy, where he learned from the Italians and in turn influenced them.

Dürer's subject matter mirrors his thoroughly European intellectual orientation. His prints deal with religion, history, mythology, and folklore. He is also one of the first great portrait engravers.

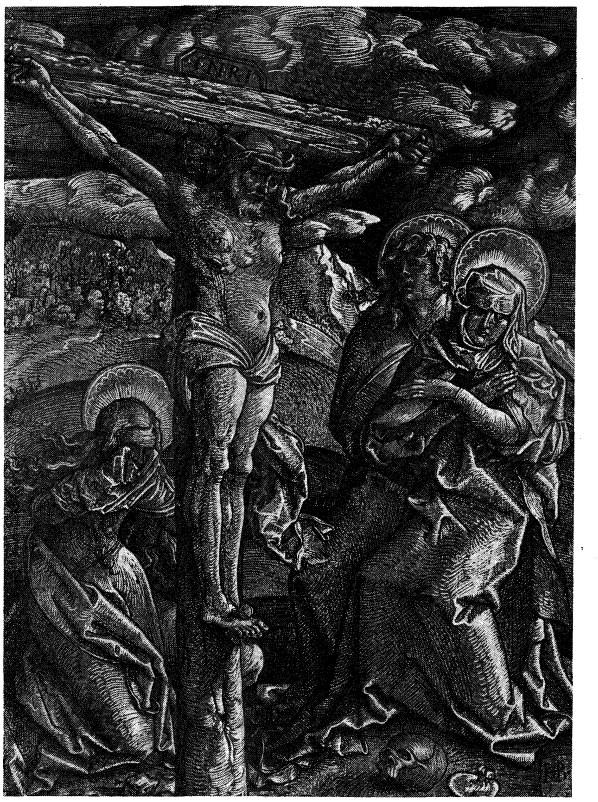
Dürer was one of the supreme draftsmen of all time and an artist of enormous imagination and sensibility. As a technician he raised the art of engraving to a height it never reached again. As an experimenter, he produced, in addition to his engravings and woodcuts, etchings and drypoints. His best works are metal engravings, which he cut himself. His woodcuts are perfect reproductions of his superb drawings.

Hans Baldung-Grien, another great German printmaker and one of the most original artists of his time, worked with Dürer. In his images of witchcraft and magic, Baldung expressed the medieval mysticism that lingered in the German Renaissance. Besides his black and white work, he produced fine chiaroscuro woodcuts in which light and shadow are produced by using different wood blocks for different tones of the same colour.

Lucas Cranach the Elder was a typical representative of the German Renaissance—much less affected by the Italian Renaissance influence than Dürer. He had a vivid imagination and an earthy imagery that was full of vitality. He made many woodcuts and relatively few metal engravings.

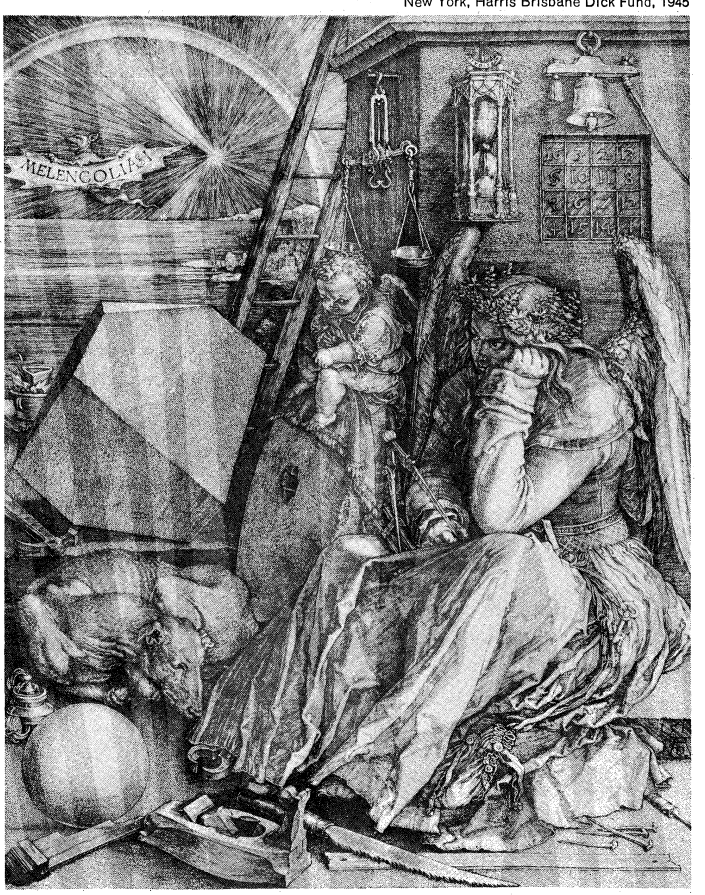
Landscape etchings of Altdorfer

Albrecht Altdorfer was one of the first great landscape artists and one of the first to make landscape etching and woodcuts for their own sake, rather than as backgrounds for figures. Under his influence, two other artists made fine landscape prints: Augustin Hirschvogel and Hans Sebald Lautensack. The spontaneity and directness of their work foreshadows the lyrical landscapes of the 18th century.



"The Crucifixion," chiaroscuro woodcut by Hans Baldung-Grien (1484–1545). 36.8 cm × 25.8 cm.
By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection

By courtesy of the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1945



"Melencolia I," engraving by Albrecht Dürer (1471–1528). 24.1 cm × 20 cm.

The 16th century also included the artists referred to as Little Masters, so called because of the size of their prints, not their stature as artists. They were Barthel Beham, Hans Beham, Georg Pencz, and Heinrich Aldegrever.

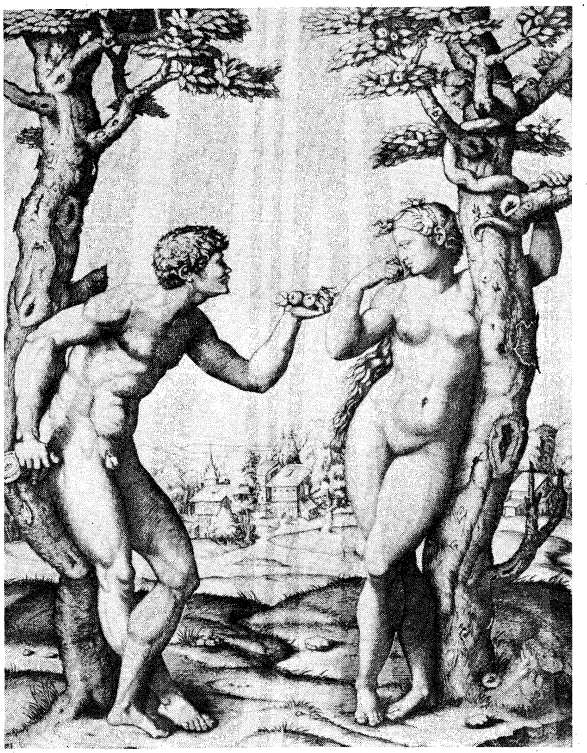
Other countries. *The Netherlands and Flanders.* The outstanding Dutch printmaker of the period was Lucas van Leyden (1489/94–1533). If the latter birth date is correct, at the age of 14 he was already an accomplished engraver. In maturity, he was a superb engraver, in many respects rivalling Dürer. Besides his metal engravings, which are characterized by a very delicate touch, van Leyden designed many wood blocks and also made a few etchings.

A virtuoso of the burin, the Flemish engraver Hendrik Goltzius (1558–1617) developed an incredible variety of cuts and textures to imitate the surface qualities of materials. Other printmakers of the period include Allaert Claesz and Cornelis Matsys.

Italy. After the death of Mantegna in 1506, Italian printmaking of the 16th century was dominated by lesser figures. During the 16th century the few etchings produced in Italy have only historical interest.

The most influential engraver of the century was Marcantonio Raimondi. Under the influence of Dürer, Raimondi became a virtuoso engraver; technically Dürer's equal, he lacked his master's originality. Raimondi eventually became the engraver of Raphael, organizing a workshop dedicated primarily to making reproductions of his work. Thus, Raimondi won the dubious honour of being the first of the many printmakers who ultimately were influential in turning the art of engraving into mere reproduction. He was followed by a whole generation of competent engravers devoted solely to reproduction.

One of the exceptions was Giorgio Ghisi of Mantua, who in his isolated regional development escaped the corrupting influence of Rome. His 1550 visit to Antwerp made Ghisi an important link between Italian and northern engraving.



"Adam and Eve," engraving by Marcantonio Raimondi (*c.* 1480–*c.* 1534). 23.8 cm × 17.6 cm.

By courtesy of the Metropolitan Museum of Art, New York, Joseph Pulitzer Bequest, 1917

Engravings of Duvet

France. The only major figure in 16th-century French engraving is Jean Duvet, whose predilection for excessive ornamentation indicates that he was trained as a goldsmith. Although Duvet's style was influenced by Mantegna, his imagery was completely original. His greatest work, "The Apocalypse," reveals a feverish, mystical imagination.

Apart from the work of Duvet, ornamental engraving was the most significant achievement of 16th-century French printmaking. Although these elegant engravings cannot be ranked with the work of the great masters, they represent a genuine expression of the French spirit. The outstanding figure of this school was Étienne Delaune. Although his motifs were influenced by those employed by Raphael for his fresco wall paintings in the Vatican, Delaune nonetheless achieved a personal style.

Trends in the late 16th century. By the second half of the 16th century, the quality of printmaking, particularly engraving, had gone into a severe decline. Masters like Dürer and Mantegna were replaced by skilled craftsmen. The trend toward reproduction that had begun with Raimondi gained ground, sapping the vitality of engraving. Yet at the same time, the quantity of production increased. Except for the modern era, this was probably the most prolific period of printmaking. Since it was the beginning of the age of travel, discovery, and religious upheaval, the demand for maps, religious pictures, illustrations, and portraits was enormous.

One after the other, print publishing houses opened all over Europe. Dutch and Flemish families dominated the new profession: in the Netherlands, the firms Cock, Galle, and Passe; in Augsburg, Dominicus Custos; in Antwerp, Brussels, Prague, and Venice, the Sadeler family. In Italy, Antonio Salamanca cornered the market and flooded it with bad reprints of Raimondi engravings.

The publishers of this period usually bought the original plates outright from the artist and issued prints on demand in unlimited quantities. If the plates wore out, they were reworked in the publishers' own workshops, a practice that was responsible for the destruction of many fine artworks. In many cases, it is no longer possible to identify the creator of the original plate.

Map engravers

In this period, map engravers were particularly important: the maps of Abraham Ortelius were engraved by Franz Hogenberg; Gerardus Mercator engraved his own map designs; and Jodocus Hondius bought the Mercator plates after their use in the edition of 1596 and introduced them in England, along with some of his own work.

PRINTMAKING IN THE 17TH CENTURY

Portrait engraving. *France.* The end of the 16th century and the beginning of the 17th were dominated by ornamental engravers and illustrators, working under Flemish influence; by the middle of the 17th century, however, a distinctly French school of portrait engraving had emerged. Although this school did not produce a major master, it represents a significant phase of European printmaking.

Michael Lesne, a French portraitist whose influence was considerable, worked for a time in the Rubens workshop, later returning to France. Claude Mellan, another major influence, was trained in Rome. Technical virtuosity dominated his prints; for example, the modelling of a face with one continuous spiral.

A superb engraver and a fine draftsman, Robert Nanteuil is considered the undisputed master of French portrait engraving. His style is simple, elegant, and free of the mannerism characteristic of his contemporaries. He and his two rivals—Gerard Edelinck, who was born in Antwerp but studied and developed his style in France, and Antoine Masson, who engraved portraits in the grand style—represented the dominant forces in 17th-century French portrait engraving.

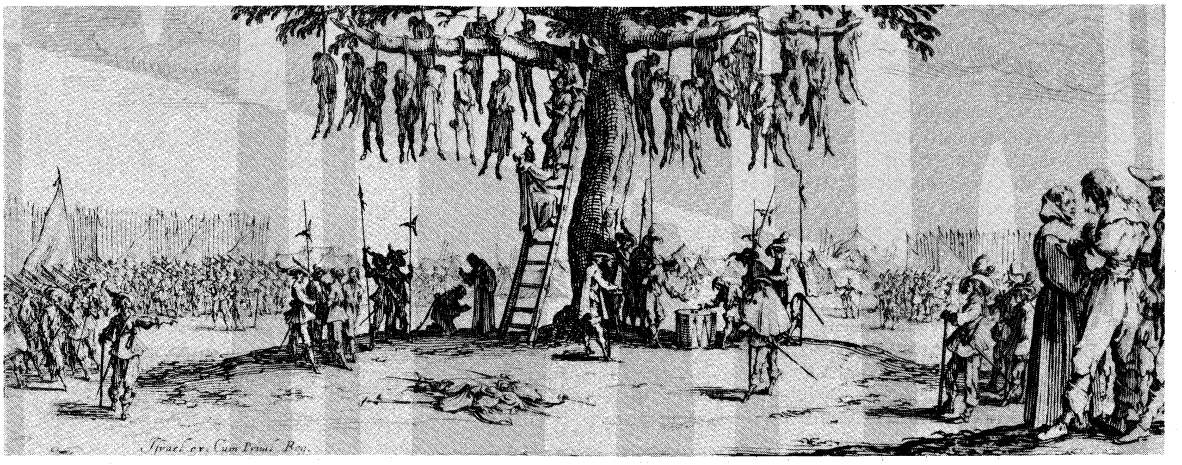
Germany. After the glories of the 15th and 16th centuries, German graphic genius was dormant for nearly three centuries. Historically, Ludwig von Siegen, a minor painter and medalist, is important for his invention of the mezzotint printing method. But the perfecting of this tonal technique increased the reproductive facility of printmaking, thus contributing to the decline of artistic creativity.

Like France, Germany produced a school of portrait engravers; but, although competent technicians, they failed to develop a distinctly national school comparable to the French. Of this group, two are significant: Jeremias Falck, a student of Hondius, and Bartholomäus Kilian, who studied in Paris and later introduced French influence into German printmaking.

By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection



"Marin Cureau de la Chambre," engraving by Robert Nanteuil (1623–78). 25 cm × 19.4 cm.



"The Hangman's Tree," etching by Jacques Callot from the series "The Miseries and Disasters of War," 1633. 6.6 cm × 19 cm.
By courtesy of the trustees of the British Museum; J.R. Freeman & Co. Inc.

The Netherlands. Portrait engraving in Holland was on a higher level than in Germany. Cornelis van Dalen was a fine engraver who emigrated to England and died there. More gifted than his father, Cornelis van Dalen II was an artist of considerable stature, who engraved some of the most powerful portraits of his time.

Abraham Blooteling, a pupil of van Dalen II, was also a fine portrait engraver. His major contribution, however, was in the development of the new technique of mezzotint—specifically, the invention of the rocker, the tool used in the technique. He also introduced the mezzotint into England, where it was adopted with such success that it later became known as the "English Manner."

The development of mezzotint

England. In the 17th century, English printmaking produced a portrait engraver of considerable stature, William Faithorne. He studied in France and initially was under the influence of Mellan and Nanteuil; in his late work, however, he developed a style independent of theirs. Faithorne was England's only major native printmaker during this period, when most prints were reproductive engravings. By the end of the century, engraving was in total decline, replaced by the fashionable mezzotint.

Flemish printmaking. One of the dominant figures of European art was Peter Paul Rubens, a painter, diplomat, and businessman. Quickly recognizing the commercial potential of printmaking, he organized a graphic workshop where, under his supervision, reproductions of his work were produced. Only one etching, "St. Catherine," is considered as his own. The quality of this one print indicates how great was the loss to the art of printmaking that this great draftsman did not make more original etchings.

Rubens' pupil Anthony Van Dyck was one of the most distinguished portrait painters of his time. At age 27 he undertook a very ambitious project: the etched portraits of the 100 most famous men of his day. For this set of prints, known as the "Iconography," he completed 18 portraits. But only five of these ("Peter Breughel the Younger," "Snellinx," "Erasmus," "Suttermans," and "Josse de Momper") remained unchanged; another five were retouched by professional engravers, and the rest were completely reworked by them.

Etching. Like the Van Dyck portraits, nearly all of the outstanding prints produced in the 17th century were etchings. Etching emerged as the dominant technique for many reasons. The fact that engraving had become a completely commercialized, reproductive method and that mezzotint had never been anything else alienated many artists. As an unexploited and relatively unexplored medium, etching intrigued the experimentally oriented. Furthermore, the fluid, flexible technique of etching was a lure for the creative painter, whose own medium had become freer and more spontaneous.

Italy. At the beginning of the 17th century, there was more etching in Italy than in any other European country. Strangely enough, probably the three most important etchers—Jacques Callot, Claude Lorrain, and José de Ribera—were foreign-born.

The Bolognese school was formed around Guido Reni, whose delicate etching style of light lines and dots became a standard technique for most Italian etchers of his time. His school, however, produced no superior printmakers.

The Spanish painter José de Ribera was the dominant figure of the Neapolitan school. Though he was the first major realist painter in Italy and a strong influence against the idealizing trend, both his paintings and his etchings were outside the mainstream of Italian art.

Next to Ribera, Salvator Rosa, an Italian, was the most notable artist of the Neapolitan school, producing a large number of etchings that are full of charm but of no great importance.

Born in Nancy, France, Jacques Callot ran away from home as a boy to study art in Italy. Of all the artists engaged in 17th-century Italian printmaking, he was historically the most significant; for he was one of the first to use repeated bitings on his plates to achieve tonal variations. His drawing style represented a transition between engraving and etching: using a specially shaped etching needle of his own invention, he imitated the swelling and tapering characteristics of the engraved line. His illustrations record and ironically comment upon the customs, historical events, and morals of his time. Callot's work was often decorative and manneristic; but, at his best, as in the series "The Miseries and Disasters of War" (1633), he transcended mere illustration and achieved powerful images of universal significance.

Callot's etching technique

Claude Lorrain, also French-born, was one of the finest landscape painters in Italy, and he had an intuitive understanding of the etching medium. His spontaneous interpretation of the atmospheric quality of his subject foreshadows the Barbizon school and Impressionism in the 19th century.

The Netherlands. In the beginning of the 17th century, Holland suddenly exploded into a frenzy of creativity in etching. The sensitive, atmospheric etchings of the brothers Esaias and Jan van de Velde can be considered the beginning of the Dutch landscape school. Others were Adriaen van Stalbent, Pieter de Molijn, and Willem Buytewech—all fine printmakers, but all eventually overshadowed by the dramatic personality of Rembrandt. Before him, however, another artist appeared who was so original that no historical precedent could anticipate him. Hercules Seghers is one of the most interesting and mysterious figures in the history of printmaking. He was a lonely, tragic man, an experimenter who was so far ahead of his time that it took centuries for the real significance of his work to become apparent. It is known that Rembrandt appreciated his work. He owned some of Seghers' prints and even reworked one of his plates, "Tobias and the Angel" (*c.* 1633); keeping the landscape, he changed the figures, making it "The Flight into Egypt" (1653).

Seghers was the first real experimenter in intaglio colour printing. His methods were completely unorthodox: he printed on tinted canvas, tried light lines on a dark background, and also mixed printing with hand colouring. Seghers' etching technique was itself very unorthodox. His eroded lines, so well suited to his subject matter, are unlike any etched line made before him, which has led some experts to the conclusion that Seghers invented the lift ground, an aquatint technique.

Most of Seghers' etchings represent craggy, arid landscapes. Everything he drew—landscapes, still lifes, even figures—seem to be made of stone. It is a world suspended in the timelessness of death.

The graphic distinction of Rembrandt

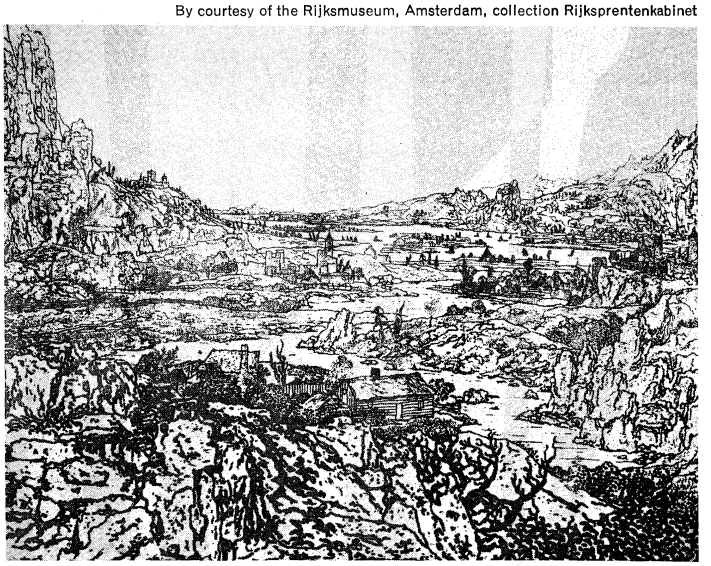
Even among the supreme artists of the world, Rembrandt van Rijn occupies a very special place. One of the most eminent painters of all time, he also left a graphic oeuvre of heroic proportions both in quantity and quality. A great innovator, he was the first artist to fully explore the possibilities of the etched line.

Rembrandt made approximately 300 plates. His subject matter represents practically every aspect of human existence: he rendered religious and historical subjects; he explored themes of love and death; and he created profound portraits and sensitive landscapes. Everything that was part of life concerned him, from the highest ideals to the most mundane bodily functions.

While Rembrandt's early prints are pure etchings, his later works frequently combine etching and drypoint. Since he often reworked his plates between printings, there are sometimes enormous variations between proofs. Rembrandt's immediate influence on his students and followers was not very productive, for his personality was so overpowering that those close to him fell under his spell and simply imitated his style. Artists like J.G. van Vliet, who copied and reworked many of the master's plates, and Jan Lievens were mere shadows of Rembrandt. Of the group closely associated with Rembrandt, probably Ferdinand Bol was the strongest, but even he is dwarfed by the inevitable comparison.

Also active during Rembrandt's time and somewhat overshadowed by him was Adriaen van Ostade, one of the most gifted of Dutch genre painters. The subjects of both his paintings and prints were taken mainly from the daily lives of simple people, usually peasants. In spirit, his work represented an important departure from the heroic orientation of historic and religious painting, reflecting a crucial social change—the emergence of a middle class in Europe. For the first time, common people replaced the clergy and the nobility as a source of inspiration for an artist. Van Ostade's etching technique was influenced by the early Rembrandt, but his drawing style was personal. It was simple, undramatic, and direct—well suited to his intimate subject matter.

By courtesy of the Rijksmuseum, Amsterdam, collection Rijksprentenkabinet



'The Rocky River Landscape," 17th-century etching by Hercules Seghers. 17.6 cm × 21.7 cm.



"Tamatori" ("Getting the Pebble"), colour woodcut by Hishikawa Moronobu, 17th century. 30.5 cm × 40 cm;
By courtesy of the Sakai Collection, Tokyo

Throughout the 17th century, landscape painting and etching thrived in Holland. Jan van Goyen and Roelant Roghman both made fine landscape paintings and etchings. In this group the most interesting figure is Jacob van Ruisdael, whose sensitive, luminous landscape etchings foreshadowed the Barbizon school.

Toward the end of the century, a strong Italian influence invaded Holland. Since the earthiness of the Dutch temperament did not mix well with the Italian tendency toward idealization, the result was an eclecticism that drained Dutch art of much of its vitality.

Japanese Ukiyo-e prints. Until the 17th century, Japanese painting was completely dominated by Chinese influence. The Japanese silk paintings and screens of idealized landscapes were hardly distinguishable from their Chinese counterparts. Then, in the early 17th century, an artist of aristocratic origin, Iwasa Matabei, started to paint images related to his environment and personal experience. Although this era of Japanese art history is rather obscure, he is credited with being one of the founders (along with Iwasa Matabei II and Iwasa Matabei of Otsu) of Ukiyo-e, whose woodcuts of the transient world or the world of everyday life represented a drastic break with the classical tradition. Of the three artists Matabei of Otsu was the most original and had the strongest influence on the development of Japanese printmaking. By standards of Western taste, the images the Ukiyo-e school produced are highly stylized and thoroughly refined. Cultured Japanese, however, found them shockingly vulgar. The very fact that ordinary landscapes and the daily life of common people, actors, and courtesans were the inspiration for the Ukiyo-e artist represented a startling departure from tradition. Just as the emerging middle class revolutionized taste in Europe, the prosperous city dwellers of Edo, Kyōto, and Ōsaka developed their own aesthetic subculture. The development of the popular Kabuki theatre, as distinct from the aristocratic Nō drama, parallels the blossoming of Japanese printmaking.

Foundation of the Ukiyo-e school

The first great master of Japanese printmaking was Hishikawa Moronobu. A creative innovator, he was the first to use street scenes, peddlers, and crowds as his subject matter and to make his prints available to the common people. As a result, he was looked upon by many as the inventor of printmaking. He illustrated more than 100 books, mirroring the culture and customs of his time. Moronobu's style was a perfect harmony of rhythm, delicacy, and monumental simplicity, leading the way toward the great flowering of Japanese printmaking in the 18th century.

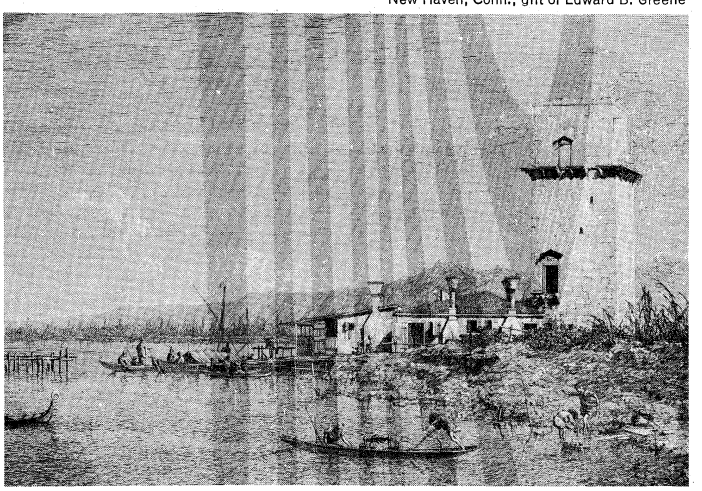
PRINTMAKING IN THE 18TH CENTURY

Italy. In the 18th century, Italy was the most fertile soil in Europe for printmaking. The first outstanding printmaker of the century was the Rococo master Gio-

vanni Battista Tiepolo. His lightly bitten, spontaneous plates reveal superb draftsmanship. With rhythmic, delicate textures, he created a living, luminous space. His 50 plates represent a major contribution to the development of etching—a contribution further enhanced by his influence on Goya (see below *Spain*). Giovanni Domenico Tiepolo, the son of Battista, made more prints than his father but remained under his influence all his life.

One of the most original printmakers of the period, Canaletto (Giovanni Antonio Canal) created lyrical etchings charged with the misty atmosphere of Venice.

By courtesy of the Yale University Art Gallery, New Haven, Conn., gift of Edward B. Greene

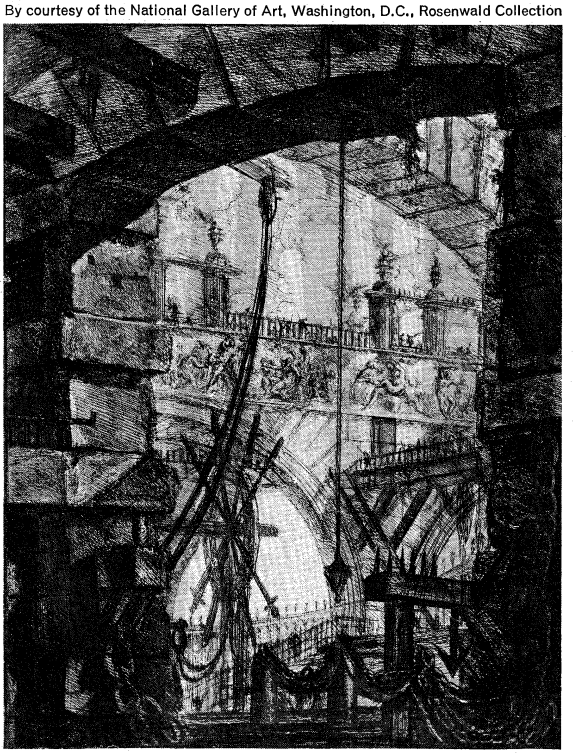


"La Torre di Malghera," etching by Giovanni Antonio Canal (called Canaletto) (1697–1768). 29.8 cm × 43 cm.

Inexhaustible in linear and textural invention, they are perfect examples of the simulation of colour and light by purely graphic means. His nephew and pupil, Bernardo Bellotto (Canaletto), who assumed his name, was a prolific printmaker but remained under his uncle's influence.

Giambattista Piranesi was the greatest architectural printmaker of his time and probably of all time. Trained as an architect, he was passionately interested in Roman antiquities. Of the approximately 3,000 large etchings completed by Piranesi, all are brilliant, and many rise above documentation. His most important work is the *Carceri d'invenzione* (imaginary prison scenes). The plates, which were made in his youth (published *c.* 1750), are personal, rich, and evocative, far surpassing anything he created after them.

By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection



Carceri d'invenzione ("Prisons"), plate 4, etching by Giambattista Piranesi, 1750(?). 55 cm × 41 cm.

England. Until the 18th century, English printmaking was dominated by foreign influences. William Hogarth, the first major English printmaker, created not only a personal style but a national school. He was a gifted pictorial satirist, belonging in some respects to the tradition of Callot and Goya. He is, however, more earthy than Callot and lacks the savagery of Goya. Hogarth was a printmaker of the people, whose work was so popular that to protect it from imitators he instigated the first engraving copyright act of 1735. Although his drawing was rather pedestrian, Hogarth's prints reveal a sharp observation, projected with robust vitality.

Pictorial satire of Hogarth

Next to Hogarth, Thomas Rowlandson is the most significant representative of English satire. A brilliant draftsman and a deft caricaturist, he spoofed the moral and social life of England with great humour.

William Blake is by far the most interesting figure in English printmaking. Poet and experimental printmaker, he was a visionary, creating his works totally outside the mainstream of art history. His printed work consisted primarily of book illustrations. Original and inventive in technique, he used a great range of media, from wood and metal engraving to relief etching. In the latter he devised a transfer method that enabled him to etch the text and the illustration on the same plate. Many of his prints were hand coloured, but he also printed colour by an offset method of his own invention.

Thomas Bewick was a provincial illustrator who made a great number of charming wood engravings, primarily of animals and rural genre scenes. He was a pioneer in the technique of wood engraving, introducing tonal variations by slightly varying the level of his blocks.

To reproduce the fashionable paintings of the day, commercial engravers perfected a whole arsenal of reproduction techniques, such as mezzotint, stipple engraving and etching, and crayon manner.

Spain. Spanish printmaking in the 17th century had been dominated by Flemish and French influences, and no printmaker of importance emerged during the period.

In the 18th-century artist Francisco de Goya, Spain had not only its first truly great printmaker but also the only printmaker whose etchings rival Rembrandt's. Moreover, he is the most eminent satirist printmaking has produced. His visual comments on human folly, war, and religious persecution are devastating.

Goya created four major cycles of prints. The first, "Los caprichos" (1797–98), consists of 80 enigmatic prints commenting on all phases of life. In 1810 he began the 83 plates of "Los desastres de la guerra," a strong visual protest against the brutality of war. After this came "La tauromaquia" (1815–16), a brilliant series on bullfighting. The last important series was "Los disparates," or "Proverbs" (*c.* 1820–24), a biting, though often humorous, interpretation of human folly.

Technically, the Goya etchings are simple and direct. He usually combined line etching with aquatint; his masterful control of the latter, a relatively new technique, has never been surpassed. Toward the end of his life, he also made a few rich, powerful lithographs.

Goya's technique

France. Most 18th-century French etchings were drawings transferred to copper, in which the effects of pencil, pen, or chalk were imitated. Although some distinguished painters, such as Antoine Watteau, made etchings, no prints of importance were produced. Jean-Honoré Fragonard made a few lovely etchings reminiscent of Tiepolo. They have a luminous, transparent quality and express the Rococo spirit but are nevertheless minor works of a major painter. Two artists are notable for technical achievements: Jean-Charles François developed the crayon manner, and Jean-Baptiste Le Prince is credited with the invention of aquatint.

Japan. The first Japanese artist to produce single prints in quantity was Torii Kiyonobu, who specialized in por-



"The Cockpit," engraving by William Hogarth, 1759. 31.7 cm × 38.4 cm.
By courtesy of the Brooklyn Museum, New York, bequest of Samuel E. Haslett

traits of actors and theatre posters. His school, the Torii, dominated printing for the theatre for two centuries. Another imaginative innovator of the early 18th century was Okumura Masanobu, who experimented with inks, embossing, and gold and silver overlays. He also invented the two-colour print and generally standardized colour printing. His studio greatly influenced the evolution of colour woodcut. Suzuki Harunobu, one of the most charming masters of Japanese woodcut, created prints of infinite delicacy and grace. In this respect he is a forerunner and rival of Utamaro. A highly gifted colourist, he was one of the first to exploit the *nishiki-e*, or full-colour print. He was also the first to colour print backgrounds and to use blind embossing extensively to give his prints three-dimensional textures. Katsukawa Shunshō is notable for his austere portraits of actors, which he designed with much strength and intensity. Some of his portraits are among the finest in Japanese printmaking.

By courtesy of the Honolulu Academy of Arts, the James A. Michener Collection



"Two Lovers Under an Umbrella in Snow," woodcut by Suzuki Harunobu (1724–70). 26.7 cm × 19.7 cm.

The period from 1780 to 1790 was dominated by Torii Kiyonaga, whose work represents the Ukiyo-e at its height. He was a great draftsman and designer and could harmonize in his prints the two seemingly contradictory qualities of elegance and power. Kiyonaga was one of the first to experiment with the compositional possibilities of the diptych, triptych, and pentaptych formats. Although he conceived each block as a self-contained unit, they functioned together in harmony. Kitagawa Utamaro can justly be called the supreme poet of Japanese art. Utamaro's prints are the most perfect expression of a tender, loving contemplation of nature, which included not only birds and flowers but women as well. At the age of 50, he was put in jail for an offending print; broken in spirit, he died shortly after his release. During his lifetime he produced over 600 series of books and albums. Toshusai Sharaku is not only one of the most distinguished but also one of the most mysterious figures of Japanese art. Seemingly out of nowhere, his magnificent, powerful portraits of actors suddenly appeared on posters. The boldness of the portraits, verging on caricature; their psychological insight; their richness in colour all represented a daring new attitude. The originality of these prints disturbed the authorities to such an extent that the police prohibited them. In less than two years of working life, Sharaku had produced approximately 145 portraits; then the prodigious flow of work stopped, and he disappeared again.

Late 18th-century Ukiyo-e

PRINTMAKING IN THE 19TH CENTURY

The 19th century was a turbulent period of art, one aesthetic revolution following the other.

France. French domination of 19th-century art is comparable to northern domination of 15th-century printmaking. Few graphic artists of importance worked outside France.

The great French painter Jean-Auguste-Dominique Ingres made only a few etchings, mainly portraits; but, as demonstrated by the lithograph "L'Odalisque" (1825), his draftsmanship was incomparable. Eugène Delacroix left a much more extensive graphic oeuvre: 24 etchings and 131 lithographs. Both in subject matter and style, Delacroix's prints are eloquent expressions of the Romantic spirit. In his tragically short life, Théodore Géricault made a series of powerful lithographs; his horses are considered classics in their genre.

At midcentury, a rebellion against studio painting took place. A group of young landscape painters, most of whom were also printmakers, formed a group that became known as the Barbizon school. The etchings of Charles-François Daubigny, Théodore Rousseau, and Jean-Baptiste-Camille Corot were close to the spirit of the 17th-century Dutch landscapes. Corot made prints whose spontaneity foreshadowed Impressionism; he also experimented with the newly discovered photographic method of *cliché-verre*.

Another member of this group, Jean-François Millet, was concerned particularly with depicting peasant life. His small but monumentally simple etchings are reminiscent of the 17th-century Dutch genre painter van Ostade at his best.

Political satire of Daumier

Honoré Daumier, one of the foremost political satirists of printmaking, was associated with the Barbizon school only through friendships. He produced over 4,000 lithographs (many of them newspaper illustrations), which are visually powerful expressions of his passionate convictions. His best work ranks with that of the greatest masters of printmaking.

A number of French artists were solitary figures working outside of any school; Charles Méryon, Rodolphe Bresdin, and Odilon Redon, for example. Méryon led a short, tragic life, living in poverty and dying insane. His major work is a series of landscapes of Paris—powerfully drawn, moody prints combining an air of mystery with morbid poetry. Bresdin was also a solitary figure, unappreciated and misunderstood most of his life. His etchings and lithographs are characterized by completely personal and fantastically rich imagery. The great symbolist



"Rue Transnonain, April 15, 1834," lithograph by Honoré Daumier. 44.5 cm × 29 cm.
By courtesy of the National Gallery of Art, Washington, D.C., Rosenwald Collection

painter Redon initially made prints under the influence of Bresdin. His graphic work—a few etchings but mostly lithographs—consists of about 206 prints, whose strange, often bizarre imagery powerfully influenced the Surrealists of the 20th century.

Although, basically, the Impressionists were concerned with the creation of light through colour, several artists identified with them made major contributions to printmaking. Of these, Édouard Manet and Edgar Degas are the most important. Both were superb draftsmen, and, in spite of their association with an avant-garde movement, their roots were firmly planted in traditional art. Manet made a few fine etchings, but his best and most personal works are lithographs, in which his swift but astonishingly precise drawing found its proper medium. Degas's drawings of horses and ballet dancers are miracles of observation and precision—as are his etchings and lithographs. Degas also made a series of monoprints, including a group of remarkably abstract landscapes. The grand old man of the Impressionists, Camille Pissarro, made 194 prints, both etchings and lithographs. His fine graphic work is representative of forceful Impressionist drawing.

The discovery of Japanese colour woodcuts was a revelation that profoundly influenced European art. Until the middle of the 19th century, Japanese printmaking was unknown to the West. As trade relations opened up with Japan, some colour prints came into the hands of young Parisian artists, who responded to the exotic images with great enthusiasm. The simple, abstract handling of colour and design represented a totally new visual experience. Paul Gauguin was one who profited greatly from their influence, which is perhaps more evident in his paintings than in his prints. Following centuries in which the woodcut was used for reproduction, Gauguin's powerful, boldly cut wood blocks were like a breath of fresh air. In the prints of Henri de Toulouse-Lautrec, Japanese influence is more immediate. Although most of his prints were lithographs, the simple bold design, the flat, decorative colour, and the startling disposition of blacks clearly show this influence, which he assimilated and turned into a thoroughly personal expression. A very strong Japanese influence can be seen also in the brilliant color aquatints of the American-born Impressionist Mary Cassatt.



"Mahana Atua," woodcut with touches of watercolour by Paul Gauguin, *c.* 1893–95. 18.3 cm × 20.5 cm.
By courtesy of the Art Institute of Chicago

Influence of Japanese woodcuts on European art

The giant of Postimpressionism, Paul Cézanne, made three etchings and three lithographs. His immense influence on modern art makes his colour lithograph "The Bathers" (*c.* 1900) an important graphic document. The Dutch artist Johan Barthold Jongkind lived in France, creating sensitive landscapes and marine etchings that were a transition between the Barbizon school and Impressionism.

By courtesy of the Art Institute of Chicago, gift of W.S. Brewster



"Le Bon Samaritain" ("The Good Samaritan"), etching by Rodolphe Bresdin (1825–85). 56.3 cm × 44.3 cm.

Japan. The most famous Japanese master of woodcut, Katsushika Hokusai, was born in Hōjō, a suburb of Edo. From the age of 13, when he became an apprentice, until his death in 1849 at the age of 89, he produced an unending stream of masterpieces—about 35,000 drawings and prints, a staggering figure even considering his long life. He also wrote books and poems. There are few masters in the history of art whose work is comparable to Hokusai's in variety and depth. His interests encompassed history and mythology, popular customs, animal life, and landscape. His output was so enormous and the quality of his work so high that it is difficult to single out individual pieces. The "Thirty-six Views of Fuji" (*c.* 1823–29) is probably his most popular set of prints. The 15 volumes of the "Mangwa" (1819) is a fascinating work, for in these rather informal woodcuts the artist gives a comprehensive record of Japanese life and culture. Of all the Japanese masters, the universal genius of Hokusai had the greatest impact on European art.

Hokusai's woodcuts

The last master printmaker of Japan was Hiroshige, whose death in 1858 ends the remarkable dynasty of art-

ists that had begun two centuries earlier. He was a great landscape painter and, with Hokusai, the first to capture the European imagination. He was also a versatile artist, famous in Europe as a painter at a time when in Japan he was known mainly as a poet. A river inspector in the Hokaido district, Hiroshige constantly sketched the landscape during his frequent trips to the river stations. These sketches formed the basis of "Fifty-three Stages of the Tōkaidō Highway," his most important series of landscapes.

Like the work of the Impressionists in Western art history, Hiroshige's prints are spontaneous both in style and in atmosphere, capturing the essence of the fleeting moments of nature.

Other countries. In Germany, Max Liebermann made a few etchings of real individuality, but the most important German achievement of the period was the invention of lithography (*c.* 1796) by Aloys Senefelder, who was not an artist. Although the Belgian artist Félicien Rops lived outside France, he was strongly influenced by the school of Paris. His witty, erotic etchings represent a minor but personal expression of the period. In Sweden, the enormously successful Anders Zorn made etchings and drypoints with great virtuosity.

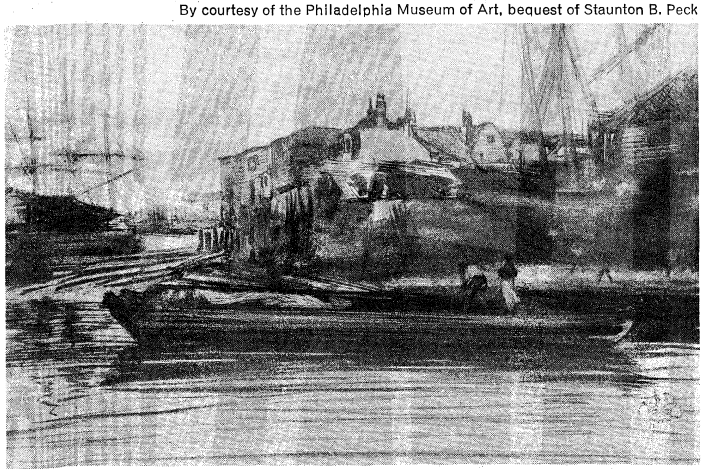
Contributions of Haden and Whistler

English printmaking of the 19th century centred around two great personalities, Sir Francis Seymour Haden and his brother-in-law, James McNeill Whistler. Haden was a Victorian country gentleman, a surgeon who loved and collected etchings. He started to make prints in his leisure time—and ultimately produced over 200 plates. His etchings, sensitively observed documentations of his environment, represent a significant contribution to the English landscape tradition. Whistler was born in America and attended West Point for a period; but he left to study art in Paris, where he met many of the leading artists, including Degas. In 1859 he went to London, where he resided until his death. Whistler was an immensely gifted, complex personality. Simultaneously with his fashionable portraiture, he did a great deal of experimentation; in the nearly abstract paintings and prints that he called "Nocturnes" (begun in 1866), for example, he was far ahead of his time. His graphic oeuvre, 442 etchings and drypoints and 150 lithographs, had great impact on modern printmaking; in particular, the freedom and painterliness of his etchings acted as a strong liberating influence.

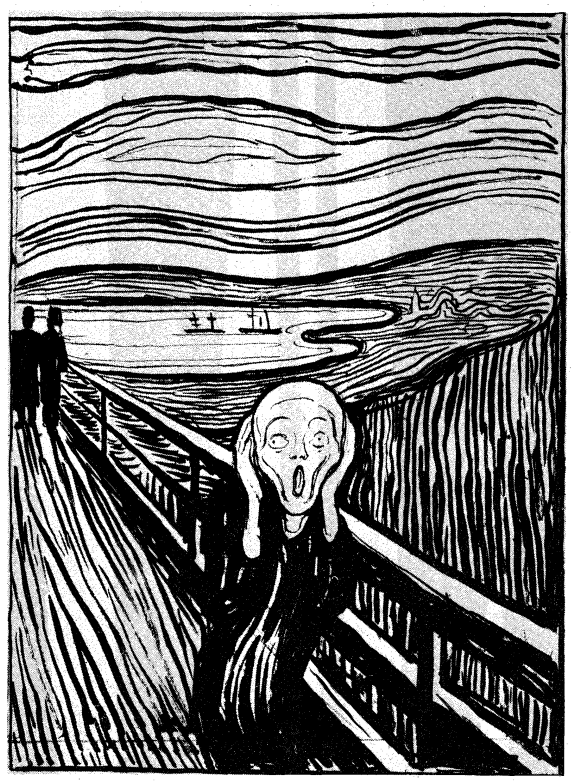
Printmaking in 19th-century America was still provincial and did not produce any artist comparable to the European masters. The colour engravings of flora and fauna executed by the naturalist John James Audubon constitutes a significant body of work, however.

In Mexico, the popular illustrator José Guadalupe Posada produced thousands of woodcuts and lead cuts for newspapers in a completely original style—a mixture of sophistication and the naïveté of popular art. His work had a substantial influence on the young Mexican revolutionary art movement.

By courtesy of the Philadelphia Museum of Art, bequest of Staunton B. Peck



"The Limehouse," lithotint by James McNeill Whistler, *c.* 1887. 17.5 cm × 26.6 cm.



"The Cry" ("Geschrei"), lithograph by Edvard Munch, 1895. 35.3 cm × 25.4 cm.

By courtesy of the Museum of Modern Art, New York, Matthew T. Mellon Fund

PRINTMAKING IN THE 20TH CENTURY

Impact of photography on printmaking

The invention of photography (1826) had a great impact on the development of the visual arts. Its effect was the most immediate on printmaking: photographic reproduction processes made reproductive printmaking obsolete, and printmaking was returned to the creative artist.

The experimental attitude that originated with the Impressionists accelerated in the 20th century. The new styles and new directions that arose with bewildering rapidity made the first half of the century one of the most exciting periods in the history of art.

Continuing the pattern set in the 19th century, France dominated the art world. Attracted by its creative climate, young artists like the Spaniard Pablo Picasso flocked to Paris from other countries and, together with the French, formed the school of Paris, which produced many first-rate artists.

At the same time, Germany became again a vital art centre. German Expressionism and later the Bauhaus school not only produced a number of distinguished artists but eventually exerted international influence.

The following discussion deals only with the "old masters" of contemporary art, those in historically secure positions. Four transitional figures are of particular importance because they represent a bridge between the 19th and 20th centuries. Edvard Munch was an extraordinarily gifted Norwegian painter and printmaker who worked in Paris and in Berlin. His intense imagery, with psychological undertones, relates him to German Expressionism. A versatile artist, he made outstanding etchings, drypoints, colour lithographs, and experimental woodcuts. The Belgian artist James Ensor made superb etchings in a style related to Impressionism, but with fantastic imagery that was close to Surrealism. Close friends, the Frenchmen Pierre Bonnard and Édouard Vuillard produced similar graphic works. Inspired by the Japanese, both made sensitive, beautiful colour lithographs.

France. Pablo Picasso is without doubt the most dramatic and monumental figure of contemporary graphic art. Besides being a superb painter and sculptor, he has created a graphic oeuvre so rich and all-encompassing that he stands alone. He has made well over 1,000 prints,

including etchings, engravings, drypoints, woodcuts, lithographs, and linoleum cuts. Georges Braque, the cofounder with Picasso of Cubism, produced ten major Cubist etchings. The distinguished French painter Henri Matisse was a remarkable colorist and a highly accomplished draftsman. Although the majority of his more than 500 prints are lithographs, he also made some outstanding line etchings and, late in his career, some cutout prints that are masterpieces of design and colour orchestration. Georges Rouault, the French Expressionist, was a solitary figure in contemporary art. The most important graphic work of this religious painter was the "Miserere," a set of etchings published in 1948. Jacques Villon, a major French printmaker, was recognized late in his life as a great painter. Early in his career he made colour aquatints, after the paintings of his more celebrated contemporaries, that raised the level of intaglio colour printing to new heights. Later he developed a completely personal style within the Cubist tradition. He made over 600 prints including engravings, etchings, drypoints, and colour lithographs. The poetic, naïve, and, at the same time, sophisticated style of Marc Chagall, a Russian Jewish member of the school of Paris, sets him apart from any art movement. In his significant body of graphic work, the most accomplished prints are illustrations of the Bible, the works of the Russian writer Nikolay Gogol, and the fables of the Frenchman Jean de La Fontaine. Like his compatriot Picasso, the Spanish painter and sculptor Joan Miró was a prolific printmaker. His witty colour etchings and lithographs represent an achievement equal to his paintings. Max Ernst was a founder of Surrealism and one of the most inventive and influential members of the group. In his extensive graphic work, he introduced a number of new techniques; most notable was his imaginative use of the "collage" in printmaking. Stanley William Hayter, an English painter-printmaker who lived in Paris, has an important position in the development of contemporary experimental printmaking. His significance lies not only in his work as an artist but also in his influence as a teacher. In the 1930s his "Atelier 17" printmaking group was the centre of experimental intaglio work in Paris. In the 1940s he came to the United States and, through his teaching in New York, exercised a powerful influence on contemporary American printmaking. Other artists who did noteworthy graphic work in France include Jean (Hans) Arp, Salvador Dali, André Derain, Jan Dubuffet, André Dunoyer de Segonzac, Alberto Giacometti, Fernand Léger, André Masson, Louis Marcoussis, and Jules Pascin.

By courtesy of the Fogg Art Museum, Harvard University, Cambridge, Mass.



Jazz, Plate XV, serigraph by Henri Matisse, 1947. 41.9 cm × 64.7 cm.

Influence of Hayter on printmaking

Germany. Unlike the extremely varied school of Paris, German Expressionism was quite homogeneous and also much less international. The Expressionists were not united by an aesthetic theory but by their human attitudes and spiritual aspirations. Nearly all of them were active in printmaking; and, although they worked in every contemporary graphic medium, the directness of drypoint and woodcut most appealed to their temperaments.

Lovis Corinth represents a transition from 19th-century naturalism to the Expressionist movement. Although Corinth made etchings, woodcuts, and lithographs, his rich, virile drypoints are his best work. Although not innovative, Käthe Kollwitz's moving, powerful protest prints against war and poverty are significant graphic statements. Ernst Ludwig Kirchner, one of the major figures of German Expressionism, produced a rich graphic oeuvre consisting of etchings, lithographs, and woodcuts. His experimental colour woodcuts represent one of the most distinguished achievements in contemporary graphic art. Emil Nolde produced prints characterized by violent imagery. He worked spontaneously, often making woodcuts without preliminary drawings. Although Nolde came late to graphic work, he left an impressive number of woodcuts, etchings, and lithographs. Max Beckmann was an outstanding draftsman, who made many woodcuts and drypoints. In the latter technique he created some of the finest portrait prints of the 20th century. During World War II, Beckmann came to the United States, where he exerted considerable influence through his teaching. George Grosz used etchings and lithographs to give savage expression to his social criticism of Germany between the wars. The following Expressionists also left significant graphic work: Ernst Barlach, Erich Heckel, Oskar Kokoschka, and Karl Schmidt-Rottluff.

Artists of the Bauhaus group

Other artists moved in a more formal, abstract direction. Based on their philosophy of "new objectivity," they founded the Bauhaus school in Germany in 1919. The two major artists in this group were the Russian Wassily Kandinsky and the Swiss Paul Klee. Kandinsky was one of the great innovators of contemporary art. In his early, lyrical paintings he was a forerunner of Abstract Expressionism, and in his late mature work he introduced Geometric Abstraction. His graphic work consists of an impressive number of woodcuts and lithographs. The whimsical, lyrical abstractions of Klee also had great influence on the course of modern art. His work—about 120 etchings and lithographs—is full of graphic invention and a rare sense of humour. Lyonel Feininger, born in the United States of German parents, studied in Europe and worked most of his life in Germany. He was associated with the Blaue Reiter group (artists who wished to express through their work the spiritual realities they felt had been ignored by the Impressionists) and then in 1919–33 with the Bauhaus. Feininger concentrated mostly on landscapes, executed in a very personal Cubist style, and was one of the most productive graphic artists at the Bauhaus. In the beginning, he made some etchings and lithographs but from 1918 worked mainly in woodcuts. Josef Albers, also associated with the Bauhaus, was born in Germany and moved to the United States in 1933. He has made a considerable number of prints including, recently, colour silk-screen prints. Rolf Nesch was born in Germany, where he started printmaking with the encouragement of Kirchner. He fled to Oslo from Germany in 1933. One of the most gifted experimental printmakers of the 20th century, Nesch developed the method called metal graphic, which he uses to make extremely intricate, heavily embossed colour prints.

Other countries. Printmaking in Italy was far behind France and Germany. The Futurist artist Umberto Boccioni made a few interesting etchings and the Cubist Gino Severini published a number of rather manneristic etchings and colour lithographs, but neither could be considered important printmakers. Giorgio Morandi is the only major Italian printmaker of this period. His intimate, delicate still-life and landscape etchings occupy a very special position in contemporary graphic art.

In Great Britain, Henry Moore, one of the great sculptors of the 20th century, published a number of strong lithographs. Graham Sutherland, a painter, made over 100 etchings and lithographs in a distinctly personal style. Anthony Gross, one of the most talented and prolific English printmakers, has published an impressive body of excellent landscape etchings and engravings.

In the United States, after the turn of the century, most

of the prominent painters became fairly active printmakers: George Wesley Bellows, in lithography; John Sloan and Reginald Marsh, in etching; Milton Avery, in drypoint; and Stuart Davis, in colour lithography. Among these painter-printmakers, two artists are particularly notable: Edward Hopper, whose few etchings are very personal and of unusually high quality; and Ben Shahn, an extremely prolific printmaker, who left an impressive graphic oeuvre in practically every medium.

Revival of the Japanese woodcut

A revival of the art of the woodcut began in Japan in the late 1920s as part of the modern art movement. Onchi Kōshirō and Hiratsuka Un-ichi were early exponents who, though working in different styles, did most for the renaissance of this national art, which thrived once again after World War II. Among the notable woodcut artists of the postwar period are Munakata Shikō and Saitō Kiyoshi.

Since the mid-20th century, there has been a spectacular increase in printmaking activity. Artists all over the world are enthusiastically working and experimenting in every conceivable medium. In this period probably more prints were made and more technical innovations introduced than in the previous history of printmaking.

BIBLIOGRAPHY. ARTHUR M. HIND, *A History of Engraving and Etching from the 15th Century to the Year 1914*, 3rd ed. rev. (1923, reprinted 1963), and *An Introduction to a History of Woodcut, with a Detailed Survey of Work Done in the Fifteenth Century*, 2 vol. (1935, reprinted 1963), cover brilliantly the whole history and development of Western printmaking. Other studies particularly recommended are: JEAN LARAN, *L'Estampe*, 2 vol. (1959), excellent documentation coupled with a volume of fine reproductions; WILLY BOLLER, *Masterpieces of the Japanese Color Woodcut* (1957), not a scholarly book but it covers well the high points of Japanese printmaking; CARL ZIGROSSER, *The Book of Fine Prints*, rev. ed. (1956), an easy-to-read introduction into the history of printmaking; *De Meisterwerken van de Europese Prentkunst 1410–1914*, Rijksmuseum, Amsterdam (1966), catalog of one of the finest shows of master prints ever assembled; A. HYATT MAYOR, *Giovanni Battista Piranesi* (1952), a fine biography of Piranesi with an excellent selection of illustrations; WOLF STUBBE, *Die Graphik des zwanzigsten Jahrhunderts* (1962; Eng. trans., *Graphic Arts of the Twentieth Century*, 1963), good introduction into the history of contemporary printmaking; E.S. LUMSDEN, *The Art of Etching* (1929, reprinted 1962), excellent document on the traditional etching techniques; WILLI KURTH (ed.), *The Complete Woodcuts of Albrecht Dürer* (1946), primarily a picture book with historical background; CARL ZIGROSSER and CHRISTA M. GAEHDE, *A Guide to the Collecting and Care of Original Prints* (1965), a wealth of indispensable information for the collector; STANLEY W. HAYTER, *About Prints* (1962), challenging ideas about printmaking by an important artist and teacher; LEO C. COLLINS, *Hercules Seghers* (1953), an excellent, well-documented book on one of the most important printmakers; J.R. HILLIER, *Hokusai*, 2nd ed. (1957), not a definitive biography but it captures the spirit of Hokusai; LUDWIG MUNZ, *Rembrandt Etchings*, 2 vol. (1949), interesting because it documents Rembrandt's influence as a teacher; K.G. BOON, *Rembrandt: The Complete Etchings* (1963), one of the finest books on Rembrandt with excellent rich reproductions; GABOR PETERDI, *Printmaking*, rev. ed. (1971), a simple but thorough book on both the traditional and experimental intaglio methods; *Redon, Moreau, Bresdin*, Museum of Modern Art, New York (1961), primarily interesting for the Bresdin documentation; CARL W. SCHRAUBSTADTER, *Care and Repair of Japanese Prints* (1948), useful information for the collector; ALBERT KOSLOFF, *Screen Process Printing*, 3rd ed. (1964), a simple, clear description of the basic techniques of screen printing; *The Complete Etchings of Goya* (1943), primarily a picture book; ANDRE MALRAUX, *Saturne: Essai sur Goya* (1950; Eng. trans., 1957), a very subjective responsive to Goya by a great writer, richly illustrated; MAXIME LALANNE, *Traité de la gravure à l'eau forte*, 2nd ed. (1878; Eng. trans., *A Treatise on Etching*, 1880), primarily interesting as a historical document on technique; FRANK and DOROTHY GETLEIN, *The Bite of the Print: Satire and Irony in Woodcuts, Engravings, Etchings, Lithographs and Serigraphs* (1963), an interesting book highlighting the militant, political aspect of prints; BERNARD S. MYERS, *The German Expressionists* (1957), a good introduction to the history of German expressionism, well illustrated; JEAN ADHEMAR, *Toulouse-Lautrec, lithographies, pointes sèches* (1965; Eng. trans., 1965), the definitive book on Toulouse-Lautrec as a printmaker; MICHAEL KNIGIN and MURRAY ZIMILES, *The Technique of Fine Art Lithography* (1970), an excellent, well-organized book on a complex subject; RALPH E. SHIKES, *The Indignant Eye: The Artist As Social Critic in Prints and Drawings from the Fifteenth Century to Picasso* (1969).

(G.F.P.)

Pripet Marshes

A vast, soggy, river-laced, and thickly forested region, the Pripet Marshes (Pripet Polesye ["woodlands"] in Russian; Polesje in the Akademiya Nauk transliteration system) sprawl across the great flat expanses of the western portion of the East European Plain (Vostochno-Yevropeyskaya Ravnina) of the Soviet Union. As Europe's largest swamp, they form a remarkable natural feature of major interest to scientists and are in large measure a legacy of the physiographic activities of the last Ice Age. For centuries their inhospitable terrain was of little interest to man, but huge reclamation projects have attempted to win some economic reward from the area during the Soviet period. Though the landscapes of the region are unique, their scientific interest is heightened because they do have parallels in other natural regions of the Soviet Union, both in the eastern European plains (Meshchorsk Lowlands and the Privetluzhye) and in the western Siberian plains (the Vasyuganye), as well as in regions of Poland contiguous to the Soviet Union. Such landscapes are often called *polesski* (*i.e.*, similar to Polesye).

Extent of the Marshes

The Pripet Marshes occupy the south of the Belorussian Soviet Socialist Republic, and the northern *oblasti* (regions) of the Ukrainian S.S.R. Physiographically, the region lies in the basins of the Pripyat and Dnepr rivers. The region is bounded on the north by the Belorussian Ridge (Belorusskaya Gryada) and on the south by the Volya, Podolsk (Podolskaya), and Dnepr uplands.

Physiography. The distinctive natural traits of the Pripet Marshes are a wide development of saturated sandy lowlands, intersected by a dense network of rivers with weakly cut river beds and wide floodlands; a prevalence of pine forests amid wide expanse of low-lying bogs and marsh-ridden lands; and a variegated and difficult landscape impeding full economic use of the land.

The underlying geological structure of the region is formed in large part by a syncline, or great downwarp, in a slab of the ancient rock structure known as the East European Platform; the downwarp is known as the Pripyat Hollow and the Brest Depression. The younger geological areas lie on the southern part of the hard mass known as the Ukrainian crystalline shield. The rocks of the region include ancient granite, gneiss, calcareous marl (a crumbly material consisting mainly of clay, sand, and calcium carbonate), and large quantities of geologically recent sandy–clayey deposits.

The natural peculiarities of the Pripet Marshes are derived in part from the environmental conditions existing in the Ice Age, when the region was smothered in ice and ice-carried material, and also from the action of contemporary physiographic processes. The geologically recent deposits that have so influenced present-day conditions have an average depth of 82 to 164 feet and are mainly represented by glacial sands, sandy–clayey moraines (mounds of material marking the rim of former glaciers), alluvial sands, and lake loams.

As a result, the characteristic relief forms are uplifted plains, river valleys cut in a series of steplike terraces by water action (and merging into flat, weakly expressed watershed areas), morainic plains and ridges (as at Volyn and Mozyr [Mozyrskaya]), and ancient wind-sculpted relief forms. The younger areas are occupied by plains worn down into outcrops of ancient Precambrian rocks. The southwest and eastern areas are characterized by the deeply eroded, strangely shaped limestone rock forms known as karst topography.

Climate and drainage. The region experiences a warm temperate climate. The average January temperatures range from 24° F (−4.5° C) in the southwest to 19.4° F (−7° C) in the east, and the average July temperature is almost 65° F (18 or 19° C). The average annual precipitation reaches 22–26 inches (550–650 millimetres) and exceeds evaporation, giving sufficient—and in some places quite abundant—moisture. In conjunction with an

abundance of subsoil waters and their close proximity to the surface, a virtually unique soil saturation and associated bogging down of the surface are thus produced.

Rivers and streams

The region is also noted for its extensive network of rivers and streams. Numerous tributaries of the Pripyat (including the Stokhod, Styr, Goryn, Ubort, Yaselda, and Ptich), the Dnepr (Sozh, Teterev, Irpen), course down into the swamps from the surrounding highlands, carrying in large amounts of water. In the spring, when snowmelt occurs, the region's rivers overflow their low-lying banks and intensify the saturation of the land. Huge marshes are developed along the course of the Pripyat itself, while the middle of the river is marked by the soggy expanses of the Pinsk Marshes. The character of the marsh regions is shaped by the local topography, itself marked by floodlands, a layer of submerged former river terraces, a second layer of submerged terraces, and the morainic plains of very old valleys and lake hollows. Low-lying marshy grasses, sedges, and sedge bogs of the type known as *gipnovo* prevail, while small areas of mossy bog and the transitional, or mesotrophic, swamp occupy the upriver areas. The numerous lakes that dot the landscape are in various stages of choking up into more bogs. The soil cover of the region consists mainly of the turf type of podzol soils, peat bog, and floodland soils: all need land reclamation measures before they can be used for agricultural purposes.

The human imprint. About one-third of the region is forested, consisting of pines (about 60 percent of the wooded area), birch, alder, oak, aspen, white spruce, and hornbeam, all of which add occasional splashes of lighter colouring against a generally dank and sombre background. The region has thus supported—where conditions permit—a diversified lumbering industry. Elk, lynx, wolf, fox, wild boar, roe, beaver, badger, and weasel are to be seen and are sometimes hunted within the forest fastnesses, and the somewhat gloomy appearance of the area is also brightened by the sounds and movements of a host of birds, including black grouse, orioles, hazel grouse, woodpeckers, owls, blue tits, and wild duck. These, too, are hunted.

The mark of man is most evident, however, in the more promising sections of the region that are being developed and transformed into agricultural lands, where rye, barley, wheat, flax, hemp, potatoes, a variety of vegetables, and fodder grasses are cultivated.

Land reclamation

Land reclamation projects have only been initiated since the late 19th century. They were begun in 1872 by a state-sponsored "western expedition for the drainage of swamps" led by the famous Russian scholar I.I. Zhilinsky. A vast amount of land reclamation has taken place in the Soviet period, especially since the reunification of western Belorussia and the western Ukraine under a socialist government. A complex series of measures for achieving this formidable goal was underway by the early 1970s. They include the regulation of water drainage and the construction of reservoirs on the rivers, the regulation of river channels, afforestation of sandy uplands, and the clearing of shrubs and other undesirable vegetation cover. Reclamation is proceeding under an overall Soviet plan, and it is hoped that the ultimate transformation of the hitherto inhospitable natural conditions will have great significance for a wide range of economic activities.

(A.M.M.)

Prisons and Penology

The prison is a relatively recent invention; as a means by which society deals with criminal behaviour, it has come into widespread use only within the last 300 years. In order to understand the nature of the prison, it must be seen as part of a larger pattern of penal practices that have undergone a number of transformations over the course of time.

Imprisonment as punishment

Imprisonment is essentially a form of punishment, although some have seen it as affording a means of reforming the criminal. Punishment is an ancient response to wrongdoing; but throughout man's history both the forms of punishment and the rationale for using it have changed markedly. Many of the changes can be traced to economic considerations, the development of a more humanitarian outlook, and the emergence of a secular viewpoint in human affairs, as well as to the development of new concepts about the nature of man and society. But if society's treatment of the lawbreaker has become, in theory at least, more rational, more benign, and more concerned with rehabilitation, this does not mean that society's reaction to the criminal has lost its punitive tone. Actually, there is little evidence as to the effect of punishment on the amount of crime in society. It is not known with any certainty that one mode of treatment is more or less effective than another. Penal systems everywhere remain largely based on tradition, untested assumptions, and inferences derived from inadequate data.

GENERAL CHARACTERISTICS OF PUNISHMENT

The rationales for punishment. Punishment inflicted by the state in response to a violation of the criminal law has been justified in various ways. It has been seen as society's vengeance upon the criminal, as atonement by the wrongdoer, as a means of deterring other criminals, as protection for the law-abiding, and as a way of rehabilitating the criminal.

Revenge. The individual who has inflicted harm on another, runs the revenge argument, must be made to suffer in return; for only an act of vengeance can undo the harm that has been done and assuage the suffering of the victim. The 19th-century English legal historian Sir Henry Maine held that punishment evolved from social necessity: in a primitive society private retaliation was all too likely to lead to feuds that could destroy the group, and, therefore, the group as a whole, acting as a formal collectivity, was slowly led to take on the role of the punishing agent. Although anthropology has not been kind to Maine's evolutionary theory, some modern commentators continue to see legal punishment as a device for forestalling disruptive acts of private retaliation.

Expiation. The concept of punishment as a form of expiation or atonement reaches far back into the history of human culture. The roots of the idea are religious; in the Judeo-Christian tradition punishment has been seen as the undoing of a wrong by means of the suffering of the wrongdoer, who thus brings his moral account with God back into balance.

Deterrence. The notion that the threat of punishment by the state will restrain the potential criminal is one of the most commonly accepted justifications for it. The idea has a philosophical base in the utilitarians' concept of the rational man acting upon a deliberate calculation of possible losses and gains. The Englishman Jeremy Bentham, the Italian criminologist Cesare Beccaria, and other members of the classical school of criminology sought to reform the criminal law of the 18th century with this idea of man as a hedonistic calculator. If men chose rationally among possible future courses of action, then surely the likelihood of a criminal course of action could be decreased by attaching to it a quick, certain, and commensurate penalty.

The use of punishment as a deterrent has been justified not only for the convicted criminal. The law-abiding man who totters on the edge of illegal behaviour is also deterred, it is said, and one must calculate the effect of society's penal system not simply in terms of its impact on the known offender who is punished but also in terms of its influence on the great mass of those who conform. As the philosopher Morris Cohen pointed out, this is ethically unsatisfactory: to justify punishment of the criminal by its effect on the noncriminal is a little like hurting Peter to keep Paul honest. As a practical matter, however, an accurate assessment of the efficacy of legal punishment would have to measure its influence on both criminals and noncriminals.

Protection of the public. Insofar as punishment by the state isolates the criminal from the ranks of the law-abiding, it has been argued, the amount of crime in society is likely to be diminished. Such isolation may include branding or otherwise marking the criminal, so that honest men may avoid him; or it may involve thrusting the criminal beyond the pale in some form of exile or con-

fining him in a prison. This argument assumes that the isolation of the offender is not outweighed by his becoming more criminal while in prison. It also assumes that the social and economic costs of isolating the criminal from the rest of society are less than those of the crimes he might have committed if he had been left free. Both of these assumptions have been seriously questioned.

Rehabilitation of the criminal. Some writers have maintained that punishment ought to be designed to transform the values and attitudes of the criminal so that he no longer wishes to commit illegal acts. The problem, of course, has been to discover how to do it. Those with a psychiatric bent have pinned their faith on programs of individual and group counselling and on the creation of what they have called a therapeutic milieu. Others, more sociologically inclined, have looked to such measures as education and job training. Still others have taken a more radical view and argued that the causes of crime do not lie with the individual so much as with the community that shapes him; if criminal behaviour is to be diminished, they say, the social context within which it occurs must be changed. Theories of rehabilitation are largely speculative, since there is a lack of scientific evidence to support them. The paucity of research funds, restrictions on social experimentation, political considerations, and other difficulties have prevented any thorough appraisal of the merits of this approach. Nevertheless, it has been influential in the development of modern penology.

The forms of punishment. Men have shown much ingenuity in devising forms of punishment. Throughout history these have ranged from ostracism and ridicule to flogging, burning, branding, mutilating, disemboweling, crushing, beheading, and mass extermination. In general, however, the many forms of punishment can be classified into five major types.

The decline of corporal and capital punishment

Corporal punishment. The infliction of physical suffering is nowadays regarded as barbarous and has been generally abandoned as a form of punishment in most modern industrial countries. The use of such devices as the knout, the cat-o'-nine-tails, the sweatbox, and deprivation of food and water were abandoned partly because of an increasing value attached to the individual and his person, as well as a growing sense that the person who had broken the law maintained his humanity despite his crime. In addition, there was an increasing acceptance of the idea that much of the physical suffering inflicted in the name of punishment was little more than sadistic torture, as degrading to the society that inflicted it as to the criminal who bore it. Cruel punishments produce cruelty in people, argued the English reformer Sir Samuel Romilly in the 18th century; and modern societies have tended to agree.

Capital punishment. The execution of criminals was carried out in the past by such methods as drowning, stoning, burning, and decapitation; in modern times, hanging, electrocution, and the gas chamber have been the most common techniques.

The use of capital punishment has declined in recent times, although in many places it is still permitted by the law. The issue of abolishing it has aroused much controversy. The advocates of capital punishment claim it is a necessary deterrent to crime and relatively painless if done properly. Their opponents argue that executions are far from physically painless, involve extreme mental anguish, may entail irrevocable miscarriages of justice, and are ineffective as a deterrent since the crimes for which capital punishment is imposed are frequently crimes of impulse or emotion beyond the individual's control.

Even where it has been legally retained, capital punishment is now seldom employed. Fewer crimes are punishable by death than formerly; the courts increasingly choose alternative penalties; and many death sentences are not carried out. In the U.S. the annual number of executions dropped from about 155 in the period 1930–34 to about 26 in 1961–65, and in some subsequent years there were none. The death penalty has been eliminated in such nations as the Scandinavian countries, West Germany, The Netherlands, Austria, Italy, Portugal, and Switzerland.

Penal colonies

Transportation. Exile and banishment have been known since ancient times and were often tantamount to a death sentence. The practice of transporting criminals beyond the national boundaries did not appear, however, until the beginning of the 17th century, when England began shipping offenders to America. The American Revolution forced the British to seek a new dumping ground in Australia. With the discovery of gold and the emergence of sheep raising and wheat growing as profitable ventures, which brought new settlers resentful of the criminal population, the English system of transportation came to an end. France established penal colonies in Africa, New Caledonia, and French Guiana (the latter the location of the famed Devil's Island). Russia set up penal colonies in Siberia under the tsars, and the system was greatly expanded by Stalin. Since World War II transportation as a method of punishment has been generally abandoned. The brutal conditions of life in convict communities, their economic inefficiency, and the lack of space in a world growing ever more populated have been factors in its decline. Rather than expelling the criminal, society has found it more convenient to imprison him.

Fines. The imposition of fines upon criminal offenders has long played a role in many societies. A number of early cultures, including the Anglo-Saxon, had intricate systems of blood money that defined in money terms the value of a man's life—for which his murderer would be forced to pay. There were also fines for personal injuries and larcenies, varying according to the social rank of both the victim and the wrongdoer. As crimes came to be defined as public wrongs against the state rather than private wrongs done to a particular individual or family, fines tended to become a subsidiary form of punishment.

In the modern era, fines remain as a penal sanction for a large number of relatively minor offenses but are seldom considered as the major element in a system of punishment except in the case of a limited number of white-collar crimes; and the money is only occasionally handed over to the person who has suffered an injury. Unless money payments are carefully regulated with regard to the ability to pay, they weigh unequally on persons from different socioeconomic levels, and may even become an informal method of licensing illicit activities. A systematic use of fines is made in some countries; for example, in Sweden, where fines are adjusted to the income of the offender and may be paid on the installment plan. The state of Wisconsin in the U.S. allows prisoners to work off their fines by holding jobs outside the institution in which they reside, paying for their maintenance and the support of their families. In general, however, in the U.S. fines remain a rather haphazard adjunct to the penal process.

Early prisons

Imprisonment. Locking up criminals in dungeons and jails is a practice that goes as far back as recorded history. Before the 18th century, however, such confinement was largely reserved for suspected offenders awaiting trial or for convicted criminals awaiting the lash or the hangman's noose. Only in the last 300 years or so has the prison been an important penal sanction. Even today it is not the method applied to most criminals: approximately two-thirds of the offenders in the U.S. penal system, for example, are under supervision in the community rather than behind walls.

Although some institutions akin to the prison existed at an early date, such as the Delle Stinche in Florence in 1300, the first modern prison is thought to have been the Walnut Street Jail built in Philadelphia in 1790. Its antecedents are to be found in the workhouses and houses of correction established in London (1557), Amsterdam (1596), Rome (1704), and Ghent (1773). In these institutions there was little segregation by age, sex, or other characteristic. The main emphasis was on strict discipline and hard labour. Those who operated the houses of correction depended for their living on fees paid by the prisoners.

By the end of the 18th century, the modern concept of the penitentiary had taken firm shape, partly through the work of John Howard in England. Upon becoming high sheriff of Bedfordshire in 1773, Howard was appalled by the conditions of the county jail and spent a number of

years travelling through Europe to study methods of penal treatment. The squalid living conditions, idleness, and immorality that he found in the crowded houses of correction distressed him. Like other prison reformers of the time, he also had a strong belief in the beneficial effects of solitary confinement. He conceived of the penitentiary as a firmly ruled institution where criminals would be kept completely apart from one another and would work at regulated labour. In the 1790s Jeremy Bentham designed the "panopticon"—a circular, glass-roofed structure with cells in tiers around the wall facing a central rotunda where guards could keep the prisoners under constant surveillance. It was a monstrosity by modern penological standards and was never built in England, although its influence is to be seen in the maximum security institutions at Breda in Holland and at Joliet, Ill., in the U.S.

THE DEVELOPMENT OF MODERN PENAL SYSTEMS

The "separate" and "silent" systems

Cherry Hill and Auburn. The ideas of solitary confinement and penal labour were first tried out in the U.S. The early part of the 19th century saw a sharp debate between two schools of penal philosophy, that of the "separate system" and that of the "silent system." The former argued for complete isolation, with each inmate working, eating, and sleeping in his own cell. The latter believed that prisoners should be permitted to work together in silence but be locked in individual cells at night. The separate system, widely known as the Pennsylvania system, was embodied in a prison architecture with rows of cells radiating from a common centre. In the Eastern State Penitentiary built on Cherry Hill in Philadelphia in 1829, seven cell blocks were arranged like the spokes of a wheel around a central administrative structure; the cells were about 16 feet high, 12 feet long, and 7½ feet wide. An enclosed exercise yard was attached to each cell, and the prisoner saw no one but his captors.

The silent system gave rise to prisons made up of cell blocks, with many small cells piled in tiers in single buildings. A model prison was built at Auburn in New York in 1825. The advantage claimed for the Auburn system over the Pennsylvania system was that it allowed prisoners to work together at industrial tasks rather than restricting them to solitary handicrafts such as weaving or shoemaking. Communication among the prisoners was forbidden, however; and as they moved back and forth between their cells and the workshops, they were required to march in lockstep to ease the guards' task of surveillance.

In retrospect, the two schools of thought seem barely distinguishable. Both emphasized strict discipline, hard labour, and elemental living conditions; both sought to prevent the inmates from associating with one another. The Auburn system came to dominate most of the thinking in the U.S. even though European observers believed the Pennsylvania system superior and promoted its adoption in Europe, where it still exists in some places in modified form.

Experiments elsewhere in the 19th century played a significant part in the further development of penological thought. At Norfolk Island, an English penal colony off the coast of Australia, Captain Alexander Maconochie developed the "mark system" through which confined criminals could reduce the length of their sentence by means of conscientious work and good conduct, for which they earned marks or credits. In France, Bonneville de Marsangy advocated the use of conditional release, or parole, for good behaviour. Occupational training for confined criminals was emphasized by a number of prison administrators, notably in Spain and Austria, not simply as a device to help make the penal institution self-supporting but as a crucial element in the offender's rehabilitation.

The "intermediate system"

The reformatory movement. These innovations were brought together in Ireland by Sir Walter Crofton in what he called the "intermediate system" to stress the place of the penal institution as a halfway point between the free community and strict confinement. Depending on the inmate's behaviour and his degree of reformation, he could move either toward greater liberty or toward greater restraint. This new ideal of the penitentiary as a place of training and performance captured the imagination of prison officials in America and in the 1870s gave rise to the reformatory movement.

The National Prison Association established in 1870 adopted a declaration of principles that called for a wide range of reforms, including the classification of prisoners, indeterminate sentences, professional education for prison administrators, industrial training for prisoners, a reward system for good conduct, and changes in prison architecture. The reformatory philosophy soon permeated most thinking in the U.S. and Europe. In practice, however, many of the world's penal institutions are quite far from being reformatories. Social inertia, lack of funds, the growth of the criminal population, and a continuing unwillingness on the part of the general public to abandon punishment as a policy have meant that a large number of prisons and reformatories make no more than a gesture in the direction of rehabilitation.

There have been some imaginative efforts. In the U.S., for example, the Highfields project in New Jersey has made thoroughgoing efforts at the reformation of juvenile delinquents, providing psychological therapy and intensive guidance. The number of offenders that can be handled in such an experimental institution is small. Some state correctional systems, including those of California and Wisconsin, have pursued policies designed to rehabilitate the inmate rather than merely keep him confined; and the federal penal system has long had a reputation for well-administered and enlightened programs.

The picture in England and western Europe is perhaps even more varied. Some institutions—such as the Herstedvester Detention Center at Albertslund, Denmark, which provides a therapeutic environment for 200 men with long criminal records and serious problems of social adjustment—are in the forefront of penological innovation. Others are relics from the 19th century, still imposing solitary confinement and lacking even the pretense of a rehabilitative program.

The trend toward small, informal, treatment-oriented institutions is probably most advanced in the Scandinavian countries, particularly among the well-staffed residential schools for juveniles.

Other approaches. An alternative theory of rehabilitation, that of the "short sharp shock," gained some adherents after World War II. According to this view, offenders of the delinquent gang type ("teddy boys") do not require residential treatment; but neither do they benefit from supervision outside the prison. The solution, it is argued, is a short, severe experience in a sternly administered detention centre of the kind found at Kidlington in England and at Hammargarded in Sweden. The merits of the theory are in dispute, and the effectiveness of the treatment remains to be demonstrated.

Labour colonies exist in Finland and the Soviet Union. Descriptions of the Soviet labour colony at Kryukovo suggest that the psychiatric interpretation of criminal causation and rehabilitation is largely ignored. The institution possesses three regimes characterized as mild, general, and strict—a modern version of the intermediate system. Inmates serve long terms and are employed in a program of industrial production geared to the economic needs of the society. Discipline is said to be maintained by a deliberate use of group pressures and political indoctrination.

The Borstal system

The Borstal system for young offenders in England is probably the most noteworthy example of an attempt to build a large-scale correctional program incorporating modern penological theories. It was established in 1908 by Sir Evelyn Ruggles-Brise, director of English prisons, after he had studied several U.S. reformatories. It consists of a variety of small institutions (usually with no more than 50 inmates) designed for specific types of offenders. Sentences are indefinite. There are facilities for investigation and diagnosis, a relatively well-trained staff including a number of psychiatrists and psychologists, extensive vocational training, and a program of lengthy aftercare that is well integrated with institutional treat-

ment. The Borstal system represents a major effort to free penal practices from the traditional punitive orientation. Its results, however, do not seem to support the hopes of its proponents. Critics have pointed out that the percentage of inmates released from the Borstal system who reappear in court for new offenses is almost identical with the percentage of inmates released from the regular prison system who run afoul of the law. In general, there is little evidence that any method of penal treatment is superior to others, as measured by the numbers of repeaters. This is not to say that one method may not be preferable on humanitarian grounds or on some other basis; it is simply that no convincing proof of rehabilitative effectiveness has been produced.

Appraisal is made all the more difficult if, as seems likely, the patterns of crime change over time. The official designation of the illegal act remains the same, but the motive for committing it changes. A case in point is the increase in serious crimes committed by delinquent gangs in Sweden and other European countries in the 1960s—gangs made up of youths alienated from modern mass society. Juvenile criminals of this sort are not likely to respond to the correctional methods that have worked with other juvenile offenders.

PENAL INSTITUTIONS

The diversity of penal practices

Penal practices. Prisons, reformatories, and other correctional institutions are extremely diverse, ranging from decaying fortresses to mobile forestry camps. It is possible, however, to make some generalizations about institutional penal practices in Western countries, keeping in mind the fact that there are numerous exceptions.

Custody. Few penal institutions can avoid a concern with maintaining custody. Juvenile reformatories are the least stringent. Many countries separate their other institutions into categories of maximum, medium, and minimum security. Preventing escapes, however, remains of primary importance and accounts for many of the rules governing the inmates. Penologists may disagree as to whether keeping the offender confined is for the purpose of punishment, deterrence, or rehabilitation. The general public usually shows no such uncertainty and quickly becomes indignant at news of any escape. It is not surprising that custody is so often given the highest priority and that the overwhelming majority of institutional personnel are devoted to maintaining it.

Work. The inmates of institutions must somehow be supported, and the traditional solution has been that they should support themselves as far as possible. There are various difficulties in this. The organization of an efficient labour force can create serious custodial problems and also arouse opposition from private enterprise outside. Many institutions limit themselves to so-called state-use industries (such as the manufacture of automobile license plates and highway signs) to avoid competing with nonprison labour, and to the routines of institutional housekeeping. These do not provide much training for employment in a modern industrial society. Special efforts in this direction are often hampered by a lack of equipment.

Reduction of sentence. It has long been the custom to reduce the offender's time in confinement as a reward for good behaviour and conscientious performance at work. The prisoner may go through the necessary motions in the hope of reducing his sentence, but the original hope that a carefully graded system of rewards would lead the offender along the path of rehabilitation has not been fulfilled. One experiment at the National Training School in Washington, D.C., has permitted inmates to earn "points"—equivalent to money—for good behaviour, and these can be used to obtain such things as individual sleeping quarters, better meals, and recreational equipment. Whether this conduces to reformation, however, is not yet known.

Education. A large proportion of offenders have a history of failure and retardation in school; and many institutions offer educational programs in the hope of remedying these deficiencies. One difficulty is that prisoners often dislike schooling; another is the prevailing low quality of the teaching.

Self-government. Approximately 95 percent of the offenders committed to penal institutions eventually return to the outside world, and it has seemed essential to many penologists that inmates learn to live in the free community by participating in some form of democratic social structure while still in the institution.

One of the first major attempts to introduce self-government was made in the Auburn and Sing Sing prisons of New York shortly before World War I. Since that time it has been tried at many institutions, particularly at honour farms and juvenile reformatories. The efforts have met with varying degrees of success, but it would appear that in many institutions it has been little more than a formality. Like the trusty systems that give power and responsibility to a few individuals, inmate self-government is subject to the sort of abuse in which some inmates exploit their fellow prisoners and use their privileged position as a means of avoiding institutional rules.

Visiting. The privilege of seeing family members and friends from the outside world exists in almost all penal institutions, although it has never achieved the clear status of a right and is hedged about with numerous restrictions. In many cases, inmates are separated from their visitors by glass partitions through which they communicate by phone. In a few institutions, such as the California Institution for Men at Chino, inmates and members of their families are allowed to mingle freely in an open setting.

The practice of conjugal or sexual visits is established in Latin America, Scandinavia, Eastern Europe, and some countries of Asia. In the U.S. it is permitted on an experimental basis under special conditions in Oregon, Mississippi, and California. A number of penologists have argued for such visits as a means of combatting the homosexuality that flourishes in prisons. An alternative approach is to allow prisoners to go home on furlough. This is easily integrated with the concept of the halfway house or prerelease detention centre that is being adopted in some countries.

Rehabilitation. As noted earlier, the view that a penologist takes of rehabilitation is likely to depend on his general approach to the problem of crime. Some institutional practices—such as vocational training, education, visiting, and self-government—are compatible with a sociological view of rehabilitation; that is, such programs presumably help the offender fit himself for the social and economic roles he must play in society. For penologists more inclined to psychology, however, such programs hardly touch the causes of most criminality. These penologists point hopefully to the many institutions that make some effort at psychological therapy. But no country has yet been able to muster the staffs that would be necessary to make psychological treatment a reality. In 1970 there were fewer than 700 psychiatrists and psychologists employed in all the prisons, reformatories, and jails of the U.S. Four times as many would be needed to meet the standards laid down by the President's Commission on Law Enforcement and Administration of Justice (1967).

A form of psychological therapy that has been regarded with much favour is guided group interaction, which seeks to involve inmates in prolonged and intensive group discussions of their behaviour and motivation. It is less demanding of staff resources than individual counselling. It also bears directly upon the nature of the offender's difficulties with the law, which generally have to do with the social groups to which he belongs and his relationships with other group members.

Institutional design. The design of the penal institution is important to the penologist from an aesthetic as well as a practical view, since the outward appearance of the prison is often taken as a symbol of its philosophy. The fortress-like buildings of the past are giving way to smaller, more informal institutions with an educational or camplike atmosphere. This trend is limited, however, by the desire to control interaction among the prisoners. Dormitory sleeping facilities in juvenile institutions, for example, are considered undesirable by many officials, not only for their lack of privacy but also because they are thought to foster the development of an independent

youth culture that can run counter to the aims of the custodians. Some institutions have provisions for single sleeping quarters that can be locked and unlocked by the inmates, although the guards hold master keys.

The prison community. There have been many illuminating accounts of life in prison by first-rate writers who have been prisoners: Fyodor Dostoyevsky, Arthur Koestler, Oscar Wilde, Miguel de Cervantes, e.e. cummings, Brendan Behan, and Jean Genet, among others. Social scientists, in their more pedestrian way, have provided a large number of detailed studies of prisons that add up to a fairly unified picture of prison life, although most of these investigations have concentrated on large institutions for adult male offenders.

The repressive nature of the prison regime

The regime of the custodians. The social order that the custodians attempt to impose on their captive population is designed to achieve a number of objectives that are often in conflict. The task of custody is frequently assigned the highest priority and is likely to dominate all other concerns. The continual checking of cells, searches for contraband, the detailed regulation of inmates' movements about the prison, the 24-hour surveillance—all are necessary, in the opinion of the custodians, to prevent escapes.

There is also the task of maintaining internal order. Under the best of circumstances it would be difficult to keep peace in a population of men with long criminal histories. The task is doubly difficult in a prison, where the irritations and frustrations of life are harder to avoid. In order to forestall thievery, sexual assaults, and other kinds of prison crime, the inmates are subjected to rules governing every aspect of life. These do much to give the social structure of the prison its authoritarian character.

A third task is self-maintenance: the cleaning, cooking, barbering, record-keeping, and providing of medical services that are needed to keep the institution running. Many inmates are employed at these functions. Others work at tailoring, the manufacture of license tags and office furniture, and whatever industrial occupations are permitted to prisoners. These are limited, both by the need to avoid competing with outside industries and by the requirements of custody and security. The prison often finds itself with too many men and too few jobs.

Finally, there is the task of rehabilitation. Limited as they are in methods and resources, many prisons can only hope that if they do not make their inmates better they at least will not make them worse. Education, recreation, visiting privileges, and similar measures may be provided not in any coherent program of rehabilitation, but simply in the hope of mitigating the more corrosive aspects of the prison environment. Even these will be carefully hedged about with restrictions.

In summary, the regime of the custodians tends to be severely repressive, requiring the minute regulation of inmates' lives. Meaningful activity is at a minimum, and the aim of reformation is likely to be honoured in words rather than in coherent programs.

The difficulty of maintaining control

The prison bureaucracy. The organization designed to put this regime into effect is what sociologists call a typical bureaucratic structure, with written rules, well-defined areas of responsibility, and impersonal standards of performance and promotion. At the bottom are the guards. Immediately above them are the sergeants and lieutenants, who in turn are responsible to the warden and his assistants. A staff of treatment employees, including teachers, clinical psychologists, group therapists, and social workers, is also under the warden's control.

In the effort to carry out their program the prison administrators encounter a number of difficulties. Their power rests on the authority of the law, but it lacks moral support among the inmates. The custodians confront a population that must be forced, bribed, or cajoled into obedience. If they attempt to rule by force they quickly discover how ineffective this is in getting the normal activities of the prison community performed. A straitjacket or a pair of handcuffs may control one rebellious inmate in an emergency, but these will be of little aid in moving hundreds of men through a mess hall. A blow with a club may check an imminent revolt, but it will not ensure the operation of machines for industrial production. If physical coercion as a means of securing compliance proves to be something of an illusion, so also does the use of punishments and rewards. The inmates of the prison are already close to the bounds of punishment allowed by society and the few restrictions left, such as solitary confinement, are limited in their use. Permissible rewards are similarly restricted, since society is opposed to indulging criminals. The consequence is that prison administrators are seldom the unquestioned rulers they are assumed to be by the public; they can maintain control over their community only by reaching a modus vivendi with the inmates.

Guards and inmates

The guard is a pivotal figure in the social structure of the prison community. He is expected to be a combination of policeman, foreman, magistrate, counsellor, and boss. But the guard, too, is in a certain sense serving time in prison. He is frequently reluctant to enforce the full range of the prison's regulations. He may fail to report infractions that he has seen; he may transmit forbidden information to inmates, neglect elementary security measures, or join the inmates in outspoken criticism of the prison administration. This "corruption" of the guard's position is seldom the result of bribery; in fact, bribery is not necessary, for there are other forces at work to bridge the gap between guards and inmates.

First, the guard is in close and intimate association with the prisoners, making it difficult for him to maintain a great social distance. He cannot physically withdraw in symbolic affirmation of his superior position, and he has no intermediaries to absorb the resentment generated by orders and regulations that are disliked. Furthermore, the guard is not necessarily hostile to all inmates, nor does he feel that all prisoners must be rigidly controlled. Some of the inmates, such as those convicted of wife desertion, are not likely to appear to him as dangerous criminals. The guard and the inmate are thus pushed in the direction of a nonauthoritarian relationship.

Second, the guard in prison is frequently dependent on the inmates for the satisfactory performance of his duties, since, like many persons in positions of power, he is evaluated in terms of the men he controls. A troublesome, noisy, dirty cell block reflects on his ability to "handle" prisoners, and this ability is an important component of the merit rating used in pay raises and promotions. The guard often finds it necessary to "trade" with the inmates, tolerating or ignoring "minor" infractions of the rules in exchange for compliance in "important" areas. Permission to gamble, pursue a forbidden hobby, or be out of the cell without authorization may be exchanged for orderly movement to and from work, good behaviour toward the guard, or a well-swept cell block before an inspection.

Finally, much of the guard's authority tends to be whittled down by the innocuous encroachment of inmates on the guard's duties. The making out of reports, the examination of cells during the periodic counts, the locking and unlocking of doors—these minor chores that the guard must perform each day may gradually be transferred into the hands of inmates whom the guard trusts.

The prison administration, in short, cannot function without some degree of uncoerced cooperation from the prisoners. What one may call the corruption of authority is greater in some institutions than in others and varies in its consequences: in some prisons the inmates may exploit one another and violate regulations with little interference from the guards, while in others there is only a slight amount of relaxation. But few prison communities bear much resemblance to the common picture of the prison in which a brutal guard tyrannizes over a criminal locked by himself in a cell.

Relationships among inmates. If interaction between guards and inmates is part of the reality of prison life, the relationships among the inmates are equally significant. Students of penology have long recognized the importance of the daily struggles and deceits, the acts of submission, friendship, and leadership that take place in the inmate social system.

A central problem is prolonged deprivation. One of the

The deprivations of imprisonment

most obvious deprivations of the prisoner is the loss of material goods and services. When he enters his civilian possessions are taken away and he is given a prison uniform. Later he will be given other things—a cup, a spoon, blankets—but never much above the level of bare necessity. Poverty is a part of prison policy.

A second deprivation is in the sexual sphere. The prisoner is denied any heterosexual contact, and the prison administrators must contend with the resulting homosexuality as best they can. Some prisoners will have been homosexuals before their confinement; but these are likely to be relatively few in number and may well be victimized or raped by aggressive prisoners who have turned to homosexuality as a temporary resort. The deprivation of heterosexual relationships must be viewed not only as a physiological loss but also as a psychological one: the inmate is shut off from the world of men and women that by its very polarity gives the word male (or female) much of its meaning.

Third, the great mass of regulations that make up the custodial regime deprives the inmate of much of his autonomy as an individual. For some inmates, of course, the authoritarian character of prison life may offer a welcome escape from autonomy; it has even been argued that certain prisoners find in the rules of the institution an external support that serves to reduce their inner anxieties. The great majority of inmates, however, feel the restrictions as a painful attack on their self-respect. As some social scientists have noted in analyzing the concentration camp, prisoners stand in constant danger of losing their identification as adults; and the imprisoned criminal is likely to find himself reduced to the weak, helpless status of childhood.

Fourth, men (or women) in prisons are forced into intimate association with other criminals for months and years. Whatever this may mean for their reformation, one important consequence is that they are deprived of the sense of security that springs from living among men who can be expected to abide by the rules of society. Living with murderers and thieves may provoke anxiety in even the most hardened criminal.

Fifth, the inmate of the prison is an outcast. For a few criminals, such as professional thieves, society's rejection may have little significance. But the evidence suggests that in most cases the moral condemnation of society is a profound threat to the individual's sense of worth. He is reminded of his degradation in many ways—his uniform, his number—and is seldom allowed to forget that by committing a crime he has forgone his claim to be a member of society in good standing.

These are the common frustrations or deprivations to which men in prison must somehow respond. To some extent they can be dealt with through the patterns of criminal behaviour that the inmates have brought with them. Some of the amenities of life can be won by physical coercion, chicanery, or simple thievery: an important feature of life in many prisons is the exploitation of one prisoner by another.

Inmate social systems

Students of the prison community have noted the argot in which the inmates describe their social roles. In the U.S. the argot includes terms such as square Johns, right guys, outlaws, gorillas, wolves, punks, fags, hoosiers, and so on. The vocabulary varies, but in penal institutions of the same type the structure of roles is evidently very similar. In adult, male prisons for serious offenders, those who use violence or the threat of violence to take things from others are called gorillas or outlaws. Those who exploit their fellows by sharp dealing and skillful trading in the stream of illicit goods that moves beneath the surface of institutional life are termed merchants or politicians. Prisoners playing these exploitative roles may in turn be exploited by an inmate—the so-called rat or squealer—who reveals their activities to the prison officials in hope of preferential treatment.

Patterns of exploitation, however, are but one answer to a threatening social environment; it is also possible to neutralize some of the frustrations of prison life through solidarity. Mutual aid, group loyalty, reciprocal protection against the inquiries of officials, and a united front against the threat of coercion exerted by other, more aggressive inmates can lessen the rigours of imprisonment. Yet such solidarity is usually more verbal than actual, and the loyalties of men in prison are fragile at best. They have no common set of beliefs, for one thing. Friendship might provide a basis for a sort of cooperative alliance, but the inmates of prisons are usually failures at interpersonal relationships. Exploitation is more common than cooperation. Faced with severe frustrations, together but apart, ill-equipped by previous experience to live in harmony, the inmates of the prison exist by manipulating their captors, coercing and defrauding one another, or withdrawing in apathy.

Empirical studies of penal institutions are too few in number to show with any precision the amount of corruption, solidarity, or exploitation that exists in different types of institutions. It is also not clear to what extent the norms of the inmate social system are derived from the world outside and to what extent they are a special product of imprisonment. Enough studies have been made, however, to indicate that women's penal institutions differ from the picture presented above. The inmate social system in prisons for women is likely to be built around artificial families. Homosexual relationships between women taking the roles of "husbands" and "wives" or "fathers" and "mothers" play a major part in organizing inmate behaviour. There is less emphasis on custody and security, and the problem of work is much changed by the ready acceptance of institutional housekeeping as appropriate employment.

In jails, the inmate social system is likely to be less clearly developed than in prisons, many of the inmates simply withdrawing into themselves and awaiting release. Jails notoriously subject their inmates to unsanitary conditions, brutality, corrosive idleness, and poor management. They usually make no effort at rehabilitation, since most of the prisoners are serving short sentences or awaiting trial.

Juvenile reformatories commonly assign custody and security a lower priority than do prisons for adult males; their rehabilitative resources are greater, and opportunities for vocational training are frequently more numerous. The inmate social system takes different and more subtle forms than in adult prisons, but it would be too much to say that coercion and exploitation do not exist in juvenile penal institutions. Some degree of coercion is made likely by the fact that many young offenders suffer more severe personality disturbances; if close supervision is abandoned, on the ground that juveniles do not require it or that it is inimical to therapy, serious disorders may result.

It is frequently argued that penal institutions cannot be expected to succeed at the task of reformation, as shown by the fact that more than two-thirds of confined criminals commit new crimes after they are released. Others answer that this figure is grossly exaggerated, that with proper resources and personnel many of the problems of penal institutions could be overcome. They also argue that it is sentimental to assume that a great many of the criminals sent to prison can be reformed, since those most likely to be confined are the hardened offenders with long careers of lawbreaking. The debate as to the efficacy of prisons revolves upon a lack of concrete evidence. If an individual does not commit a crime after being imprisoned, has he reformed because of his experience in the penal institution or in spite of it? If he does commit a crime after being imprisoned, has the penal institution exercised a harmful influence or no influence at all? It is possible that the prison reforms but that its achievements are negated or undone by other forces that come into play when the offender returns to the community.

Whatever the merits of these arguments, the fact is that many criminals are dealt with outside of institutions. Any discussion of penology must take account of the trend away from institutional treatment.

CORRECTION OUTSIDE THE PRISON

Of the nearly 1,300,000 offenders handled by the penal systems of the U.S. each day more than half are under

Probation and parole

supervision in the free community. The great majority of these convicted criminals outside penal institutions are on probation rather than parole—that is to say, they have been placed under supervision in the community directly after conviction—rather than, as in the case of parolees, being first sent to prison and then given an early release. The number of offenders on probation will probably continue to increase relative to the number who pass through prison.

The theory underlying the use of probation and parole is that the best path to rehabilitation for most criminals lies in the community, where the offender can be reintegrated into society, and that the problems of the criminal are better dealt with in their social context than in the artificial environment of the penal institution. Probation and parole are also far less expensive, it is argued; the cost of supervising a youthful offender on probation, for example, is less than a tenth of what it costs to keep him in a training school.

The development of probation. Probation evolved from Anglo-American judicial practices worked out to avoid automatic application of the harsh penalties prescribed in the statutes. These practices originated in the early 1800s in response to humanitarian concern over the effects of imprisonment on the young or first offender. Magistrates in such cases imposed only token sentences, remanding the offenders to the custody of their parents or other guardians for more careful supervision. As the practice evolved, an offender could be released on his acknowledgment of obligation until some later date set by the court for final action. He was often required to post bond or deposit bail to guarantee his reappearance.

John Augustus, a Boston boot maker, initiated the practice he called "probation" when in 1841, through his interest in the temperance cause, he appeared in court and posted bail for a man charged with drunkenness. Successful in helping the man get established, he proceeded to assist hundreds of men, women, and children charged with various offenses.

The voluntary work of Augustus and his successors led the Massachusetts legislature to authorize in 1878 the appointment of a paid probation officer to serve the Criminal Court of Boston. Subsequent legislation made probation services mandatory in most Massachusetts courts. Other states were slow to follow the example until the turn of the century, when the children's court movement began to sweep the U.S. In England, the efforts of temperance volunteers in the police courts paralleled the development of probation in the U.S. On the continent of Europe progressive elements had also long supported the mitigation of punishment for juvenile and first offenders. Europeans were primarily interested in a substitute for short-term imprisonment rather than in extended probation services. They adopted a procedure whereby sentence was imposed but contained provision for suspending execution of the punishment subject to the offender's good behaviour. First authorized by Belgian law in 1888, the conditional sentence was adopted by France shortly thereafter and eventually spread to nearly every European country.

Most continental European countries have delegated the personal supervision of probationers to social welfare organizations dependent on volunteer services. A general trend toward professional probation services has appeared in the Belgian, French, Dutch, and Swedish systems, where trained consultants are responsible for indoctrinating volunteer agents, supervising their work, and in many cases giving direct services to probationers. The move toward professionalization has been accompanied by increased central government responsibility for coordination, regulation, or administration of probation services. In most English-speaking countries probation is under the supervision of professional probation officers.

Problems of supervision. Seldom, however, do the courts and prosecutors adhere to all of the principles that are considered essential to the proper adminstration of probation: (1) careful selection of cases amenable to supervision, with consideration given on the basis of the offender's likelihood to improve under treatment, rather than the seriousness of his offense; (2) suspension of sentence for persons so selected; (3) supervision by professionally trained personnel; and (4) release of the probationer at the end of a specified period set by the court, provided his behaviour has been satisfactory, or revocation of his probation if he has committed a new crime or failed to comply with the conditions of his probation. It is widely agreed that a probation or parole officer should supervise no more than 35 to 40 persons at a time, but case loads are rarely that small. Successful supervision also requires a professional staff, a wide range of community services and facilities, and careful administration.

The administration of probation and parole must guard against two types of error. On the one hand, offenders who should be incarcerated may be set at large in the community; on the other, offenders who can be rehabilitated in the community and present no danger may be kept locked up when they should be free.

An accurate means of predicting criminal behaviour would be invaluable in deciding when offenders should be released from prison on parole and which offenders should be placed on probation rather than in a penal institution. The efforts of criminologists to develop statistical methods for predicting criminal behaviour have not succeeded in providing officials with practical guidance. Aside from the scientific difficulties in making accurate predictions, parole and probation officials often are suspicious of statistical formulas. The administrators of penal systems, for reasons of training and intellectual orientation, are far more receptive to clinical judgments, informed guesses, and the "wisdom of experience." As a consequence, prediction tables have generally remained an adjunct to the decision-making process, when they have been used at all.

Among the services needed for the supervision of the offender in the community, job training and job placement are viewed by many penologists as of prime importance. The problem of employment for convicted offenders is not easily solved, since they tend to come from minority groups and the poor, have little education, and often suffer from behaviour disorders, mental retardation, and poor health. To these disabilities must be added the stigma of the criminal that makes many potential employers turn them away. Penologists have given increasing attention to the importance of working with the family of the offender, particularly in the case of juveniles.

TRENDS IN PENOLOGY

There are many indications that modern societies are at the beginning of a significant transformation in their penal systems. Imprisonment as a principal method of handling the convicted criminal has proved to be costly, and its effectiveness has been increasingly questioned. Theories of the causation and prevention of crime now place less stress on moral factors that might be dealt with by awaking penitence or fear in the criminal. The segment of the population from which the majority of convicted criminals are drawn has ceased to be regarded as an alien, subhuman lower class, and the demands for just and humane treatment have increased.

The need for professionally trained people in the field of correction is recognized; there is a growing demand for the services of social workers, psychologists, sociologists, and persons with a background in public administration. The field is extremely diverse, and it cannot be said that there is a clear-cut correctional career; but the cynicism and sense of futility so common in the past are diminishing. It seems probable that in the near future penal institutions will come to be used to a greater extent as diagnostic and classification centres rather than as places of long-term confinement. The variety of institutions will continue to grow; there will be more use of temporary detention homes and halfway houses where custody and security no longer take first precedence. Finally, it is likely that institutional treatment will diminish in importance in comparison with supervision in the community.

BIBLIOGRAPHY. H.E. BARNES and N.K. TEETERS, *New Horizons in Criminology*, 3rd ed. (1959), an excellent general

textbook in criminology, with strong emphasis on the history of penology; DONALD CLEMMER, *The Prison Community* (1940), the first detailed sociological study of an American prison; R.A. CLOWARD *et al.*, *Theoretical Studies in Social Organization of the Prison* (1960), a discussion of important efforts to link the analysis of prison systems to sociological theory; AMERICAN CORRECTIONAL ASSOCIATION, *A Manual of Correctional Standards* (1956), a description of the needs of corrections as an emerging form of professional work; J.P. CONRAD, *Crime and Its Corrections: An International Survey of Attitudes and Practices* (1965), one of the few books that attempts to provide a comparative approach to penology, although largely limited to Europe and the Scandinavian countries; D.R. CRESSEY (ed.), *The Prison: Studies in Institutional Organization and Change* (1961), also presents important efforts at linking the analysis of prison systems to sociological theory; EUROPEAN COMMITTEE ON CRIME PROBLEMS, COUNCIL OF EUROPE, *The Effectiveness of Punishment and Other Measures of Treatment* (1967), a statistical view of what is known about the effect of penal sanctions; R. GIALLOMBARDO, *Society of Women* (1966), one of the few full-length studies of a penal institution for women; D. GLASER, *The Effectiveness of a Prison and Parole System* (1964), a comprehensive survey of prisons in the United States; L.W. MCCORKLE, A. ELIAS, and F.L. BIXBY, *The Highfields Story: An Experimental Treatment Project for Youthful Offenders* (1958), a lucid account of a juvenile institution organized around the concept of therapy; P.W. TAPPAN, *Contemporary Correction* (1951), a systematic review of the range of modern techniques used in dealing with convicted criminals; G. RUSCHE and O. KIRCHHEIMER, *Punishment and Social Structure* (1939), a scholarly effort to uncover the causes of variation in penal sanctions; TASK FORCE ON CORRECTIONS, PRESIDENT'S COMMISSION ON LAW ENFORCEMENT AND ADMINISTRATION OF JUSTICE, *Task Force Report* (1967), an invaluable source of recent data on the treatment of criminals in the United States; G.K. STURUP, *Treating the "Untreatable": Chronic Criminals at Herstedvester* (1968), a description of a modern, small institution in the forefront of psychiatric treatment; G.M. SYKES, *The Society of Captives: A Study of a Maximum Security Prison* (1958), another attempt to link sociological theory to the analysis of a prison system; GUSTAVE DE BEAUMONT and ALEXIS DE TOCQUEVILLE, *Du système pénitentiaire aux États-Unis, et de son application en France* (1833; Eng. trans., *On the Penitentiary System in the United States and its Application in France*, new ed., 1964), important primarily for its historical interest and its impact on European penal philosophy.

(G.M'C.S.)

Probability, Theory of

A description of probability that was first presented by the French mathematician Pierre-Simon Laplace (*q.v.*) in 1795 reads as follows: If an experiment (or operation or trial) can produce a number of different and equally likely outcomes, some (but usually not all) of which can be considered favourable, then the probability p of a favourable outcome is defined as the ratio of the number of favourable outcomes to the total number of possible outcomes. The most familiar example of such an experiment is the tossing of a coin. If the call is heads, then of the exactly two possible outcomes, a head or a tail, one is favourable to the caller. Provided the coin is fair, the probability of success for him is thus $\frac{1}{2}$. If a player is to bet that an ordinary cubical die will fall with the one spot uppermost, he wishes to know the odds, or probabilities. The assumption is that there are six equally likely ways in which the die can come to rest, only one of which is favourable; hence the probability of throwing a one is $\frac{1}{6}$.

While Laplace's description of probability cannot be considered a mathematical definition because it depends on the non-mathematical idea of "equally likely outcomes," it can be used as a guide in the calculation of many simple problems. An acceptable mathematical definition was formulated in the 20th century (see below *Foundations of 20th-century probability theory*).

The requirement that all the possible outcomes be equally likely is of major importance. There may be, for example, three cards, one red on both sides, one red on one side and white on the other, and the third white on both sides. After shuffling them in a closed bag, the experimenter reaches into the bag, draws out one card, and without looking at it, lays it on the table. Only then does he observe that the visible (upper) side of the card he has drawn is white. The problem is to find the probability that among all such experiments the underside is red. It is certain that the card in view is not the red-red card. Therefore it is either the white–red card or the white–white card. In the one case the underside is red; in the other case it is white. It follows that the face that is showing is one of the three white faces and, because of the way the cards are shuffled and drawn, each is equally likely. For only one of the three is the other side red, resulting in a probability of $\frac{1}{3}$.

Inadequacy of Laplace's definition

Unfortunately, an attempt to define what is meant by the phrase "equally likely" leads into a logical trap, and for this reason Laplace's approach is mathematically inadequate. In the coin-tossing experiment it can be assumed that the coin is unbiassed; that is, it is just as likely to fall heads as tails. If it is said, however, that this means that the two probabilities are equal, the reasoning is circular. The experiment is so simple and the result so obvious that the dilemma is concealed or overlooked. To define probability empirically in some such way as this might be tried: "Toss a coin a large number of times and divide the number of heads that appear by the total number of tosses. If this quotient approaches a fixed limit as the number of tosses is increased, call that limit the probability of obtaining a head in one toss." The difficulty with this definition is that it is not well suited for calculation. It is at the root, however, of the mathematical concept of probability. Along with a modified form of Laplace's description, it has been incorporated into a basic law of probability, "the law of large numbers," that essentially identifies probability in the Laplace sense (stripped of the "equally likely" ambiguity) with this limit of a quotient.

Probability on finite-dimensional spaces

ELEMENTARY PROBLEMS

An important principle of probability theory concerns the concept of statistical independence and is suggested by the statement that "the coin doesn't remember." Even if there has just been a run of 10 heads in a sequence of tosses of a fair coin, the probability that the next toss will be a head is neither greater nor less than usual: it is still exactly $\frac{1}{2}$. After a long run of red at roulette, if the wheel is fair, red is exactly as likely as black to win the next play. A formalization of this principle states that two events, such as outcomes of tosses of a coin or spins of a wheel, are statistically independent if and only if the probability of their joint occurrence is the product of their separate probabilities.

Statistical independence in dice problems

If two dice are tossed, an interesting problem is to determine the probability that they will both turn up ones. Resorting to Laplace's approach, it is found that there are altogether 36 outcomes of this experiment: six ways for the first die to fall, and each of these paired with six ways for the second. Only one of all these shows two faces turned up ones so that the probability is $\frac{1}{36}$. Another way to solve the problem using the principle of statistical independence is to think of the outcomes of two dice as two statistically independent events. The probability of an ace in the first throw is $\frac{1}{6}$; that is, using the law of large numbers referred to above, in a large number of throws $\frac{1}{6}$ of the throws will be successes and $\frac{5}{6}$ failures. A product of two such separate probabilities is $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. In general, if two independent events have probabilities p_1 and p_2, respectively, then the probability of both occurring jointly is p_1p_2.

In another problem the 13 cards of one suit in a 52-card deck are shuffled together. The probability of drawing first the ace and then the king in exactly two draws is desired. The probability of the ace on the first draw is $\frac{1}{13}$, but now the pack has been reduced to 12 so that the probability of the king on the second draw is $\frac{1}{12}$. It is not possible to solve the problem with the principles introduced thus far because the second drawing of the card was not independent of the first. The principle of conditional probability can be introduced in which the probability of two events occurring jointly can be calculated as the probability of one event times the conditional probability of the second event given the first. In this example, the desired probability becomes $\frac{1}{13} \times \frac{1}{12} = \frac{1}{156}$.

Next, the probability of drawing the ace and the king in either order from the 13 cards of one suit is desired. Stated another way, two cards are picked, face down, from the 13 cards. The probability that they are the ace and the king is needed. Problems of this kind are most simply answered in terms of permutations and combinations, which are discussed under COMBINATORICS AND COMBINATORIAL GEOMETRY. The total number of ways (combinations) of drawing two cards from the 13 cards is 78, of which only one consists of the ace and the king (remembering that order does not matter). Therefore $p = 1/78$.

The birthday problem

There is an amusing and instructive probability question that has become known as the "birthday problem." Among 24 people chosen at random (say, a dozen couples at a party), it is difficult to estimate the probability that some two of them will have the same birthday. Contrary to intuition, the probability is slightly larger than ½ and rises rapidly if the number (at the party) is increased. The probability that some two among 40 people will have the same birthday is about 8/9. The proof illustrates a principle upon which the solution of many probability problems is based: if p is the probability of an event, then $1 - p$ is the probability of the negation of that event. In the stated problem the ways in which it is possible for no two of the 24 people to share a birthday can be examined. For simplicity, anyone born on February 29 may call his birthday February 28. There are, consequently, 365 days on which the first person in the party can have a birthday, and the probability that his birthday is on some day is certain, or $^{365}/_{365}$. The conditional probability that the second person has a birthday on some other day, given that the first birthday is specified, is $^{364}/_{365}$. The conditional probability that the third person has a birthday on some day other than these two, given that the first two birthdays are specified, is $^{363}/_{365}$, and so on. Thus, using the principle of conditional probability, the probability that no two have the same birthday is $q = 365 \cdot 364 \cdot 363 \cdot \cdots 342/365^{24}$. It is this fraction that is slightly less than ½ so that $p = 1 - q$ is slightly greater than ½.

If six coins are tossed simultaneously (or one after the other), it is of interest to know the probability that just three will show heads. The possible successful outcomes are the ways of finding three heads among six coins, or the combinations of six things taken three at a time, which is 20. The total number of outcomes is the totality of ways of tossing six coins, or 2^6. Hence, the required probability is $20/2^6 = 5/16$.

If a perfectly fair coin were tossed a large number of times, say, 1,000, it would be difficult (though possible) to calculate the probability of exactly 500 heads. If two coins are tossed ($n = 2$), the probability of half heads is ½, but this probability is never again so high and decreases indefinitely with increasing n. For $n = 6$ it is only 5/16, as calculated above. The probability of obtaining exactly half heads can be made arbitrarily small by tossing a sufficiently large number of coins. The intuitive reason for this is, of course, that there are many ways of tossing almost exactly half heads when n is large. Mathematically, the reason is concerned with the coefficients in a binomial expression, $(a + b)^n$, that is written as a sum of powers of a and b. A deep understanding of these facts was obtained in the first quarter of the 18th century and led to approximations of the binomial coefficients in terms of what is now known as the probability density of the normal distribution. The normal plays a central role in modern probabilistic and statistical theory.

BASIC CONCEPTS

Contributions of Pascal and Fermat

In the early contributions to mathematical probability theory, primarily those of the French mathematicians Blaise Pascal (*q.v.*) and Pierre de Fermat (*q.v.*) and the Dutch mathematician, astronomer, and physicist Christiaan Huygens (*q.v.*), the main consideration was the expected gain in games of chance and considerations other than the direct calculation of probability itself. In Huygens' treatise *De Ratiociniis in Ludo Aleae* (1657; "On Ratiocination in Dice Games") there appeared the first documented reference to the concept of mathematical expectation. Huygens' formulation was if, among N cases all having equal chances of occurring, a player can gain the amount a in n of them and the amount b in the rest of them, then his expectation is given by $[na + (N - n)b]/N$. Historically, calculation concerning expectation came first, and these calculations implied secondary consequences about chance. Though the presentation of the subject is now frequently in the other order, probability first and expectation second, probability is logically a special case of mathematical expectation. To illustrate, it is possible to let C be a symbol to designate a particular collection of n equally likely events (or cases) out of a total of N. In the language of mathematical expectation, a player may receive amount 1 if any of the events composing C occur and the amount 0 otherwise. For such a game the expectation according to Huygens becomes $[n \cdot 1 + (N - n) \cdot 0]/N = n/N$, which is precisely the chance (or probability) for some one of the events composing C to occur. In modern notation if X is the amount that a player in such a game is to receive (X assuming the values 1 or 0 in the contingencies described), then $\mathbf{E}X = \mathbf{P}C$, showing the basic relationship between mathematical expectation, **E**, and probability, **P**.

Mathematical expectation. As an example from Huygens in the use of his concept of expectation, the following game is useful. Two dice are rolled. If a seven is thrown, player A wins the amount a; if a ten is thrown, player B wins the amount a; and if any other number is thrown, the amount a is divided evenly between the two players. The objective is to show that the ratio of the expectations of the two players is 13/11. The symbols "X_A" and "X_B" may be used to stand for the amounts that players A and B will receive, respecitvely. (The same convention will be used also in another example given below.)

Contribution of Huygens

There are six out of 36 equally likely ways to roll a seven, three out of 36 equally likely ways to roll a 10, and a remaining 27 ways for the dice to fall. From Huygens' definition of expectation, it follows that the expectations of A and B, respectively, are $13a/24$ and $11a/24$ (see Box, formula 1) so that the ratio of the two is 13/11.

The following is another example: Two players, A and B, play with a pair of dice on the condition that A is to win the stake of amount b if he throws a six before B throws a seven and B is to win the stake b if he throws a seven before A throws a six. The rules of the game specify that they are to throw alternately, with A beginning. It is desired to find the expectations of A and B.

It is possible to fix attention on B's expectation and to note that it will have one value when A is about to throw and another value when B is about to throw. These two expectations of B's can be designated x and y, respectively. When A is about to throw the chances for each player to win and their expectations are the same as at the beginning of the game, and whenever B is about to throw the two players' chances and expectations are as they were just before the second toss. Whenever A is about to throw, then, there are five cases in which B can win nothing (A throws a six), but there are 31 cases (those in which A would fail to throw a six) in which B can win the amount y, namely, the expectation that accrues to him when his throw comes up. Thus, according to Huygens' definition of expectation, B's expectation whenever A is about to throw is expressible in terms of y (see 2). By the same reasoning, whenever B is about to throw there are six cases in which he can win b and 30 cases in which he can win the amount x, namely, the expectation that accrues to him when A throws. Thus y is

$$(1)\quad \begin{cases} \mathbf{E}(X_A) = (6a + 27a/2)/36 = 13a/24 \\ \mathbf{E}(X_B) = (3a + 27a/2)/36 = 11a/24 \\ \text{in which } \mathbf{E} \text{ stands for "expectation of"} \end{cases}$$

$$(2)\quad x = (5 \cdot 0 + 31 \cdot y)/36$$

expressible in terms of x (see 3). There are two linear equations in two unknowns x and y, b being a known constant. Their solution is obtainable by elementary methods (see 4). Only the x value is needed because, if B's expectancy is x, A's must be $b - x$, and the desired expectations are determined (see 4). In this problem in which some one player gets all the stakes and in which there is no case in which the stake is divided, expectation is proportional to probability with proportionality constant b (see 4). This proportionality is a guide in the derivation of probabilistic laws below.

The additivity assumption. Alternate methods of solution to some of the same problems that occupied Huygens were later given by the Swiss mathematician Jakob Bernoulli (see BERNOULLI FAMILY) in his famous *Ars Conjectandi* ("The Conjectural Arts"), written in 1705 and published in 1713. By Bernoulli's reasoning player A in the above problem can win in any one of an infinite number of ways: six on the first toss; or no six on the first toss, no seven on the second toss, and six on the third toss; and so forth. A bar over a word can be used to symbolize its negation and a dot between words to symbolize succession of events; thus, a list of these separate ways for A to win reads as follows:

Contribution of Jakob Bernoulli

six; $\overline{\text{six}} \cdot \overline{\text{seven}} \cdot$ six; $\overline{\text{six}} \cdot \overline{\text{seven}} \cdot \overline{\text{six}} \cdot \overline{\text{seven}} \cdot$ six; $\cdots$.

The number of distinct configurations can be counted in the successive rolls of two dice that would lead to each of these ways for player A to win (see 5). These numbers, when taken relative to the total number of possible configurations of one roll of two dice, three rolls of two dice, five rolls of two dice, etc., yield a list of probabilities for all the separate ways in which player A can win, the sum of which should be the total probability (or chance) for A (see 6), the symbol "A" also being used in this context to stand for the event that player A wins. Applying the formula for the sum of a series yields $^{30}/_{61}$ as the chance for A to win (see 7). A similar line of reasoning for player B yields $^{31}/_{61}$ as the chance for B to win.

In Huygens' use of the definition of expectation he allowed that the amounts that could be gained by a player under various circumstances might well be, themselves, expectations. In fact, it was this flexibility of interpretation that led to the two simultaneous equations of his solution. If it is observed further that in some cases, notably in the problem Huygens was concerned with, that expectation is proportional to probability, an equation applicable to a variety of cases can be found that involves products of probabilities. To be precise, either of the two simultaneous equations of Huygens' solution (see 2, 3) involves probabilities instead of expectations when both sides of the equation are divided by the amount a (see 8). The resulting analysis of a probability into additive and multiplicative parts definitely entails an assumption about the ways in which probabilities can be composed. Such an assumption would be as follows:

If an event A, the chance of which occurring is $\mathbf{P}(A)$, will be realized in the presence of either of two mutually exclusive events, A_1 and A_2, with corresponding probabilities $\mathbf{P}(A_1)$ and $\mathbf{P}(A_2)$; and if the realization of A in the presence of A_1 has probability $\mathbf{P}(A|A_1)$ and in the presence of A_2 has probability $\mathbf{P}(A|A_2)$, then the relationship (see 9) between the probabilities of these events identifies $\mathbf{P}(A)$ with the sum of the two products $\mathbf{P}(A|A_1)\mathbf{P}(A_1)$ and $\mathbf{P}(A|A_2)\mathbf{P}(A_2)$.

Though Huygens did not formulate the statement as precisely as this in terms of probabilities, his thinking, as well as that of probabilists to follow him, was consistent with such a formulation, which will be used in the discussion below of Huygens' important "fifth problem."

Bernoulli, in achieving his solution to the problem already presented above (see 6), made the infinite addition of probabilities a plausible step in his solution. Bernoulli assumed that if an event A can be realized in any one of a possibly infinite number of mutually exclusive ways, $A_1, A_2, A_3, \cdots$, then the probability $\mathbf{P}(A)$ and the probabilities $\mathbf{P}(A_1), \mathbf{P}(A_2), \mathbf{P}(A_3), \cdots$ are related by identifying $\mathbf{P}(A)$ with the sum of the latter (see 10).

(3) $y = (6b + 30x)/36$

(4) $\begin{cases} \mathbf{E}(X_A) = 30b/61 \\ \mathbf{E}(X_B) = 31b/61 \end{cases}$

(5) $5,\ 31 \cdot 30 \cdot 5,\ 31 \cdot 30 \cdot 31 \cdot 30 \cdot 5, \cdots$

(6) $\begin{cases} \mathbf{P}(A) = (5/36) + (31 \cdot 30 \cdot 5/36^3) + \\ \qquad + (31^2 \cdot 30^2 \cdot 5/36^5) + \cdots \\ \text{in which P stands for "chance of"} \end{cases}$

(7) $\begin{cases} \mathbf{P}(A) = (5/36)[1 + 31 \cdot 30/36^2 + \\ \qquad + 31^2 \cdot 30^2/36^4 + \cdots] \\ \qquad = (5/36)[1 + (31 \cdot 30/36^2) + \\ \qquad + (31 \cdot 30/36^2)^2 + \\ \qquad + (31 \cdot 30/36^2)^3 + \cdots] \\ \qquad = (5/36)[1 - 31 \cdot 30/36^2]^{-1} = 30/61 \end{cases}$

(8) $p = p_1p_2 + p_3p_4$

(9) $\mathbf{P}(A) = \mathbf{P}(A|A_1)\mathbf{P}(A_1) + \mathbf{P}(A|A_2)\mathbf{P}(A_2)$

(10) $\mathbf{P}(A) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) + \cdots$

It was a Soviet mathematician, A.N. Kolmogorov (*q.v.*), who formulated this assumption, called "the additivity assumption," of the addition of an infinite number of probabilities as an unqualified postulate in his *Grundbegriffe der Wahrscheinlichkeitsrechnung* (1933; Eng. trans., *Foundations of the Theory of Probability*, 1950), but it is important to notice that when he did so he said, "It is almost impossible to elucidate its empirical meaning." Notwithstanding the fact that the assumption has now been cast into a basic postulate of measure-theoretic probability theory, there are later probability problems (particularly in reference to chance phenomena as they evolve on computing machines) in which a Bernoulli-like operation would lead one into contradiction (see AUTOMATA THEORY: *Computable probability spaces*). Such problems may arise when a computing machine generates an infinite (but countable) sequence of random numbers, each one of which has an associated probability of 0 but all of which combine to make an event with probability of 1. Interestingly enough, it follows that the acceptability or unacceptability of the additivity assumption in unqualified form rests upon the tacit assumption that random phenomena are not generated by automata and is, thus, allied with the tenability or untenability of a theory of nature in which chance and causality are distinct. That is, the Bernoulli postulate, as stated in modern terms by Kolmogorov, is in keeping with the classical bifurcation of the Greek goddesses Tyche (chance) and Moira (fate), which still permeates scientific thinking, rather than with the postulates of computability that have been formulated in the 20th century.

Huygens' fifth problem

As an illustration of the use of products of probabilities, Huygens' famous fifth problem will be considered. This problem evolved through various generalizations and extensions into the problem of gambler's ruin, duration of play, simple random walk, and many others that have occupied many mathematicians on different occasions from the time of Huygens to the middle of the 20th century. In its simplest original form, it is as follows:

Two players each take 12 counters and play with three dice on the condition that if 11 is thrown the first player gives a counter to the second and if a 14 is thrown the second player gives a counter to the first. That player wins the game who first obtains all the counters. The chances of each player to win are desired.

A difference equation from which will follow the desired chances for the first player to win can be deduced

from the composition of probabilities developed above. If $P(n)$ is the chance for the first player to win when he has n counters (or, in alternate convenient language, when the game is in state n), then in the next play of the game following state n the first player must either give or receive a counter, and these eventualities have the corresponding probabilities $27/216$ and $15/216$, respectively (as an enumeration of the various ways three dice can fall easily shows). In the terminology of the above assumption involving products of probabilities (see 9), these two eventualities comprise two mutually exclusive events, and the event that the first player will win will be realized in the presence of either one or the other of the two with corresponding probabilities $P(n-1)$ and $P(n+1)$. A direct application of the above reasoning yields an equation in which $P(n)$ is expressed in terms of $P(n-1)$ and $P(n+1)$. Such an equation, called a "difference equation," is first supplemented with boundary values of $P(n)$ for values of $n=0$ and $n=24$ (see 11). It is then solved by known methods (see NUMERICAL ANALYSIS), and the chance of the first player winning is obtained by setting $n=12$ in $P(n)$.

Conditional probability. There is an advantage in noticing that the term $\mathbf{P}(A|A_1)\mathbf{P}(A_1)$ used above is the probability for both A and A_1 to occur, with a similar meaning holding for the second term, $\mathbf{P}(A|A_2)\mathbf{P}(A_2)$. This fact is an immediate consequence of the rule on the addition of probabilities (see 10) applied to two mutually exclusive events, namely, the events "A and A_1" and "A and A_2." For any two events A and A_1, the probability for the occurrence of both A and A_1, $\mathbf{P}(AA_1)$, and the product, $\mathbf{P}(A_1)$ times $\mathbf{P}(A|A_1)$, are equal (see 12), the latter being the probability for the occurrence of A in the presence of A_1. In this statement the events A and A_1 are any events and may be interchanged without changing the meaning (see 13). In contemporary terminology, $\mathbf{P}(A|A_1)$ is the probability of A conditioned on the occurrence of the event A_1 or, more briefly, the conditional probability of A given A_1.

Contribution of De Moivre

Following Bernoulli, the work of a French-born British mathematician, Abraham De Moivre, explicitly introduced the concepts of statistical independence and conditional probability. Though probabilistic statements such as the above are implicit in the writings of Huygens and Jakob Bernoulli, it was unequivocally presented in De Moivre's *The Doctrine of Chances* of 1718. In the symbolism $\mathbf{P}(A|A_1)$, used above, De Moivre's definition becomes equivalent to its modern formulation:

Two events A and A_1 are statistically independent if and only if the conditional probability of A given A_1 is equal to A and the conditional probability of A_1 given A is equal to A_1 (see 14). Equivalently, A and A_1 are statistically independent if and only if $\mathbf{P}(AA_1)$ is equal to the product of the separate probabilities (see 15).

It is interesting to see that in the same way that Pascal and Huygens derived many of their results on probability from a formulation predicated upon the more intuitively immediate concept of expectation, so also De Moivre derived the relation of independence from a rule on the composition of expectations:

> If the obtaining of any sum (amount) requires the happening of several events that are independent of each other, then the value of the expectation of that sum is found by multiplying together the several probabilities of happening, and again multiplying the product by the value of the sum.

As an important corollary to the above principle, De Moivre adds:

> If we make abstraction of the value of the sum to be obtained, the bare probability of obtaining it will be the product of the several probabilities of the (independent) events.

De Moivre also pointed out, but less clearly, that the principle of composition of probabilities for independent events can be extended to more than two events. A modern formulation is the following:

The events $A_1, A_2, A_3, \cdots$, are mutually independent if and only if every pair, as well as every pair of single or joint occurrences, B and C, is independent (see 16). For example, B or C in the above definition could be chosen as A_1 or as A_5 or as "A_1 and A_5" or as "A_2 and A_7 and A_9," or as any joint occurrence of one or more events. For N mutually independent events $A_1, A_2, \cdots, A_N$, it follows from the last definition that, in particular, $\mathbf{P}(A_1A_2\cdots A_N)$ is the product of the probabilities for separate events (see 17).

$$(11)\quad \begin{cases} P(n)=(27/216)P(n-1)+ \\ \qquad +(15/216)P(n+1), \qquad n=1,2,\cdots,23 \\ P(0)=0, \qquad P(24)=1 \end{cases}$$

$$(12)\quad \mathbf{P}(AA_1)=\mathbf{P}(A|A_1)\mathbf{P}(A_1)$$

$$(13)\quad \mathbf{P}(AA_1)=\mathbf{P}(A_1|A)\mathbf{P}(A)$$

$$(14)\quad \mathbf{P}(A|A_1)=\mathbf{P}(A) \text{ and } \mathbf{P}(A_1|A)=\mathbf{P}(A_1)$$

$$(15)\quad \mathbf{P}(AA_1)=\mathbf{P}(A)\mathbf{P}(A_1)$$

$$(16)\quad \begin{cases} \text{The events } A_1, A_2, A_3, \cdots \text{ are mutually independent} \\ \text{if and only if} \\ \mathbf{P}(BC)=\mathbf{P}(B)\mathbf{P}(C), \\ \quad B=A_iA_j\cdots A_k, \qquad C=A_mA_n\cdots A_p, \\ \text{for any choice of distinct subscripts } i,j,\cdots,k, \\ m,n,\cdots,p \text{ from the integers } 1,2,3,\cdots. \end{cases}$$

$$(17)\quad \mathbf{P}(A_1A_2\cdots A_N)=\mathbf{P}(A_1)\mathbf{P}(A_2)\cdots\mathbf{P}(A_N)$$

FOUNDATIONS OF 20TH-CENTURY PROBABILITY THEORY

Toward the end of the first quarter of the 19th century most of the significant work on the theory of probability had been summarized by Laplace in his great *Théorie analytique des probabilités* (1812; "Analytic Theory of Probability"). It was in Laplace's introduction to this work on chance that the striking pronouncement for causality was found. For Laplace, probability theory was a device for coping with reality, in view of the inadequate nature of man's knowledge, and his reasoning was entirely mechanistic:

Contribution of Laplace

> We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence. Its discoveries in mechanics and geometry, added to that of universal gravity, have enabled it to comprehend in the same analytical expressions the past and future states of the system of the world.

With the beginning of the 20th century, striking discoveries in the physical sciences and in mathematics were responsible for basic innovations in probability theory. The statistical mechanics of the German and U.S. physicists Ludwig Boltzmann and J. Willard Gibbs (*q.v.*) expressed in more far-reaching fashion Laplace's conviction that the approximate nature of measurement was responsible for chance phenomena. The Boltzmann ergodic hypothesis, which identifies two types of averages, as does the weak law of large numbers, and Gibbsian phase space (appropriate for the statistical description of gases) provided the physical starting point from which, in 1923, the U.S. mathematician Norbert Wiener (*q.v.*) formulated the first mathematically precise description of Brownian motion (the motion seen in a microscope to be conveyed to tiny particles by the action of surrounding molecules). The mathematical tools used by Wiener were those of the

Contribution of Wiener

(18) A function f defined on S to R is interval measurable if and only if it possesses the following two properties:
(i) The function f has at most a countable number of discontinuities,
(ii) The sets $\{w:f(w)=t\}$ that are inverse images of points in the range of f are interval measurable.

(19) $$\mu(A)=(b_1-a_1)+(b_2-a_2)+\cdots+(b_n-a_n)+\cdots$$

French mathematician Henri-Léon Lebesgue and included measure theory, a formal extension of the idea of length (see below *The measure-theoretic reasoning*). While Wiener's analytical approach was being placed on a postulational basis by Kolmogorov in 1933, the German and Austrian quantum theorists Werner Karl Heisenberg and Erwin Schrödinger (*qq.v.*) discovered that probability was entering in the description of atomic phenomena. The measure-theoretic ingredients of probability, known to the mathematicians, did not seem to be a part of the quantum theory. This apparent paradox continues to inspire mathematical research into the nature of probability theory even today, and for that reason introductory remarks will be made about each approach, the measure-theoretic and the quantum-theoretic.

For other approaches to probability including the frequency theory of probability, the range theory of probability, and the belief theory of probability, see MATHEMATICS, HISTORY OF: *Various mathematically based probabilities.*

The measure-theoretic reasoning. The present discussion will be concerned with functions f defined on the space S composed of the unit interval, such that each set in S, constituted of all points that are mapped by f into the same real value, is of simple enough structure and provided there are not too many such sets. (Both these qualifications are to be defined more rigorously.) In physical language, such a function f corresponds to a measurable property of a system. A function f is thus being used to label sets in S, and it is assumed that the values that serve as labels are real numbers. It is seen that distinct values of f correspond to disjoint sets of points in S. That is to say, f induces a partition of S in the following sense: For each function f and for each real number r there is a set $A^{(r)}$ of points in S, such that $f(a)=r$ if and only if $a \in A^{(r)}$; moreover, each point of S will belong to one and only one $A^{(r)}$. Now if all investigations could be confined to one function and to simple statements involving it, then the sets $A^{(r)}$ of the partition induced by f would be as large a class of sets as necessary. On the contrary, it will be found necessary to consider many functions, and, once these are available, it will be necessary to observe complicated functions of these. This situation means that not only the partition induced by one f but the sets composing many partitions, as well as various combinations of these, must be considered. It is not enough to define the functions individually, but they must be defined as a class, and strict cognizance must be taken of the logical structure of the sets induced on S by this class. In order to do this, it may be convenient first to establish a class of sets in S and afterward to define a function f in reference to this class. It is this class of sets in S that it is desirable to define in more detail and use as an essential mathematical ingredient in the study of probability theory.

In the above approach to a class of sets in S it has not been specified what is meant by "simple structure" and "not too many." With some arbitrariness the conditions will be imposed that the sets of the class be simple in the following sense: they are to be built out of intervals with end points a, b contained in (the closure of) S. By the term interval is understood either open, closed, or half-open interval. It can now be made specific what is meant by an interval-measurable class of sets in the real line R. (An equivalent definition holds for the interval-measurable class of sets in any interval subset of R, such as S.) The interval-measurable class of sets in R is the class $\mathfrak{F}$ composed of the empty set and all sets in R that can be partitioned into at most a countable number of intervals in R. In this definition it is understood that a set A is partitioned into subsets $A_1, A_2, \cdots, A_n$ if and only if each point of A is a member of one and only one A_k and each point of any A_k is a member of A. In order to extend measurable concepts to functions, it is necessary to use the concept of an inverse image of a set. The inverse image of a set A, as mapped through f, is composed of the set of points a, such that $f(a)$ belongs to A.

Interval measurable sets and functions

A function f defined on S to R is interval measurable if and only if for any set A in $\mathfrak{F}$ the inverse image of A, as mapped through f, is in $\mathfrak{F}$. A basic theorem can be proved that gives deeper insight into the meaning of interval-measurable functions than the definition does (see 18).

The definition of $\mathfrak{F}$, in this simplified context, can be used as the first building block for the more complicated idea of the class of Lebesgue measurable sets, named after Lebesgue. The fact will later be used that most of the more complicated sets required by probability theory can be approximated to arbitrary accuracy (assuming an appropriate definition of approximation is given) by sets in $\mathfrak{F}$. Thus, $\mathfrak{F}$ will be considered intuitively meaningful and at the same time representative of abstractly defined classes of sets. Similarly, an interval-measurable function has explicit meaning for probability theory and is close to the idea of a random variable. According to a reasonable criterion, most of the more complicated functions of probability theory can be approximated to arbitrary accuracy with the interval-measurable functions defined here.

The Lebesgue measure $\mu(A)$, defined on sets A in $\mathfrak{F}$, is simply length in the case where A is an interval and sums of lengths in the case where A is partitioned into a countable number of intervals. Thus, in general, if A is partitioned into subsets $A_1, A_2, \cdots, A_n, \cdots$, a subset A_k being an interval from a_k to b_k, the Lebesgue measure of A is defined to be the sum of differences of the type $(b_k - a_k)$ (see 19).

Though S has been identified with the unit interval, a similar definition for a measurable function, with a more general class of sets replaced for $\mathfrak{F}$, would be meaningful if no topology had been imposed to define distance between points in S. The freedom from a topology in S (or a means of defining distance in S) is characteristic of measure-theoretic approaches. Significantly, it will later be convenient to use just such a topology in studying what is meant by measurement in the quantum-mechanical sense.

The quantum-theoretic reasoning. It is important to consider some of the aspects of an individual precision measurement, either classical or quantum-mechanical. First, it must be performed at an instant of time or over a period of time. Because an instant of time cannot be known to an arbitrary degree of accuracy, it must be assumed, in effect, that the measurement relates to a period of time. Next, a single measurement has the interpretation that, if it were performed again under the same circumstances, it would yield the same result or a second sample value from a fixed statistical distribution. If this were not the case and if experiments were not at least conceptually repeatable, an individual measurement and thus all measurement would be meaningless. Finally, the repetition of a precision measurement can at most be expected to give an identically repeated answer in the case in which a measured value is one of a finite or countable number of values. It is significant that it has not been possible to describe precision measurement without reference to time and that, with the exception of countably many possible measured outcomes, precision measurement has attached to it some error.

Effect of time on measurement

The system composed of the measured object and the measuring device must next be considered. The theory is to be guided with sufficient generality to ensure meaningfulness in quantum theory as well as classical science. Probabilistic analysis will be of basic importance in the study. Curiously enough, the purely mathematical works

on probability theory that relate directly to the lineage of Pascal, Laplace, Lebesgue, Wiener, Kolmogorov, and others have to a large extent become isolated from the physical sciences that reason in terms of quantum theory. In the words of the U.S. physicist Richard P. Feynman:

> The quantum mechanical laws of the physical world approach very closely the laws of Laplace as the size of the objects involved in the experiments increases. Therefore, the laws of probabilities which are conventionally applied are quite satisfactory in analyzing the behaviour of the roulette wheel but not the behaviour of a single electron or a photon of light.

On the other hand, a mathematical theory of chance phenomena that draws from the best mathematical work of the past and that is to be consistent with the physical postulates of 20th-century physics, including quantum theory, considers at the outset what is meant by the measurement process in all generality.

The space S composed of the points a in the unit interval $(0, 1)$ can now be re-examined. It is possible to retain the concept that points in S can be put into correspondence with the states of a physical system (though the phrase classical-physical system is avoided) and to retain the concept that the real value $f(a)$ of a function corresponds to a set in S and thus, intuitively, identifies incomplete information about the system.

Many-one functions and incomplete information

In the mathematical description of the classical process of measurement, a function f defined on S assumes a value when the system being measured occupies a specific state. From the many-one character of a function f, it can be inferred that a given value r of f is consistent with a set $A^{(r)}$ of points in S. This statement, though phrased in the classical setting, is meaningful whether the entity f, representing the property being measured, is characterized mathematically as a measurable function or as a function possessing certain local properties such as continuity.

The condition $f(a) = r$, for some fixed real value r, may be considered. This condition determines the set $A^{(r)}$, and the points composing $A^{(r)}$ will be called the roots of the equation $f(a) = r$. Physically, these roots are separate states of the system under observation that, individually, would contribute to the measured value r. There may be a finite, a countably infinite, or an uncountably infinite number of roots composing a set $A^{(r)}$. Whatever the case, they are not distinguishable in terms of f—that is, no condition imposed on f alone characterizes a subset of an $A^{(r)}$. To speak again in physical language, the condition $f(a) = r$ has reference to information about the state of a physical system, this information being nevertheless incomplete.

If in the spirit of 20th-century physics measurement on a space-time system is considered, a condition of the type $T(a) = t$, in which T is a real-valued function on S, can represent an observation on time. Consistent with the above remarks, a set $A^{(t)}$ that is the inverse image of the point t, as mapped through the function T, thus takes on the interpretation of incomplete information about the time of an observation. In this more general setting, in which time is represented mathematically as a function rather than a parameter, measurement on a physical system that, realistically, is inexact both in space and time is represented in terms of restrictions on two functions. If the function f represents one measurable (spatial) property of the physical system and T represents time, averages become replaced by a sum of all terms of the type $w_k f(a_k)$, for which a_k belongs to $A^{(t)}$ (see 20), in which, for the present discussion, it is assumed that the a_k are all the roots of the equation $T(a) = t$, and in which the w_k are weights, to be determined from the theory or observation, that depend on the functional structure of the time function T. Such averages are interpreted as the average (or expected) value of f, given that $T = t$.

Use of time as a function

It will occur that a knowledge of the coefficients w_k can be made tantamount to a knowledge of the square of the absolute value of the wave function—that is, the solution of the central equation of quantum mechanics. Thus, it may be expected that some analytic structure in the time function T similar to the analytic structure of the quantum-mechanical wave function is essential in deducing knowledge about these coefficients. The time function T is also needed to obtain the roots a_k, just as knowledge of the set of equations known as the Hamiltonian dynamics is essential to obtain the successive state values of a classical mechanical system. At least one way to obtain such structure in T is to make use of a topology in S for exploring the continuity and differentiability of T, considered as a function. From the discussion of the above section, however, it is seen that measure theory makes no essential use of a topology on S or of differentiability or continuity of functions defined on S. For these reasons it has already been emphasized that alternatives to the measure-theoretic approach are of value particularly in grasping the aspects of physical observation and measurement that supersede the classical approach of the last century and the early part of the present one.

Probability on infinite-dimensional spaces

It was clear from the work of Wiener in 1923 that, if any precise mathematical description of a random phenomenon as important to science as the Brownian motion (see above *Foundations of 20th-century probability theory*) was to be achieved, then Lebesgue measure must be extended to infinite dimensions. The reason for this was clear: The Brownian motion paths were observed in the laboratory to be highly irregular, and mathematically they were found to be expressed by functions that were so irregular as to be non-differentiable (that is, without tangents) except for a set that is negligible, technically of Lebesgue measure zero. The mathematical analysis of the Brownian-motion process could be achieved by dividing the time axis into a finite number of parts, by imposing a probability measure to the class of functions that conformed to certain spatial restrictions during those specified finite segments of time, and by passing to the limit in the number of temporal subdivisions. (Other techniques have since been found possible but are equivalent to this one, which was Wiener's first approach.) This analysis meant that probability measure, applied to as many dimensions as there were time intervals, namely, a finite number, must be extendible, in the limit, to infinite dimensions. Wiener found that the Daniell integral, already defined in order to treat functions in infinite dimensions, was a suitable tool with which to compute averages of functionals over the function space of Brownian-motion paths.

Brownian motion

The mathematical description of the Brownian motion constituted the first construction of an infinite collection of random variables ordered with increasing time and jointly satisfying certain measurability properties, now called a stochastic process. The techniques of Lebesgue measure theory, including the theorem that extends measure to infinite dimensions, are now the familiar beginning to the study of stochastic processes in general.

ADVANCED PROBABILITY AND STOCHASTIC PROCESS THEORY

Succeeding remarks will require some notation from the algebra of sets (see 21).

The definition of the interval-measurable class $\mathfrak{F}$ can be formulated by saying that $\mathfrak{F}$ is the smallest collection of sets in the unit interval that is "closed under countable unions" (made no larger by forming such unions)

(20) $$\sum_{a_k \in A^{(t)}} w_k f(a_k)$$

(21)
- $A \cup B$ is the set of points in A or B or both.
- $A \cap B$ is the set of points in both A and B.
- A^c is the set of points not in A.
- $A \supset B$ if and only if each point in B is also in A.
- $B \subset A$ if and only if $A \supset B$.

and that includes the empty set, $\emptyset$, the whole space, S, and all subintervals within S. The same definition for $\mathfrak{F}$ can also be stated symbolically (see 22).

The importance of this expression for $\mathfrak{F}$ is that it brings into evidence the close relationship between $\mathfrak{F}$ and two important classes of sets in S that are widely used throughout probability theory. Specifically, in the language of the above definition, there can be the introduction of classes $\mathfrak{A}$ and $\mathfrak{B}$, such that $\mathfrak{A} \subset \mathfrak{F} \subset \mathfrak{B}$ (see 21).

Let $\mathfrak{B}$ be defined as the smallest collection of sets in S including $\mathfrak{F}$ that is closed under countable-set operations. An equivalent symbolic definition is also possible (see 23). The existence of a smallest field with these properties follows from the fact that the class of all subsets in S includes $\mathfrak{F}$ and is closed under countable-set operations and the fact that the intersection of any two classes possessing such properties yields a class with the same properties.

In a definition similar to that for $\mathfrak{F}$, it is possible to define $\mathfrak{A}$ as the smallest collection of sets in the unit interval that is closed under finite unions and that includes the empty set, $\emptyset$, the whole space, S, and all subintervals within S. As a consequence, $\mathfrak{A}$ is closed under all finite set operations and is contained in $\mathfrak{F}$. Defining properties for $\mathfrak{A}$ can be expressed purely symbolically (see 24).

The class $\mathfrak{A}$ or any collection of sets with the same closure properties is called a field, while class $\mathfrak{B}$ that has more extensive closure properties is called a σ-field. If $\mathfrak{J}$ is the class of subintervals in S, then F is sometimes denoted $\mathfrak{J}_\sigma$, consistent with the notation of classification of sets due to the French mathematician Émile Borel, to indicate that it is composed of all countable unions of sets from $\mathfrak{J}$. As such, members of $\mathfrak{F}$ are identified to be of rather simple structure compared to sets in a hierarchy of more and more complicated sets leading up to those comprising $\mathfrak{B}$.

σ-field of sets

In the present context $\mathfrak{B}$ could be called the σ-field generated by $\mathfrak{F}$ or $\mathfrak{A}$. It should be emphasized, as it was by Kolmogorov, that, while a field can be considered observable, the generated σ-field is purely a mathematical convenience. In Kolmogorov's words:

> Even if the sets (events) A of $\mathfrak{A}$ can be interpreted as actual and (perhaps only approximately) observable events, it does not, of course, follow from this that the sets of the generated field, $\mathfrak{B}$, reasonably admit of such an interpretation. . . . Thus, sets of $\mathfrak{B}$ are generally merely ideal events to which nothing corresponds in the outside world. However, if reasoning which utilizes the probabilities of such ideal events leads us to a determination of the probabilities of an actual event of $\mathfrak{A}$, then, from an empirical point of view also, this determination will automatically fail to be contradictory.

The intermediate class $\mathfrak{F}$ has been chosen because it is still close to the intuition and to a class of actual events in Kolmogorov's sense and, at the same time, it comes closer to providing the mathematical flexibility of the more complex class $\mathfrak{B}$. Deeper reasons for considering $\mathfrak{F}$ have to do with the basic fact that the measure-theoretic approach to the concept of probability is not sufficiently general for some purposes in quantum theory (see above *The quantum-theoretic reasoning*) and automata theory (see above *The additivity assumption*).

The broadening from $\mathfrak{F}$ to $\mathfrak{B}$ calls for an extension of the domain of definition of the Lebesgue measure. Earlier (see above *The measure-theoretic reasoning*), so long as sets in $\mathfrak{F}$ were the only ones involved, the intuitive notion of Lebesgue measure as the length of an interval or the sum of such lengths sufficed. It now becomes necessary to invoke the extension theorem (reserving the complete statement until later; see below *Extension of probability measure to infinite-dimensional spaces*) that assures that a measure can be defined uniquely for all sets in $\mathfrak{B}$ in such a way that it coincides with the intuitive notion of Lebesgue measure for sets of $\mathfrak{F}$. This theorem depends strongly on the property of Lebesgue measure on $\mathfrak{F}$ called complete additivity and defined by identifying the measure of the union of sets with the sum of the measures (see 25), for every disjoint sequence $F_1, F_2, \cdots$ of sets in $\mathfrak{F}$ the union

(22) $\mathfrak{F}$ is the collection of sets in S such that:
(i) $\emptyset \in \mathfrak{F}, S \in \mathfrak{F}$.
(ii) $I \in \mathfrak{F}$, for I any interval in S.
(iii) If $A_k \in \mathfrak{F}, k = 1, 2, 3, \cdots$, and $A = A_1 \cup A_2 \cup A_3 \cup \cdots$, then $A \in \mathfrak{F}$.
(iv) If $\mathfrak{F}'$ satisfies i), ii), iii), then $\mathfrak{F}' \supset \mathfrak{F}$.

(23) The collection $\mathfrak{B}$ has the three properties:
(i) $\mathfrak{B} \supset \mathfrak{F}$.
(ii) If $B_k \in \mathfrak{B}, k = 1, 2, 3, \cdots$, and if $B = B_1 \cup B_2 \cup \cdots$, then $B \in \mathfrak{B}$ and $B_k^c \in \mathfrak{B}$.
(iii) If $\mathfrak{B}'$ has properties i) and ii), then $\mathfrak{B}' \supset \mathfrak{B}$.

(24) Defining properties for $\mathfrak{A}$:
(i) $\emptyset \in \mathfrak{A}, S \in \mathfrak{A}$.
(ii) $I \in \mathfrak{A}$ for I any interval in S.
(iii) If $A_k \in \mathfrak{A}, k = 1, 2, \cdots, n$, and $A = A_1 \cup A_2 \cup \cdots \cup A_n$, then $A \in \mathfrak{A}$.
(iv) If $\mathfrak{A}'$ satisfies i), ii), iii), then $\mathfrak{A}' \supset \mathfrak{A}$.

(25) $$\mu(F_1 \cup F_2 \cup \cdots) = \sum_{k=1}^{\infty} \mu(F_k)$$

of which belongs to $\mathfrak{F}$. The complete additivity assumption had been anticipated by Bernoulli (see above *The additivity assumption*). With the use of the extension theorem, there is a mathematical entity, the Lebesgue measure, $\mu(B)$, to go with the members, B, of the mathematical abstraction $\mathfrak{B}$. The Lebesgue measure is simply a non-negative set function satisfying the property of complete additivity on $\mathfrak{B}$ and coinciding with the intuitive notion of measure on $\mathfrak{F}$.

Lebesque measurable sets

The distinction between the σ-field, $\mathfrak{B}$, and the class of Lebesgue measurable sets can be found only by examination of the subsets of measure zero in $\mathfrak{B}$. The broadening from $\mathfrak{B}$ to the class of Lebesgue measurable sets is called completing the field, $\mathfrak{B}$, and the extension of the measure to the completed field is called completing the measure.

Probability spaces, random variables, and distribution functions. The two entities $(S, \mathfrak{B})$ when taken together are a special case of a Borel measurable space, and the three entities $(S, \mathfrak{B}, \mu)$ when taken together are a special case of a Borel measure space. Because $\mu(S) = 1$, the term probability space also applies, and $(S, \mathfrak{B}, \mathbf{P})$ is written for the previous triplet. The restriction of S in these definitions to the unit interval rather than to an abstract space is seldom a limitation, and in many cases it offers an enormous simplification. Not only can almost all of the mathematical ingredients of theorems and proofs of theorems be found in this setting, but they appear without the complications of generality that sometimes obscure the meaning. The results achieve no greater generality when stated with the aid of measure spaces $(\Omega, \mathfrak{B}, \mu)$, for which Ω can be carried onto S by a one-to-one mapping that preserves measure (sets that correspond through the mapping having equal measure). Such measure spaces are called regular.

Just as a class of sets wider than $\mathfrak{F}$ is mathematically essential, the class of Lebesgue measurable functions is an essential generalization of the interval-measurable class. (In the following definition, as anticipated above, the σ-field, $\mathfrak{B}$, is assumed to be completed and, thus, is composed of the class of Lebesgue measurable sets in the unit interval.)

A function f defined on S to R is Lebesgue measurable if and only if for any set A in $\mathfrak{F}$ the inverse image of A, as mapped through f, is in $\mathfrak{B}$. (The term measurable func-

tion, if not otherwise qualified, will mean Lebesgue measurable function.)

A measurable function defined on S is called a random variable.

In a purely mathematical sense, the class $\mathfrak{F}$, however simple, is representative of $\mathfrak{B}$ in a way that formal definitions and theorems make precise (see 26, 27). Similarly, though interval-measurable functions appear to be rather special in comparison with measurable functions, they are representative of the latter in the sense of a formally expressed theorem (see 28).

After the sets that are to provide the mathematical generalization of events are defined, it is necessary to introduce a convenient means of calculating probabilities for representative sets. Such means are supplied by distribution functions F, which are defined over the entire real line and possessing the following properties: A distribution function F vanishes at negative infinity, it is a monotone nondecreasing function as its argument increases, it is continuous from the left (though continuity from the right is also acceptable), and it is bounded by unity (that is, it converges to 1 as its argument approaches positive infinity). A distribution function is generated by each random variable X, though more than one random variable can possess the same distribution function. For a given random variable X, one of the familiar algorithms that is used to produce the relevant distribution function F_X identifies the value of the distribution function at point t with the probability measure of the set of points in the space S, for which X assumes a value less than t (see 29). Thus, with the device of distribution functions, measure relevant to X is transferred from the space S to the real line. A number of calculational details concerning special distribution functions, particularly in the n-dimensional case, can be found in the article STATISTICS.

Distribution functions

It is not incidental that the properties of a distribution function are the same (except for the unity bond) as the properties of a Lebesgue–Stieltjes measure that associates a number, not necessarily the length, with each interval of the real line (see ANALYSIS, REAL: *Measure and integral calculus*). The importance of this fact from the probabilistic point of view is found in a basic theorem concerning the mathematical expectation.

The mathematical expectation $\mathbf{E}X$ of any random variable X is defined as the Lebesgue integral of the random variable over the space S, integrated with respect to the probability measure $\mathbf{P}$ (see 30).

The basic theorem just referred to states that mathematical expectation of a random variable X, or a function h of it, can be calculated as the Lebesgue–Stieltjes integral of the identity function, or the function h, integrated over the whole real line with respect to the measure induced by the distribution function of X (see 31). This integral is the more familiar form of expectation.

Random variables that assume only a finite number of values—examples can be constructed from elementary problems in games of chance (see above *Elementary problems*)—have distribution functions composed of "steps" and "jumps." The random variable with binomial distribution function F_B (see 32) is especially important and is used in the calculation of probabilities for the number k of successes in a sequence of N experiments (such as flips of a coin) that are statistically independent events with constant probability p of success in a single trial. For such variables the Lebesgue–Stieltjes integral in the basic theorem cited reduces to a sum (see 33) and the mathematical expectation becomes identified with the form first discovered by Huygens (see above *Mathematical expectation*).

Binomial and normal distributions

Random variables that assume a continuum of values generally have distribution functions that are absolutely continuous. An important example is the random variable with normal distribution function F_N (see 34), first discovered by De Moivre, studied by Laplace and by the German mathematician Carl Friedrich Gauss (*q.v.*), incorporated in the basic work of the French mathematician Paul-Pierre Lévy, and found by Wiener to be an integral part of the stochastic process that is used to describe the Brownian motion. For such random variables the mathematical expectation reduces to a Lebesgue–Riemann integral (see ANALYSIS, REAL) of the product of the identity function and the derivative f_X of the distribution function called the probability density function (see 35).

The concept of statistical independence generalizes in a natural way from independent events (see above *Conditional probability*) to independent random variables by defining the variables as mutually independent if representative inverse images, as mapped through the variables, are mutually independent events. Basic theorems concerning sequences of mutually independent random variables can then be stated and proved.

Law of large numbers and central limit theorem

One of the most important of these theorems is a generalization of the law of large numbers that was discovered by Bernoulli called "the strong law of large numbers." One form of the generalization (see 36) states that, for a sequence of mutually independent random variables $\{X_k\}$ with common distribution function and finite mathematical expectation $\mathbf{E}X$, the difference between the arithmetic average of the number N of the variables $(X_1 + X_2 + \cdots + X_N)/N$ and the expectation common to the variables converges to (or intuitively comes closer to) zero as N increases (except for certain points in the space S, the totality of such exceptional points constituting a set with zero probability measure).

(26) Let the measure space $(S, \mathfrak{B}, \mu)$ be given. A class $\mathfrak{F}$ of subsets of S is dense in $\mathfrak{B}$ if $\mathfrak{B} \supset \mathfrak{F}$ and if for each $B \in \mathfrak{B}$ and for any positive number ϵ there exists a set $F \in \mathfrak{F}$ such that $\mu(F \Delta B) < \epsilon$, where $F \Delta B = (F \cap B^c) \cup (F^c \cap B)$ is the symmetric difference.

(27) The interval measurable class, $\mathfrak{F}$, is dense in the Lebesgue measurable class $\mathfrak{B}$.

(28) Let f be a measurable function defined on the measure space $(S, \mathfrak{B}, \mu)$. For any positive number ϵ, there is a set $B \in \mathfrak{B}$, $\mu(B) < \epsilon$, and an interval measurable function g (dependent upon ϵ) defined on $(S, \mathfrak{F}, \mu)$ such that $|f(a) - g(a)| < \epsilon$, except for $a \in B$.

(29) $$F_X(t) = \mathbf{P}[X < t], \qquad -\infty < t < \infty$$

(30) $$\mathbf{E}X = \int_S X \, d\mathbf{P}$$

(31) $$\mathbf{E}X = \int_{-\infty}^{\infty} t \, dF_X(t), \qquad \mathbf{E}h(X) = \int_{-\infty}^{\infty} h(t) \, dF_X(t)$$

(32) $$F_B(t) = \sum_{k<t} \binom{N}{k} p^k (1-p)^{N-k}$$
$$0 \leq t \leq N, \qquad 0 < p < 1$$

(33) $$\int_{-\infty}^{\infty} t \, dF_X(t) = \sum_{k=1}^{N} x_k \, \mathbf{P}[X = x_k],$$
if X assumes only the values $x_1, x_2, \cdots, x_N$.

(34) $$F_N(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} \, dx, \qquad -\infty < t < \infty$$

(35) $$\int_{-\infty}^{\infty} t \, dF_X(t) = \int_{-\infty}^{\infty} t f_X(t) \, dt$$
if F_X is absolutely continuous

(36) Strong law of large numbers: If the random variables $\{X_k\}$ are mutually independent with common distribution function and finite mathematical expectation $\mathbf{E}X$, and if $\overline{X} = (X_1 + X_2 + \cdots + X_N)/N$ with values $\overline{X}(a)$, then except for points a in a set O for which $\mathbf{P}(O) = 0$:
$$[\overline{X} - \mathbf{E}X] \to 0.$$

(37) Central limit theorem: If the random variables $\{X_k\}$ are mutually independent with common distribution function, if $\mathbf{E}X < \infty$, $\mathbf{E}[X - \mathbf{E}X]^2 < \infty$, and if $\overline{X} = (X_1 + X_2 + \cdots + X_N)/N$, then for $-\infty < t < \infty$ it follows that as N increases:

$$\mathbf{P}[\sqrt{N}\,(\overline{X} - \mathbf{E}X) < t\,\mathbf{E}^{1/2}(X - \mathbf{E}X)^2] \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\,dx.$$

Another important theorem attributed to De Moivre and Laplace, called the central limit theorem, is also expressed in terms of a sequence $\{X_k\}$ of mutually independent random variables. One of the generalized forms of this theorem (not the most general) states that, if the variables have common distribution function, finite expectation $\mathbf{E}X$, and finite variance $\mathbf{E}[X - EX]^2$, then a modified form of the difference between the arithmetic average of the number N of the variables and the expectation common to the variables (a form of the difference that is in effect multiplied by the square root of N) behaves like a random variable with normal distribution for large enough value of N (see 37). The applications of this theorem, particularly in the field of statistics (*q.v.*), are multifold.

Extension of probability measure to infinite-dimensional spaces. If two set functions s_1 and s_2 are defined on $\mathfrak{F}$ and $\mathfrak{B}$, respectively, it may be said that s_1 is the restriction of s_2 on $\mathfrak{F}$, provided s_1 and s_2 agree on $\mathfrak{F}$. Similarly, s_2 is the extension of s_1 to $\mathfrak{B}$, provided s_1 and s_2 agree on $\mathfrak{F}$. The concepts of restriction and extension are used similarly in reference to set functions defined on any two classes of sets, one of which is contained in the other.

A class of sets closed under formation of monotone limits is sometimes naturally associated with a field or a σ-field. A class of sets in S is closed under formation of limits of monotone sequences if the logical union of the sequence of sets $\{A_n\}$ belongs to the class, whenever $\{A_n\}$ belongs and $\{A_n\}$ is nondecreasing; and the logical intersection of the sequence of sets $\{A_n\}$ belongs to the class, whenever $\{A_n\}$ belongs and $\{A_n\}$ is nonincreasing.

A monotone field has the properties of a monotone class given above as well as the field properties. A monotone class generated by a field is the smallest monotone class containing the field.

It can be shown that a monotone field is a σ-field (the converse being obvious). It can also be shown that the monotone class generated by a field and the σ-field generated by the same field are identical. These facts are important in the construction of proofs of relevant theorems.

In the terminology of this section, the smallest σ-field containing the class of open sets in the n-dimensional Euclidean space R^n is the Borel field in R^n. The Borel field can also be obtained by starting with the class of closed sets in R^n or the class of closed or open intervals (n-dimensional intervals in the case of n dimensions). The elements of the Borel field are Borel sets.

An important example in the relation between fields and σ-fields is built up from the concept of a product set. Given two spaces S_1 and S_2, and two sets A_1 and A_2 chosen from fields of sets in S_1 and S_2, respectively, the product set $A_1 \times A_2$ is composed of the set of all ordered pairs (a_1, a_2), with a_1 in A_1 and a_2 in A_2. The class of all finite unions of product sets composed in this way is a field of sets in $S_1 \times S_2$. It is called a product field. If the field from which A_1 is chosen is $\mathfrak{F}_1$ and the field from which A_2 is chosen is $\mathfrak{F}_2$, then the product field is written $\mathfrak{F}_1 \times \mathfrak{F}_2$.

If $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are σ-fields, the product σ-field $\mathfrak{B}_1 \times \mathfrak{B}_2$ is obtained by first composing the product field and then identifying $\mathfrak{B}_1 \times \mathfrak{B}_2$ with the smallest σ-field containing the product field.

Generalizations of these concepts to countable and even uncountable dimensional spaces will be of importance in studying measure on function space. For example, if the spaces S_t and sets within them A_t are indexed by a parameter t, in which t belongs to a countable or uncountable subset T in the real space, then the above definitions are extended in the obvious way to product sets ΠA_t, product fields $\Pi\mathfrak{F}_t$, and product σ-fields $\Pi\mathfrak{B}_t$, in which the product is extended to all choices of the parameter t in the index set T. In more detail, if the index set T is composed of the positive integers, a product set ΠA_t will be defined as the collection of all sequences $(a_1, a_2, \cdots)$, with a_k in A_k, the A_k belonging to the fields $\mathfrak{F}_k$, and $A_k = S_k$ for all but a finite number of values of k. The class of all finite unions of product sets formed in this way is the product field $\Pi\mathfrak{F}_t$. The smallest σ-field containing this field is the product σ-field $\Pi\mathfrak{B}_t$. As another example, if the index set T is composed of all finite real numbers, the product field $\Pi\mathfrak{F}_t$ is formed of all finite unions of sets ΠA_t, in which a typical set ΠA_t is defined as the collection of all sequences of elements, each from some A_t, with subscripts of such a sequence constituting a countable set of points $(t_1, t_2, \cdots)$. (Restrictions on the choice of the A_t hold as before.) From such a field the product σ-field is defined as before.

The construction of measure on product measurable spaces as well as on σ-fields in general can follow in a unique sense from the properties of a (σ-finite) measure on the generating field. This fact, at the root of much of the work in measure theory, is expressed in the extension theorem.

Techniques of the extension theorem

The techniques that make up the statement and proof of the extension theorem come from the properties of outer measure as defined on real point sets. The discussion of extension of measure in the article ANALYSIS, REAL should also be consulted. If one identifies (following Lebesgue) with any set of real points the infimum (greatest lower bound) of sums of lengths of sequences of open intervals the union of which contains the given set, one obtains the real-value set function m^* called outer measure. As immediate properties of outer measure defined in this real setting, one obtains: first, $m^*(0) = 0$; second, $m^*(A) \leq m^*(B)$ for sets A and B, such that $B \supset A$; and third, $m^*(\cup A_k) \leq \Sigma\, m^*(A_k)$ for any sequence $\{A_k\}$. Moreover, having proved these properties, it is natural (after the ingenious stroke of the German mathematician Constantin Carathéodory) to consider the behaviour of m^* on the class of sets A that satisfy, for every set D, the relation: fourth, $m^*(D) = m^*(A \cap D) + m^*(A^c \cap D)$ for every set D. If a motivation is sought for the Carathéodory property, it might ensue from the inequality (not strict) of property preceding, together with the equality that might be expected to result if the sets A and D are not too much more complicated than intervals.

It is proved—still in the context of the real space—that the class of sets A satisfying the Carathéodory property is a σ-field and m^* on this σ-field is countably additive.

Clearly, the class of sets satisfying the fourth property includes all intervals, and thus, by restriction of m^* to the smallest σ-field including all intervals, $\mathfrak{B}$, an extension of measure is obtained from the class of intervals to the Borel field $\mathfrak{B}$.

These are the techniques that suggest the following theorem, which is known as the extension theorem:

If a measure m is defined on a field $\mathfrak{F}$, there exists a measure m' on the σ-field $\mathfrak{B}$ generated by $\mathfrak{F}$ that is an extension of m to $\mathfrak{B}$; and if m is σ-finite, then m' is σ-finite and unique.

The mathematical analysis represented by the extension theorem, or a technique that is equivalent, provides sufficient generality to define a random function or, equivalently, a stochastic process. Knowledge of stochastic processes is largely composed of knowledge of certain classes of processes. It is convenient to define certain properties and then confine attention to the processes, the second-order processes, the Markov processes that satisfy these properties. Other classes are processes with independent increments, processes with orthogonal increments, martingales, and stationary stochastic processes. A discussion of an important part of the theory of stationary stochastic

Major classes of stochastic processes

processes is included under the headings *Control and single-series prediction* and *Multiple prediction theory* in the article AUTOMATA THEORY. In view of the breadth of the subject matter that has now developed in this field, it is necessary to limit the scope of the remarks to be presented here. Of basic significance are the Markov processes. Moreover, in the literature there are theorems that suggest an underlying relationship between Markovian and non-Markovian processes. It would seem valuable to review the precise meaning of these theorems not only in order to convey an introductory idea of the Markovian class but to characterize the nature of other processes as well.

MARKOVIAN AND NON-MARKOVIAN PROCESSES

Just as the historically important problems that drew Pascal and others to the theory of probability served to illustrate the basic concepts of probability theory in an elementary sense, an important modern problem in cosmic-ray physics will serve to illustrate the relationship of Markovian to non-Markovian processes.

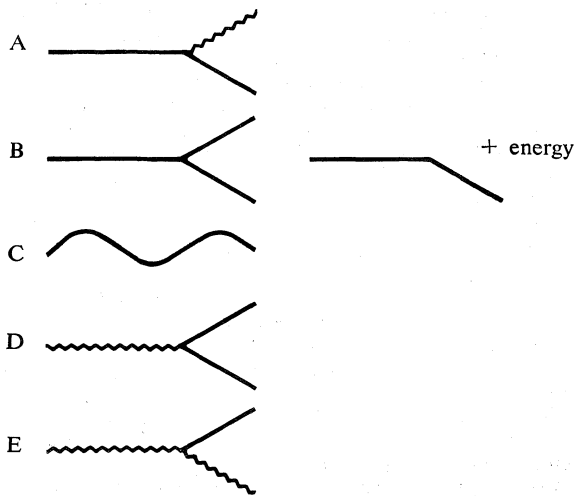
The electron–photon shower problem. In the early 1930s a number of experimental observations on cosmic rays (the radiation detected from outer space) suggested some kind of multiplication, or population-growth process specific to the cosmic rays as they passed through the Earth's atmosphere. Principal experimental observations were those of the Italian-U.S. and German-U.S. physicists Bruno Rossi and Victor Regener. Such experiments led to a long-standing problem in theoretical physics that was to explain the growth and death of a population of fast electrons and photons, two of the components of the cosmic rays, as they passed through matter. Work on the theoretical problem began in about 1934 with the ideas of men in Germany, Scotland, India, and other countries, primarily the physicists Lothar Nordheim, Hugh Carmichael, Homi Bhabha, and Walter Heitler. It became apparent that the growth of cosmic-ray populations was caused by successive branching processes and that these processes, the radiation loss by electrons and materialization by photons, could be predicted from what was then the new quantum theory.

Increasing evidence for a shower phenomenon in cosmic rays accumulated with more experimentation so that by 1937 the cascade-shower problem, as it came to be called, could be enunciated as a theoretical, though physically meaningful, objective: under the assumption of well-specified initial conditions, to predict the number of electrons and photons, descendants of an initial high-energy particle, that emerge with specified energies from a given thickness of absorbing material. The clear statement and partial solution of this problem were accomplished in 1937 by two teams of research physicists, Bhabha and Heitler as well as J. Franklin Carlson and J. Robert Oppenheimer (*q.v.*).

The reasons for the mathematical difficulties in the cascade-shower problem will be clear after the process for electron-photon development has been described mathematically and a general equation for it has been written down. First, it is necessary to remark that the basic processes that contribute to the development of a shower are: (A) radiation loss by an electron (an electron branches into an electron and a photon); (B) collision loss by an electron (an electron branches into two electrons, as in collision with a free electron, or it simply loses a fraction of its energy to excitation of an atom, as in collision with a bound electron); (C) deflection of an electron in the Coulomb field of a nucleus without the electron losing any energy (the electron path becomes curved at random from multiple scattering); (D) pair production by a photon (a photon branches into two electrons); and (E) Compton collision by a photon (a photon branches into an electron and a photon).

Basic processes of electron–photon interaction

Although only two of these processes, (A) and (D), are important at high energies in which the shower begins, all of the processes will contribute before the energy of the shower has been completely dissipated through excitation. Pictorially, these processes can be illustrated in the following way:



Ignoring for the time being process (C) and ignoring the fact that there is actually an angle of emission at each branch point, it is possible to specify the state, σ, of a shower at the point $t \geq 0$ of penetration by giving some particular values to a set of parameters. A suitable set of parameters might be n, m, $\bar{e}$, and μ, representing, respectively, the number of electrons, the number of photons, an $(n+m)$-dimensional energy vector, $\bar{e} = (e_1, e_2, \cdots, e_{n+m})$, and the ordering of electrons and photons among the energy components of $\bar{e}$. The ordering parameter μ can take on any of 2^{n+m} permissible values, each value representing one of the 2^{n+m} ways in which $n+m$ particles can be distributed among $n+m$ energies. A possible state of a shower is written: $\sigma = (n, m, \bar{e}, \mu)$. If I is the initial energy of the shower at zero penetration, then no individual energy, e_k, can exceed it.

In what follows, the random function $W(t)$ can be thought of as a random variable for each fixed value of t. (For a rigorous definition of random function, see below *Stochastic processes: general definitions.*) With these conventions, it can be said that a shower is completely described by the random function $W(t)$, in which $W(t)$ may take its values in the space $\Sigma = \{\sigma\}$ of all possible states σ. The mathematical problem is to determine the conditional probability, $P(\Delta\sigma, t|\sigma_0, 0)$, that $W(t)$ will belong to some set of states $\Delta\sigma$, given that $W(0)$ coincided with some initial state σ_0. This will be written as $P(\Delta\sigma, t)$, without the initial condition appearing explicitly when the initial condition is constant and understood. The symbol $\Delta\sigma$ will be understood as the set of states, σ, with common n, m, and μ but with energy components taking values anywhere in the intervals $(e_k, e_k + e'_k)$, $e'_k \geq 0$, $k = 1, 2, \cdots, n+m$.

Because it can be assumed, first, that the processes (A, B, C, D, E) all occur at some discrete points in space, namely, the points of impact with the Coulomb field of the

(38)

$$P(\Delta\sigma, t|\sigma', s) = [1-\alpha(\sigma')(t-s)]\delta(\sigma', \Delta\sigma) + \pi(\sigma' \to \Delta\sigma)(t-s) + o(t-s),$$

$$\delta(\sigma', \Delta\sigma) = \begin{cases} 1 & \sigma \in \Delta\sigma \\ 0 & \text{Otherwise.} \end{cases}$$

In general $o(t-s)$ will depend on σ' and $\Delta\sigma$. The explicit forms of $\alpha(\sigma')$ and $\pi(\sigma' \to \Delta\sigma)$ can be given when they are needed for calculation. If $\Delta\sigma$ is fixed, $\pi(\sigma' \to \Delta\sigma)$ is measurable with respect to some σ-field, $\mathfrak{B}$, of sets in Σ. For fixed σ', $\pi(\sigma' \to \Delta\sigma)$ is a completely additive, non-negative function of sets in $\mathfrak{B}$ with

$$\pi(\sigma' \to \Sigma) = \alpha(\sigma'),$$
$$\pi(\sigma' \to \sigma') = 0.$$

$\pi(\sigma' \to \Delta\sigma)$ is independent of s and t.

nucleus or the electron, and, second, that before each process the particle "remembers" only what its energy is and whether it is an electron or a photon, the following assumptions I and II can be made.

Purely discontinuous Markovian processes, the diffusion equation. I. The process is purely discontinuous and temporarily homogeneous. That is to say, the conditional probability of $W(t)$ being contained in $\Delta\sigma$, given that at the fixed point s, $0 \leq s < t$, the variable $W(s)$ coincided with the state σ', is a continuous function of t and s, independently, and has a specific limiting form (see 38).

II. The process is Markovian. That is to say, the conditional probabilities of $W(t)$ being contained in $\Delta\sigma$, given $\{W(t): t \leq s\}$ and given $W(s)$, are the same. In other words, the future state of the system is independent of the history prior to s, so long as $W(s)$ is known (see 39).

Once properties I and II are assured, the most obvious approach to the calculation of $P(\Delta\sigma, t)$ is to write down and solve the diffusion equation that it satisfies (see 40). Following the treatment of the U.S. mathematician William Feller for a purely discontinuous Markov process, it is possible to write $P(\Delta\sigma, t)$ as an infinite sum (see 41) of probabilities, $P^{(N)}(\Delta\sigma, t)$, each summand thus being given in terms of an integro-differential-difference equation that can be solved in successive steps in N. The $P^{(N)}(\Delta\sigma, t)$ can be interpreted as the compound probability that, while penetrating x thickness of material, the system will experience exactly N transitions and that $W(t) \in \Delta\sigma$, if it is known that $W(0) = \sigma_0$.

The integro-differential-difference equation that appears in this context evolved first in the researches of Sydney Chapman, an English mathematician, and Kolmogorov and is called the Chapman–Kolmogorov equation. It might appear here that one had only to solve this equation and the electron-photon problem would be completed. In reality, this method of solution presents too difficult a procedure because the variety of energies that the electrons and photons might assume at each point in the development of the cascade shower is too large. Though, as a Markov process, the structure of the cascade shower is describable in mathematical terms, the detailed nature of the state space Σ makes the solution of the equation unduly complicated.

Chapman–Kolmogorov equation

By omitting detailed reference to the energies, the complicated structure of each state disappears. The simplification of the state space is accomplished, however, at the possible expense of the Markov character of the process as a whole. For this reason, the cascade-shower problem was one of the earliest sources of ideas for handling non-Markovian problems and problems that could be expressed in terms of functions of Markovian processes that resulted from providing an incomplete description of an underlying Markovian process. Other sources of ideas of this kind became apparent in the probabilistic studies on growth of biological populations.

Stochastic processes: general definitions. It is possible to suppose that the set of possible states that a stochastic process assumes at a fixed time t forms a space and that this space is the same for any time t. As time progresses the process will experience transitions from point to point in the space. This space may now be partitioned into subsets, and it is possible to consider the event that at time t the state of the process belongs to a specified set of the partition. This and other like events corresponding to all different points in time, if collected, will form a sample value of a new stochastic process. The new process is, of course, determined by the first and by the partition and may be said to describe the transitions of the original process from set to set of the partition. The particular transformation involved in relating two processes in the above way is simply a mapping of the space of states of the original process onto a partition. Such a mapping is an appropriate tool with which to study the interrelation between processes and, more specifically, Markovian and non-Markovian processes.

Let $(\Omega, \mathfrak{B}_\Omega, \mathbf{P})$ be a probability space, (Γ, B_Γ) a measurable space, and R^+ the space of positive real numbers. Let $G(t)$ be a measurable function defined on $\Omega \times R^+$ to Γ. The symbols Ω and Γ represent some general spaces made up of points ω and γ, respectively, while $\mathfrak{B}_\Omega$ and B_Γ are some σ-fields of sets in Ω and Γ, respectively. $\mathbf{P}$ is a probability measure defined for all sets in $\mathfrak{B}_\Omega$. The function $G(t)$ defined above will be called a random function or a stochastic process. Γ will be called the range space of $G(t)$.

$$\mathbf{P}[W(t) \in \Delta\sigma \mid W(s)] = \mathbf{P}[W(t) \in \Delta\sigma \mid W(r), r \leqslant s], \qquad t \geqslant s \geqslant 0 \tag{39}$$

$$(\partial/\partial t)P(\Delta\sigma, t) = \int_\Sigma P(d\Sigma_{\sigma'}, t)\pi(\sigma' \to \Delta\sigma) - \int_{\Delta\sigma} P(d\Sigma_{\sigma'}, t)\alpha(\sigma'). \tag{40}$$

$$P(\Delta\sigma, t) = \sum_{N=0}^{\infty} P^{(N)}(\Delta\sigma, t) \tag{41}$$

It should be further noticed that, if Λ is chosen to be a measurable partition of Γ, and if the random function $G(t)$ is given, then it is always possible to treat (Λ, B_Λ) as a new measurable space and to define a new random function $L(t)$ as follows: $L(t)$ is the measurable function on $\Omega \times R^+$ to Λ, such that $L(t) = \lambda$ whenever $G(t) \in \lambda$. In other words, if $\phi(\gamma)$ denotes the function on Γ to Λ defined by: $\phi(\gamma) = \lambda$ for all $\gamma \in \lambda$, then $L(t) = \phi[G(t)]$.

Functions of Markovian processes. Two important facts about the related random functions $G(t)$ and $L(t)$ can now be expressed. First, the $G(t)$ process may be such that the probability of its future is equally well conditioned on the present or the present combined with the past—that is, it may be Markovian. If so, the Markovian assumption relating conditional probabilities, as expressed for $W(t)$ above, applies (see 39). If such is the case, $L(t)$ may or may not be Markovian. The information in the process $L(t)$ is not, in general, sufficiently complete to permit a reconstruction of the process $G(t)$, and, as such, that component of it that relates to the present state may not offer a sufficiently rich condition on which to evaluate the probability of the future without the addition of information from the past. Second, though in general a stochastic process is not Markovian, any stochastic process, say, $L(t)$, can be analyzed in such a way that it is related to a Markovian process $G(t)$ through some function ϕ. These facts provide guidelines for extending analytical methods that were originally established by Kolmogorov for Markovian processes to functions of Markovian processes and, thus, potentially to many types of processes.

Loss of information and loss of the Markovian property

The class of sets $\Lambda = \{\lambda\}$ in the space Γ is said to be enchained by the random function $G(t)$ if the following conditions hold:

A. Λ is measurable in the sense that $\lambda \in B_\Gamma$ for each $\lambda \in \Lambda$.

B. Λ forms a partition of Γ. That is to say, the sets of Λ are disjoint and their union covers the whole space Γ.

C. If $L(t) = \phi[G(t)]$ is the random function defined earlier, then for all t, s, $0 \leq s \leq t$, and for all sets $\Delta\lambda \in B_\Lambda$: $\mathbf{P}\{L(t) \in \Delta\lambda | L(s)\} = \mathbf{P}\{L(t) \in \Delta\lambda | G(s)\}$, with probability 1.

The purpose of this definition is to place conditions on Λ and $G(t)$ that permit the random function $L(t)$ to be Markovian if $G(t)$ is Markovian. This conditioning is done essentially through C, which is equivalent to asking that, for each fixed t and s, the conditional probability $\mathbf{P}\{L(t) \in \Delta\lambda | G(s)\}$ is measurable not only with respect to the smallest σ-field generated by $G(s)$ but also with respect to that generated by $L(s)$. Intuitively speaking, condition C extends the Markov property from $G(t)$ to $L(t)$ by demanding that any transition of the process $G(t)$ from points in $\lambda \in B_\Gamma$, at t_1, to a set $\Delta\lambda \in B_\Lambda$, at t_2, $t_2 > t_1 \geq 0$, have the same probability. Though one ultimately wishes to write down an equation for the transitions of the process $G(t)$ through the sets of Λ, the new

Implications for calculational methods

(42)

$$\begin{cases} (\partial/\partial t)P_1(\Delta\lambda, t \mid \lambda_0, 0) \\ \qquad = \int_\Lambda P_1(d\Lambda_{\lambda'}, t \mid \lambda_0, 0)\,\pi'(\lambda' \to \Delta\lambda) \\ \qquad\quad - \int_{\Delta\lambda} P_1(d\Lambda_{\lambda'}, t \mid \lambda_0, 0)\,\alpha'(\lambda') \\ \lim_{t\to 0} P_1(\Delta\lambda, t \mid \lambda_0, 0) = \delta(\lambda_0, \Delta\lambda) \\ \text{The functions } \pi'(\lambda' \to \Delta\lambda) \text{ and } \alpha'(\lambda'), \text{ as functions of } \lambda', \\ \text{are unique up to sets of measure zero and are defined} \\ \text{by:} \\ \pi'(\lambda' \to \Delta\lambda) = \pi(\gamma \to \Delta\lambda) - \pi(\gamma \to \lambda')\delta(\lambda', \Delta\lambda), \\ \alpha'(\lambda') = \alpha(\gamma) - \pi(\gamma \to \lambda'), \\ \text{for any } \gamma \in \lambda', \text{ where } \pi(\gamma \to \Delta\lambda) \text{ and } \alpha(\gamma) \text{ are the} \\ \text{functions appearing in the definition of } G(t) \text{ as a purely} \\ \text{discontinuous and temporally homogeneous process.} \end{cases}$$

function $L(t)$, itself a function of the process $G(t)$, becomes a conventional tool to help one do so.

The following facts can now be proved:

If (a) $G(t)$ is a Markovian random function with range space Γ and if (b) $\Lambda = \{\lambda\}$ is a class of sets in Γ that is enchained by $G(t)$, then $L(t)$ (the random function defined in condition C above) is Markovian.

Suppose (b) in the above statement is true. If (c) $G(t)$ is purely discontinuous and temporarily homogeneous, then $L(t)$ is also.

If (a), (b), and (c) in the above statements are true, then for each $\lambda_0 \in \Lambda$, for all $\Delta\lambda \in B_\Lambda$, and for all $t > 0$, the function $P_1(\Delta\lambda, t|\lambda_0, 0)$—that is, the conditional probability for $L(t)$ to belong to $\Delta\lambda$, given that $L(0)$ coincides with λ_0—satisfies an equation (see 42) of the Chapman–Kolmogorov type (see 40).

It can be seen in these facts that the Markovian property of $L(t)$ (and also its purely discontinuous property) leads to a diffusion equation, a linear equation for the transitions of the process $G(t)$ through the sets of Λ. As such, it provides a convenient alternative to the diffusion equation governing $G(t)$ because any specific realization of it will involve only enough parameters to specify the sets of Λ rather than the detailed structure of the states γ. When applied to the cascade-shower problem, for example, this technique of taking a function of a Markovian process makes solutions possible.

BIBLIOGRAPHY. F.N. DAVID, *Games, Gods and Gambling: The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era* (1962), for the early problems of probability theory prior to Laplace; MARK KAC, *Statistical Independence in Probability, Analysis and Number Theory* (1959), a highly attractive small book showing the influence of probability in other areas of mathematics; WILLIAM FELLER, *An Introduction to Probability Theory and Its Applications*, 3rd ed., 2 vol. (1967), the best reference for discrete probability and introduction to Markovian processes; MICHEL M. LOEVE, *Probability Theory*, 3rd ed. (1963), the leading reference for advanced probability based on a rigorous measure-theoretic foundation; MURRAY ROSENBLATT, *Random Processes* (1962), a presentation in a limited number of pages of some of the basic problems of stochastic processes, including reference to the relation between Markovian and non-Markovian processes; BAYARD RANKIN (ed.), *Differential Space, Quantum Systems, and Prediction* (1966), includes quantum-theoretic as well as measure-theoretic foundations of probability, current reconstruction and application of Wiener's original study of Brownian motion; I.I. GIKHMAN and A.V. SKOROKHOD, *Introduction to the Theory of Random Processes* (1969; orig. pub. in Russian, 1965), a survey of the modern results in the theory of stochastic processes, appropriate for readers who have a sound background in measure theory.